

Fall 12-2018

SEX-SPECIFIC EFFECTS AND GENE-SEX INTERACTIONS IN SERUM METABOLOME LEVELS: THE ATHEROSCLEROSIS RISK IN COMMUNITIES STUDY

Zhe Wang
UTHealth SPH

Follow this and additional works at: https://digitalcommons.library.tmc.edu/uthsph_dissertsopen



Part of the [Community Psychology Commons](#), [Health Psychology Commons](#), and the [Public Health Commons](#)

Recommended Citation

Wang, Zhe, "SEX-SPECIFIC EFFECTS AND GENE-SEX INTERACTIONS IN SERUM METABOLOME LEVELS: THE ATHEROSCLEROSIS RISK IN COMMUNITIES STUDY" (2018). *UT School of Public Health Dissertations (Open Access)*. 2.

https://digitalcommons.library.tmc.edu/uthsph_dissertsopen/2

This is brought to you for free and open access by the School of Public Health at DigitalCommons@TMC. It has been accepted for inclusion in UT School of Public Health Dissertations (Open Access) by an authorized administrator of DigitalCommons@TMC. For more information, please contact digitalcommons@library.tmc.edu.

**SEX-SPECIFIC EFFECTS AND GENE-SEX INTERACTIONS IN SERUM
METABOLOME LEVELS: THE ATHEROSCLEROSIS RISK
IN COMMUNITIES STUDY**

By

ZHE WANG, BM, MSC

APPROVED:

ERIC BOERWINKLE, PHD

HAN CHEN, PHD

ALANNA C. MORRISON, PHD

MICHAEL SWARTZ, PHD

DEAN, THE UNIVERSITY OF TEXAS
SCHOOL OF PUBLIC HEALTH

Copyright
by
Zhe Wang, BM, MSc, PhD
2018

SEX-SPECIFIC EFFECTS AND GENE-SEX INTERACTIONS IN SERUM
METABOLOME LEVELS: THE ATHEROSCLEROSIS RISK
IN COMMUNITIES STUDY

by

ZHE WANG

BM, Peking University Health Science Center, 2012
MSC, Wageningen University and Research Center, 2014

Presented to the Faculty of The University of Texas

School of Public Health

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

The University of Texas
School of Public Health
Houston, Texas
December, 2018

ACKNOWLEDGEMENTS

I would like to express my deepest appreciation to my academic advisor, Dr. Eric Boerwinkle, who provide persistent support and great advice for my education and research work over the years, which shaped me as a student, scholar and person. Without his guidance, completion of this doctoral dissertation would not have been possible. I would also like to express my sincere gratitude to my dissertation committee members, Drs. Han Chen, Alanna C. Morrison, and Michael Swartz, who offered me not only valuable comments for my dissertation, but also encouragement, advice to pursue my career goals.

I would like to acknowledge Dr. Bing Yu, for her tremendous help in building my skill sets on metabolomics and epidemiological analyses. I am also thankful to Michael Brown, for his help in cloud programing; and Megan Grove, for her contribution to exome chip genotyping. My appreciation also goes to all those who helped me to finish my Ph.D. training, for instance the great faculty and staff at Human Genetics Center and School of Public Health, and all my fellow friends and peers at University of Texas Health Science Center at Houston. Their supports made my challenging PhD training a journey of joy.

Finally, I am very much indebted to my loving parents, who encouraged and helped me at every stage of my personal and professional life with their unconditional love and solid support. My fiancé Wenjun, who is always there standing by me through numerus working days and nights. I am so grateful to have them as my wonderful family.

SEX-SPECIFIC EFFECTS AND GENE-SEX INTERACTIONS IN SERUM
METABOLOME LEVELS: THE ATHEROSCLEROSIS RISK
IN COMMUNITIES STUDY

Zhe Wang, MB, MSc, PhD
The University of Texas
School of Public Health, 2018

Dissertation Chair: Eric Boerwinkle, PhD

Metabolomic signatures associated with complex disease have been identified. Metabolomic profiling and the integration of genomic data have proven to be powerful tools to investigate genetic effects underlying intermediate phenotype levels and may facilitate improved understanding of pathophysiologic processes of disease. However, most published studies did not consider sex as an effect modifier, analyze sex-specific effects, nor gene by sex interactions. One reason can be incomplete knowledge of the power of statistical methods used in a given dataset.

I first investigated sex-specific genetic effects by performing sex-stratified exome-wide association studies for 271 chromatography-mass spectrometry measured metabolites in the Atherosclerosis Risk in Communities (ARIC) study, followed by a conventional Z test to evaluate the heterogeneity of genetic effects between men and women. We used African-Americans as the discovery sample and pursued exome-wide significant (false discovery rate $Q \leq 5\%$) genes for replication in European-Americans. Overall, we identified and replicated variants in 12 genes associated with metabolite levels, one of which, rs11555566 in ADA,

was a novel common variant suggesting a larger effect in men compared to women for association with N1-methyladenosine levels.

I then focused on rare genetic variants and sex interactions on serum metabolite levels and evaluated the joint effect of genetic main effects and gene-sex interactions in the same discovery and replication population. Using gene-based rareGE and MiSTi approaches, we observed and replicated 14 gene-metabolite associations through joint test, three of which were novel, including PLA2G7- arachidonate (20:4n6), PTER- N-acetyl-beta-alanine and NPC2- leucylserine. Significance of the NPC2- leucylserine association arose from both genetic main effects and gene-sex interaction effects.

Finally yet importantly, I carried out a simulation study to investigate the performance of two aforementioned emerging methods in detecting rare variant gene-sex interaction effects on a quantitative phenotype. Compared with conventional burden tests, rareGE and MiSTi have more power under a wide range of scenarios. Simulation results also illustrate that an approach that jointly tests the genetic main effects and gene-sex interactions increases statistical power and has the potential to uncover novel genetic signals that have not been identified previously.

In conclusion, our study suggests sex-specific genetic effects on the metabolome, and reports novel genetic variants associated with metabolite levels. Use of simulated data provides insights into the power and desired sample size in conducting rare variant G×E interaction studies for these newly introduced methods, justify their use in practice.

TABLE OF CONTENTS

List of Tables	i
List of Figures	ii
List of Appendices	iii
Chapter I. Background	1
Literature Review.....	1
Sex Difference in Cardiovascular Disease.....	1
Gene-Environment Interactions	3
Rare Variant G×E Interactions	5
Metabolomics and the metabolome by sex	6
Public Health Significance.....	10
Chapter II. Sex-Specific Genetic Effects on the Serum Metabolome in the Atherosclerosis Risk in Communities (ARIC) Study	12
Abstract	13
Introduction.....	15
Methods.....	17
Study Sample	17
Measurements of Metabolites	17
Genotypes	18
Statistical Analyses	19
Results.....	21
Common single variant results.....	21
Gene based rare and low-frequency variants	22
Discussion	23
References	26
Chapter III. Rare and Low-Frequency Genetic Variant × Sex Interactions Identify Novel Loci influencing the serum Metabolome in the Atherosclerosis Risk in Communities (ARIC) Study	35
Abstract	36
Introduction.....	37
Methods.....	39
Study Sample	39
Measurements of Metabolites	39
Genotypes	40
Statistical Analysis:.....	41
Results.....	42
Discussion	44

References	49
Chapter IV. Power of Two Emerging Methods for Detecting and Characterizing Gene×Sex Interaction Effects for Rare Variant Analyses Compared to Standard Stratified Analyses	58
Abstract	59
Introduction	60
Methods- Simulation Studies	62
Study Sample	62
Simulation Design	63
Part 1. Detecting gene-sex interaction	63
Part 2. Effect of sample size on power	64
Results	66
Part 1. Performance in detecting gene-sex interaction	66
Part 2. Effect of sample size on power	68
Discussion	69
References	72
Chapter V. Synthesis	85
Summary of results	85
Strength and Limitations	88
Future Directions	89
Conclusions	92
Appendices	93
References	135

LIST OF TABLES

Table II-1 Baseline characteristics of analyzed participants in the ARIC study	33
Table III-1 Genes discovered and replicated through jointly testing the genetic main effects and gene-sex interactions in the ARIC study	57
Table IV-1 Characteristics of the 10 selected genes in European- Americans from the ARIC exome chip data.....	76
Table IV-2 Power of various methods under scenarios with 2 settings of genetic main effects (setting 1 genetic main effects in the same direction and setting 2 genetic main effects in opposite directions) and two scenarios of interaction effects (GxE effects in the same direction and GxE effects in opposite directions), respectively. The significance levels α are 0.05, and bonferroni corrected 4.95×10^{-6} , respectively. 2a. the size of interaction effect $c=0.5$; 2b. the size of interaction effect $c=1$; 2c. the size of interaction effect $c=1.5$; 2d. the size of interaction effect $c=2$	77

LIST OF FIGURES

Figure I-1 Sex difference in cardiovascular disease (CVD) at a glance	2
Figure II-1 Significantly identified and replicated gene-metabolite pairs revealed by sex-stratified exome-wide association studies.	34
Figure IV-1 Power of three methods with positive genetic main effects and two scenarios of interaction effects (1a & 1c. GxE in the same direction; 1b & 1d. GxE in opposite directions). The significance threshold for 1a & 1b is Bonferroni corrected (4.95×10^{-6}), for 1c & 1d is 0.05.....	79
Figure IV-2 Power of three methods with genetic main effects in both directions and gene-sex interaction effects in the same direction ($c=1$). The significance threshold for 2a is 0.05 for 2b is Bonferroni corrected 4.95×10^{-6}	80
Figure IV-3 Power of three methods with genetic main effects in both directions and gene-sex interaction effects in the same direction ($c=2$). The significance threshold for 3a is 0.05 for 3b is Bonferroni corrected 4.95×10^{-6}	81
Figure IV-4 Power of three methods with genetic main effects and gene-sex interaction effects in different directions ($c = \pm 2$). The significance threshold for 4a is 0.05 for 4b is Bonferroni corrected 4.95×10^{-6}	82
Figure IV-5 Power of three methods with genetic main effects in the same direction and gene-sex interaction effects in different directions ($c = \pm 2$). The significance threshold is Bonferroni corrected 4.95×10^{-6}	83
Figure IV-6 Power comparisons of three methods with positive genetic main effects and two scenarios of interaction effects (6a & 6b. GxE in opposite directions $c = \pm 1$; 6c & 6d. GxE in the same direction $c = 1$) under different significant thresholds and varied sample size.	84

LIST OF APPENDICES

Appendix A. supplemental materials for Chapter 2.....	93
Appendix B. supplemental materials for Chapter 3.....	105

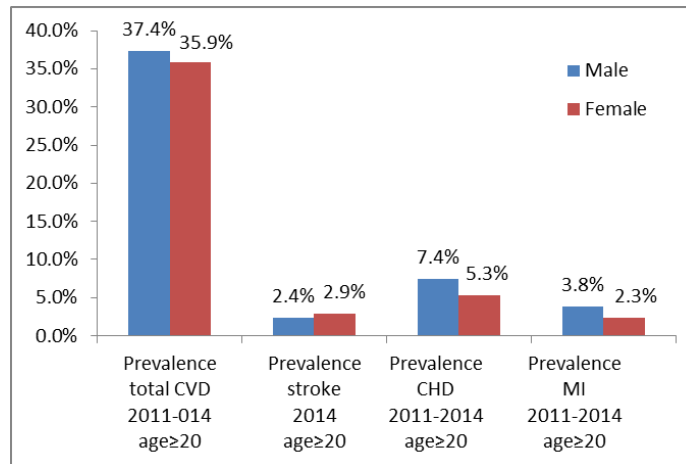
CHAPTER I. BACKGROUND

Literature Review

Sex Difference in Cardiovascular Disease

Cardiovascular disease (CVD) remains the leading cause of mortality across race groups for both women and men in the United States (1), but there are substantial sex differences in the prevalence and presentation of different CVD (**Figure I-1**). The percentage of adult men living with major manifestations of CVD exceeds those of adult women, except for stroke. Men develop coronary artery disease (CAD) earlier and usually present with more severe atherosclerosis in their coronary arteries than women, but the risk of heart failure and mortality rate following myocardial infarction is higher in women than men (2, 3). Despite unfavorable progression of CVD, women may be less likely to receive optimal diagnosis and timely treatment because the presentation of symptoms in women with acute coronary disease is “atypical”- women are significantly less likely to report chest pain or discomfort compared with men (4). Although sex differences in prevalence, age of onset, progression and outcome of CVD have been well-documented (5), the biological underpinnings of these differences are not well-understood.

Figure I-1 Sex difference in cardiovascular disease (CVD) at a glance



Note: Data adapted from *Heart disease and stroke statistics 2017 update: a report from the American Heart Association* (5)

Sex differences in the epidemiology of CVD may arise from different exposures of environmental and lifestyle risk factors, for instance, heavy alcohol consumption, tobacco use, and physical inactivity (6, 7). In addition to differences in health-related lifestyles, to understand the biological determinants of observed sex differences in CVD, a natural starting point is the biological effects of the sex chromosomes (8), and related sex hormones effects, such as estrogen levels (9). However, it is also important to consider genetic variants on the autosomes that may affect risks of developing CVD differently in women and men. Previous studies have identified different genetic variants influencing CVD risk in men compared to women, or found significant genotype-sex interactions for CVD or related traits (10-13). For example, Silander et al. reported that variants in *CPB2* and *USF1* genes have a female-

specific risk for coronary heart disease (CHD) and/or CVD, while a variant (rs2069840) in *IL6* shows strong association with CVD in men but not in women (12). In a large study of more than 200,000 individuals, 49 loci were found to be associated with waist-to-hip ratio, an independent risk factor to CVD; 20 of the 49 loci show significant sexual dimorphism, 19 of which present a stronger effect in women (13).

Despite previously reported sexually dimorphic genetic-disease associations (14), most large published meta-analyses (15-19) do not take sex differences into account beyond adjusting for sex as a covariate (15, 19, 20). Thus, potential sex differences are important components for further genetic epidemiologic research of CVD and its risk factors.

Gene-Environment Interactions

Traditional genome-wide association studies (GWAS) utilizing common genetic variants have successfully identified a large number of loci associated with complex diseases and traits. However, a large proportion of the heritability of these diseases/traits remains unexplained (21). To find the “missing heritability”, rare variants, structural variations, as well as gene-environment (G×E) interactions have been suggested to extend beyond straight-forward genome-wide association approaches (22).

Gene-environment interactions are defined in this proposal as different effects of a genotype on disease risk between differing environmental exposures, including sex. Equivalently, G×E interactions may be defined by different effects of an environmental exposure on disease risk in persons with different genotypes (23, 24). Studying G×E

interactions is important, as it may extend our knowledge of the genetic architecture of complex traits and improve our understanding of the underlying mechanisms of common diseases for novel and known loci (25-27). Although premature at this point, knowledge of gene-sex interactions could lead to different genetic risk algorithms and treatment recommendation in men compared to women.

Since 2010, several large-scale genome-wide G×E studies have successfully identified novel loci accounting for the modifying effects of environmental exposures such as age, sex, BMI, alcohol consumption, smoking status on CVD-related intermediate traits (28-37). Although none of the studies directly model clinical defined CVD, using risk factors such as blood pressure, lipid profiles, and obesity, these studies have successfully identified novel common variant loci related to CVD risk that were not detected via analysis of main effects alone. For example, a genome-wide meta-analysis of 114 studies in up to 320,485 European-ancestry individuals reported 4 novel loci for BMI that showed age-specific effects, and 17 novel loci with sex specific effects on BMI (29). Though it is tempting to consider conducting an association test within each stratum of environmental exposures, a recent study (38) compared a stratified analysis approach and a 2 degree of freedom (DF) joint test for studying G×E interactions and suggested that inclusion of G×E interactions is important in terms of identifying novel signals, particularly for rare and low-frequency variants.

Rare Variant G×E Interactions

Availability of high throughput DNA sequencing technologies and large-scale imputation reference panels (39) offer an opportunity to investigate rare and low-frequency genetic variants with minor allele frequency (MAF) $\leq 5\%$ across the genome. Analysis of G×E interactions involving rare variants may identify novel loci, and characterize rare variant G×E interactions in previous loci identified by GWAS of common variants. However, unlike well-established G×E interaction tests for common variants (40, 41), methods development for detecting rare variant G×E interactions is challenging for several reasons. First, considering typical sample sizes of most published GWAS studies, a single marker test is underpowered for rare and low frequency variants. Second, conventional burden tests that simply summarize the total number of variants within a region and fit a model for the rare variant burden by environment interaction term, often result in inflated type 1 error rates and biased estimates when the rare variants and environment are not independent (i.e. G×E correlation) (42).

Recently developed novel approaches for testing rare variant G×E interaction effects (42-47) face limitations. Jiao and colleagues (45, 46) treated genetic main effects as fixed effects, which may suffer from inflated type I error. Lin et al proposed an interaction Sequence Kernel Association Test (42) that is powerful when both positive and negative directions of G×E effects exist, yet loses power when the variants in the set have the same direction of G×E. Finally, Tzeng (43) assumed comparable magnitude of the variance component parameters for genetic main effects and G×E interactions, which may not be true.

Further work is underway to overcome the aforementioned limitations. Su et al. proposed a novel and rigorous framework to derive independent score statistics for fixed effects and the variance component that is more powerful to test $G \times E$ interaction terms of rare variants (48). A joint test that allows one to simultaneously test genetic main effects and interaction effects and requires no assumption about the magnitude of the variance component parameters for the genetic main effects and $G \times E$ interactions was proposed and successfully implemented by Chen and colleagues (49). The former interaction-only test allows detecting $G \times E$ interactions regardless of the genetic main effect, while the latter joint testing approach aims to detect associated genetic effects allowing for $G \times E$ interactions. Applying these newly developed methods to study rare variant $G \times E$ interaction in CVD-related traits, for instance the metabolomic data that will be reviewed in the next section, may improve our understanding of the underlying pathophysiology of disease.

Metabolomics and the metabolome by sex

Metabolomics is one of the “-omics” disciplines that systematically studies small-molecule metabolites found in biological samples, such as cells, biofluids, tissues or organisms. These metabolites are produced and modified by a variety of chemical and physiologic processes, such as amino acid and lipid biosynthesis, carbohydrate anabolism, and xenobiotic metabolism. The entire ensemble of small-molecule metabolites presented in a biological sample is generally referred as the metabolome. These small-molecule

metabolites may reveal pathologic or etiologic pathways to complex diseases because they represent intermediates that profile biological status closely related to phenotypes (50).

At present, there are two major instrument platforms for measuring metabolite levels in biological samples, namely nuclear magnetic resonance (NMR) and chromatography combined with mass spectrometry (MS)-based metabolic profiling (51-53). There are also two major distinct technologic approaches, “untargeted” and “targeted”, to metabolite measurements (54). Untargeted metabolomics aims to analyze all of the measureable analytes in a sample including unknown chemicals, and targeted metabolomics means to measure an *a priori* defined group of chemically characterized metabolites (e.g. lipids). Several review papers have described and contrasted these platforms and approaches (55, 56).

Acknowledging concerns about the semi-quantitative nature of the untargeted MS-based approach, it has notable advantages for detecting and quantifying (at least relative quantification) as many metabolites as possible in a biological sample with high sensitivity. Therefore applying such an approach is able to achieve high-throughput profiling of the metabolome.

In the past few years, numerous epidemiological studies utilizing metabolomics have successfully linked metabolite levels to the etiology and progression of complex diseases such as hypertension, CVD and diabetes in individuals with and without European-ancestry (57-69). The identified CVD metabolomic signatures include dietary phosphatidylcholine metabolites, acylcarnitines (61, 65), as well as several other lipid classes such as polyunsaturated fatty acids (FAs) (66, 68, 69). Such metabolomic signatures were involved

in CVD risk via various potential mechanisms. For example ω -3 FAs may prevent arrhythmias, lower heart rate and blood pressure, decrease platelet aggregation, and lower triglyceride levels (70). The latter is accomplished by reducing hepatic very-low-density lipoprotein and triglycerides synthesis and secretion and enhancing the triglycerides clearance from chylomicrons (71, 72).

In addition, both traditional GWAS and sequence analyses across the whole genome or exome have successfully identified and verified hundreds of genetic loci associated with metabolite levels (73-85), and many of them can be further linked to clinically relevant factors of disease development. An example of integrating genomics and metabolomics to promote novel biomarker discovery and better understand etiologic pathways of complex disease is the story of hexadecanedioate. In a whole exome sequencing study of African-Americans, Yu et al. identified a loss-of-function (LoF) variant in *SLCO1B1* that was associated with increased levels of hexadecanedioate, which for the first time, was reported for its relationship with heart failure risk (81). Hexadecanedioate, a long-chain dicarboxylic acid, was also reported to be significantly associated with increased blood pressure and mortality (59, 81). The aforementioned genetic and metabolomic evidence together implicated a potential pathway for heart failure and opened up the possibility of further hypothesis tests and experimental studies.

Limited work has been done that shows that the metabolomic profiles of men and women are different, and sex-specific metabolism-related genetic polymorphisms have been identified through sex-stratified GWAS in European-ancestry populations (86, 87). Pathway

analysis has revealed gender-specific pathway differences in the serum metabolome (87). Moreover, a single-nucleotide polymorphism (SNP) in *CPSI* rs715 that previously showed a strong sex difference in association with glycine (86), yields a strikingly significant and protective association with decreased risk of CAD only in women (88). Metabolomics studies also reveal the sex-specific effects of a SNP rs646776 in *SORT1*, a known low-density lipoprotein cholesterol locus (89). Additional systematic studies are needed to better understand the modifying effect of sex on the human metabolome and its genetic determinants with a particular focus on rare variants. Additionally, current understandings of sex differences in the metabolome have solely originated from studies in European-Ancestry populations and there is a need for expanding these studies to underrepresented populations, such as African-Americans (AAs).

Public Health Significance

There are encouragements from the National Heart, Lung, and Blood Institute (NHLBI) for conducting additional research addressing the public health concerns of sex differences in cardiovascular diseases (90). Although significant sex-related differences in CVD epidemiology are appreciated (1, 5), less effort has been devoted to uncovering its etiology. Studies that go beyond common single nucleotide variants to investigate the role of rare genetic variants as well as studies with more complex statistical analyses to evaluate the impact of sex alteration of disease phenotypes are warranted.

It is generally accepted that complex diseases such as CVD are not only caused by genetic or environmental factors alone, but also the interactions between them (26). Serum metabolite levels ultimately are the reflections of functional activities of genes and environmental exposures (91, 92). Given the nature of metabolite levels and numerous aforementioned work that have linked metabolites to complex diseases, they can serve as ideal intermediates to understand the effects of G×E interactions on complex diseases. Rare and low frequency ($MAF \leq 5\%$) variants make up the vast majority of the genetic variation in the genome (93), and may account for part of the missing heritability along with G×E interactions (22). Studying the integration of rare and low frequency genetic variants, sexual dimorphism, and metabolomics may improve our comprehensive understanding of the underlying pathophysiology. To date, there is no study systematically utilizing and comparing methods developed for testing rare variant G×E interactions in large-scale human population data. In addition, investigating gene-sex interactions or sex-specific genetic

variants related to CVD-related traits has not been done in AAs. New studies focusing on AAs will help address this important knowledge gap.

This dissertation research leverages existing data from a large population-based multi-ancestry cohort: the Atherosclerosis Risk in Communities (ARIC) study that contains well-characterized AA and European-American (EA) participants to investigate whether there are sex-specific differences in the genetic determination of serum metabolome levels. As mentioned above, the serum metabolome may serve as an ideal surrogate/biomarker for disease status, including CVD. Therefore, Chapter 2 of this dissertation describing the sex-specific genetic effects on the serum metabolome responds to the aforementioned rising public health concerns on sex-related differences in CVD-related phenotypes (20, 22, 90). Chapters 3 and 4 address the challenge of testing rare variant G×E interactions through estimating gene-sex interactions on the serum metabolome with a particular focus on rare and low-frequency genetic variants using two emerging methods, and comparing the power of differing methods in simulation studies. Chapter 5 presents a summary of the findings and a perspective about further rare and low-frequency genetic variant G×E interaction studies.

The results of this dissertation are expected to offer new evidence about sex-specific genetic influences on the human metabolome and report novel genetic variants that were not previously identified when gene-sex interaction parameters were omitted in previous studies. This dissertation will also provide insights into the power and desired sample size for conducting rare variant G×E interaction studies, which may advance the understanding of current G×E interaction results and benefit future studies.

**CHAPTER II. SEX-SPECIFIC GENETIC EFFECTS ON THE SERUM
METABOLOME IN THE ATHEROSCLEROSIS RISK IN COMMUNITIES (ARIC)
STUDY**

Abstract

Metabolomic signatures associated with complex disease, such as cardiovascular disease, have been identified. Metabolomic profiling and the integration of genomic data have proven to be powerful tools to investigate genetic effects underlying intermediate phenotype levels and may facilitate improved understanding of pathophysiologic processes of disease. However, most published studies did not consider sex as an effect modifier, analyze sex-specific effects, nor gene by sex interactions. This study investigated sex-specific genetic effects on serum metabolite levels and evaluated the estimated heterogeneity of genetic effects between men and women. We analyzed 3,540 individuals from the Atherosclerosis Risk in Communities (ARIC) study with metabolite measurements and exome chip genotyped data. We performed sex-stratified exome-wide association studies for 271 chromatography-mass spectrometry measured metabolites in ARIC African-Americans as the discovery sample and pursued exome-wide significant (false discovery rate $Q < 5\%$) genes for replication in European-Americans. We identified and replicated variants in 12 genes through either common single variant analysis or gene-based burden tests in sex-stratified exome-wide association analyses. For example, rs11555566 in *ADA* is a novel common variant associated with N1-methyladenosine levels. Results of rs11555566 suggested a larger effect in men (estimated effect size 0.18-0.22) as compared to women (estimated effect size 0.14-0.17), but the difference was not significant. Variants in 6 genes suggested differing genetic effects on metabolite levels observed through testing for difference of the effect size estimates in sex-stratified results, although the difference

between sexes was not replicated. Our study suggests that sex-specific genetic effects of metabolites may exist, but the lack of consistency in testing sex differences of the genetic effects between discovery and replication samples underscores that future studies should consider sex-specific effects with enhanced statistical methods and tools.

Introduction

Metabolomics is one of the “-omics” disciplines that systematically studies small-molecule metabolites found in biologic samples such as cells, biofluids, tissues or organisms. These metabolites are produced and modified by a variety of chemical and physiologic processes, such as amino acid and lipid biosynthesis, carbohydrate anabolism, and xenobiotic metabolism. The ensemble of small-molecule metabolites presented in a biologic sample is referred to as the metabolome. These small-molecule metabolites may reveal pathologic or etiologic pathways to complex diseases because they represent intermediates that at least partially profile the biological status of an individual and are closely related to a number of risk factor and disease-related phenotypes (1).

Numerous epidemiologic studies utilizing metabolomics have successfully related metabolite levels to the etiology and progression of complex diseases such as hypertension, cardiovascular disease CVD, and diabetes in both Whites and non-Whites (2-14). In addition, both traditional genome-wide association studies (GWAS) and sequencing analyses across the whole genome or exome have successfully identified and verified hundreds of genetic loci with metabolite levels (15-25), and many of them can be further linked to clinically relevant factors of disease development. An example of integrating genomics and metabolomics to promote novel biomarker discovery and better understand the etiological pathways of complex disease is the story of hexadecandioate. In a whole exome sequencing study of African-American population, Yu et al. identified a loss-of-function (LoF) variant in *SLCO1B1* that was associated with increased levels of hexadecanedioate, which for the first

time, was reported to be associated with incident heart failure (23). Hexadecanedioate, a long-chain dicarboxylic acid, was also reported to be significantly associated with increased blood pressure and mortality (4, 23). The aforementioned genetics and metabolomics evidence together implicated a potential novel pathway for heart failure and opens up the possibility of further hypothesis testing and experimental studies.

Limited work has shown that the metabolomic profiles of men and women are different, and sex-specific metabolism-related genetic polymorphisms have been identified through sex-stratified GWAS (26, 27). Pathway analysis has also revealed sex-specific pathway differences in the serum metabolome (27). Moreover, a single-nucleotide polymorphism (SNP), rs715 in *CPS1*, that previously showed a strong sex difference in association with glycine (26), yields a strikingly significant and protective association with decreased risk of coronary artery disease only in women (28). Metabolomics studies also reveal the sex-specific effects of a SNP rs646776 in *SORT1*, a known low-density lipoprotein cholesterol locus (29). Additional and systematic studies are needed to better understand the modifying effect of sex on the human metabolome and its genetic determinants. Additionally, current understanding of sex differences in the metabolome has solely originated from studies in Whites, and there is a need for expanding these studies to underrepresented populations, such as African-Americans (AAs). Therefore, we investigated whether there are sex-specific differences in the genetic effects on the metabolome using a subset of AAs in the Atherosclerosis Risk in Communities (ARIC) study, and examined the replication of these sex-specific effects in European-Americans (EAs) from the ARIC study.

Methods

Study Sample

The ARIC study is a population-based prospective cohort study of 15,792 adults from four U.S. communities (Forsyth County, NC; Jackson, MS; suburbs of Minneapolis, MN; and Washington County, MD), which has been described in detail previously (30). ARIC included both EAs and AAs aged 45-64 at the baseline examination (1987-1989).

Participants completed three additional triennial follow-up examinations, a fifth exam in 2011-2013, and a sixth exam in 2016-2017. Included in this analysis were 3,540 participants with metabolite measurements and exome chip genotyped data at the baseline examination.

The ARIC study has been approved by the institutional review boards at each site, and written informed consent was obtained from all participating individuals.

Measurements of Metabolites

Metabolite profiling was completed in 2010 (batch 1) and 2014 (batch 2) using fasting serum samples that had been stored at -80°C since collection at the baseline examination. Batch 1 were all AAs and Batch 2 included both AAs (24.8%) and EAs (75.2%). In total, 602 metabolites were detected and semi-quantified by Metabolon (Durham, USA) using untargeted, gas- and liquid- chromatography-mass spectrometry (GC-MS and LC-MS)-based protocols (31, 32). To evaluate batch effects, a set of 97 samples were measured in both the 2010 and 2014 batches. There were 384 named metabolites that were

identified to be present in both batches and these metabolites will be used for this thesis research.

In the present study, sample-level quality control was performed to remove individuals with missing values for more than 40% of the measured metabolites (1 sample was removed from batch 2). After sample-level quality control, metabolomic profiles were available in 2479 AAs and 1553 EAs. Exclusion criteria for metabolites includes: 1) six-there metabolites were excluded as more than 40% of the samples have missing values or values below the detection limit (BDL) within each batch; and 2) fifty metabolites were excluded as the Pearson correlation coefficient (r) between 2010 and 2014 measurements on the same stored sample (at least 46 out of the 97 pairs) is less than 0.30. Thus, this study was based on an evaluation of 271 named metabolites. Metabolite levels were analyzed as continuous variables, where missing/BDL values were imputed using random forest imputation based on the remaining observed measurements (33, 34).

Genotypes

Genotyping was performed with the Illumina HumanExome BeadChip v1.0 (“exome chip”) querying 247,870 single nucleotide variants (SNVs) at the baseline examination in 11,071 EAs and 2,953 AAs in the ARIC study. The exome chip data was selected rather than sequence data that is also available in ARIC because using exome chip data maximizes the available sample size for this proposed analysis. To improve accurate calling of rare variants, genotyped data from ARIC along with 10 other studies from the Cohorts for Heart and Aging

Research in Genomic Epidemiology (CHARGE) Consortium were pulled together for joint calling, details were described elsewhere (35). A total of 8,994 variants were excluded after laboratory quality control steps, for instance call rate <95%, Hardy-Weinberg equilibrium test P value (pHWE) < 1×10^{-6} , and poorly clustering variants (35). SNVs with missing rate >5% were removed from analysis.

Exome chip variant annotation was completed using the Whole Genome Sequencing Annotation (WGSA) pipeline v055 (36), including dbNSFP v2.9 (37). Functional variants and genes were determined using ANNOVAR (38) according to the reference genome GRCh37/hg19 and National Center for Biotechnology Information RefSeq.

Statistical Analyses

Metabolite levels were winsorized (99%) within each batch, respectively. Due to right-skewed distributions of many metabolite levels, natural log transformation was applied to most metabolites prior to analyses. For metabolites that were still not normally distributed, a rank based inverse normal transformation was used. The transformation methods applied to each metabolite are provided in Appendix A- Supplemental Table 1.

Race-specific exome-wide association studies for each metabolite level were conducted in men and women separately. Linear regression analyses were performed for the continuous metabolite levels. For common variants with MAF>5%, single variant association tests assuming an additive genetic model were conducted. Because our primary focus was on rare and low-frequency variants, we aggregated rare and low-frequency variants (MAF \leq

5%) in groups based on a gene's exons using burden tests (39). The unit-of-analysis is an annotated gene. All annotated coding variants, such as splicing, stop-gain, stop-loss, nonsynonymous, and indels within the gene were aggregated for the analysis. Genes with cumulative minor allele count < 3 within men or women of each race group were not analyzed. Models were adjusted for age and population substructure using the first three ancestry principal components (PCs) (40), with additional adjustment for estimated glomerular filtration rate (eGFR) and batch effects for metabolites.

To test each SNV/gene and metabolite for difference of the effect size estimates for the SNV/gene calculated in the sex-specific analyses, we used an approximately normally distributed test statistic, Z (41). This Z test was selected as opposed to a sex-pooled multiplicative interaction test because traditional linear regression assumes homoscedasticity across all combinations of G and E which is often violated for rare variant burden test (42).

$$Z = \frac{\hat{\beta}_{men} - \hat{\beta}_{women}}{\sqrt{se(\hat{\beta}_{men})^2 + se(\hat{\beta}_{women})^2}}$$

We used AAs as the discovery sample and conducted replication in EAs. Using a false discovery rate (FDR) to correct for number of genes/SNV tested while considering the 271 metabolites, we defined exome-wide significant genes/SNVs as those with FDR $Q \leq 5\%$ in discovery AAs; these genes/SNVs were pursued for replication analyses in EAs. Replication was defined as those genes/SNVs with consistent directions of the effect, and FDR $Q \leq 5\%$, corrected for the number of genes/SNVs taken forward to evaluate in EAs. All

statistical analyses were conducted in R version 3.4 (R Foundation for Statistical Computing, Vienna, Austria).

Results

We conducted sex-specific exome-wide association analyses in 1292 women and 720 men among ARIC AAs. A total of 827 EA women and 701 EA men were analyzed for replication. The baseline characteristics of both men and women in AAs and EAs were shown in **Table II-1**. The average age of women and men were comparable in both AAs (52.9 ± 5.6 vs. 53.0 ± 5.8 years) and EAs (54.3 ± 5.8 vs. 54.9 ± 5.8 years).

Common single variant results

Sex-stratified exome-wide association analysis identified and replicated (across race groups) common variants in 9 genes (*ADA*, *ALMS1*, *DMGDH*, *DUSP11*, *FBX07*, *GCKR*, *KLKB1*, *LACTB*, and *VNN1*) that were shown to be associated with metabolite traits in women. Common variants in the first four genes (*ADA*, *ALMS1*, *DMGDH* and *DUSP11*) were significant and replicated (across race groups) in men (Appendix A- Supplemental Table 2-3, **Figure II-1**). Eight out of the nine observed significant genes were consistent with those previously identified through analyses using pooled samples of men and women. One novel missense variant rs11555566 (*ADA*) with MAF ~6.2% - 7.8% was associated with N1-methyladenosine in women (AA: $\hat{\beta} = 0.14$ $p = 2.44 \times 10^{-11}$ FDR $Q = 4.67 \times 10^{-6}$, EA: $\hat{\beta} = 0.17$

$p = 7.55 \times 10^{-5}$) and men (AA: $\hat{\beta} = 0.18$ $p = 2.09 \times 10^{-7}$ FDR $Q = 0.04$, EA: $\hat{\beta} = 0.22$ $p = 9.71 \times 10^{-6}$), respectively.

In order to reveal sex-specific effects, we evaluated the estimated common genetic effects for heterogeneity between men and women. One common variant, rs3746414 (*ZFP64*) showed a significant ($p = 1.42 \times 10^{-8}$, FDR $Q = 0.05$) difference between men and women for its association with propanediol in AAs. The significant association between rs3746414 and propanediol was only observed in AA men ($\hat{\beta} = -0.30$, $p = 7.00 \times 10^{-10}$). For AA women, the observed effect was positive but not significant ($\hat{\beta} = 0.04$, $p = 0.22$). Although we observed a similar difference in the direction of effects in EAs (negative effect in EA men and positive effect in EA women, data not shown), the sex-difference in genetic effects of rs3746414 on propanediol was not significantly replicated in EAs.

Gene based rare and low-frequency variants

For the gene-based approach, we report four known metabolite genes (*ACAD8*, *CCBL1*, *ACY1* and *DMGDH*) that were successfully identified and replicated in female-only burden tests. The latter two (*ACY1* and *DMGDH*) pass the significance thresholds of identification and replication using male-only burden tests as well (Appendix A-Supplemental Table 4, **Figure II-1**). In the Z tests that evaluated the aggregated gene effects for heterogeneity between men and women, we observed significance (FDR $Q < 0.05$) sex-

difference in genetic effects of five genes on metabolites in AAs, however, none of them was successfully replicated in EAs (Appendix A- Supplemental Table 5).

Discussion

We performed sex-stratified exome-wide association analyses for 271 GC-MS/LC-MS measured named metabolites in ARIC AAs, and identified a novel common variant rs11555566 in the *ADA* gene associated with N1-methyladenosine levels in both men and women. The association between rs11555566 and N1-methyladenosine levels was successfully replicated in independent samples of ARIC EA men and women, respectively. In AAs, we observed variants in 6 genes using common single variant tests or burden tests suggesting differed genetic effects on metabolite levels through testing for difference of effect sizes in sex-stratified results, although the difference between sexes was not shown to be consistent in an independent sample of EAs.

In total, we identified and replicated variants in 12 genes through either common single variant analysis or gene-based burden test in sex-stratified analyses. Eleven of them were previously reported genes known to be associated with one or more metabolites in non-sex-stratified genetic association studies (16, 19, 21, 22). Variants in 6 genes reached FDR-corrected exome-wide significance for testing the difference of effects between men and women but we failed to replicate the sex differences in EAs. A previously reported sex-specific genetic effect of a non-coding variant in *CPSI* associated with glycine (26) was not

observed in our data. One difference between the analysis presented by Mittelstrass et al. (26) and that presented here is we focused on only coding functional variants in our analyses.

In the results presented here, we identified a novel variant rs11555566 in the *ADA* gene to be associated with N1-methyladenosine. *ADA* encodes the enzyme, adenosine deaminase that catalyzes the hydrolysis of adenosine to inosine and plays a critical role in purine metabolism and adenosine homeostasis (43, 44). The metabolite we observed to be associated with rs11555566, N1-methyladenosine, is one of the modified nucleosides that contains adenosine as its core base. N1-methyladenosine modification regulates transfer ribonucleic acid (tRNA) and messenger RNA (mRNA) stability (45, 46), and impacts a wide array of gene expression (47). Although *ADA*'s primary function is developing and maintaining the immune system in human (48), the metabolic basis and full physiological role of *ADA* is not completely understood. Evidence has been reported for a role of *ADA* in male fertility (49, 50). Our results suggest a slightly larger effect in men ($\hat{\beta}=0.18, 0.22$ in AAs and EAs, respectively) as compared to women ($\hat{\beta}=0.14, 0.17$ in AAs and EAs, respectively).

There may be lack of consistency in testing sex differences of genetic effects between AAs and EAs due to: 1) the genetic architecture of the serum metabolome is consistent between men and women, and/or- 2) a lack of statistical power to detect small differences in genetic effects between men and women. Sex-stratified analyses followed by a z test testing for difference of effect sizes may not be powerful enough to test sex difference in genetic

effects on metabolome, particularly for rare and low-frequency genetic variants. Enhanced statistical methods and tools with sufficient power and flexibility for testing heterogeneity of genetics effects are needed. Finally, our study's design with discovery in one race group and replication in another may not be ideal. The discovery sample for this study was AAs, a population with high level of genetic diversity to promote novel findings (51). However, the replication sample was EAs. Rare variants aggregated in genes may differ between the two races, and ancestry-specific rare variants may contribute to sex-specific effects on metabolites, which will not be consistent between races and missed in our analyses.

In summary, we identified a novel variant in *ADA* associated with N1-methyladenosine levels in both race groups, suggesting that sex-specific genetic effects of metabolites may exist. The lack of consistency in testing sex differences of the genetic effects between discovery and replication samples underscores that enhanced statistical methods and tools are warranted for further sex-specific effect studies.

References

1. Villas-Bôas SG, Mas S, Åkesson M, Smedsgaard J, Nielsen J. Mass spectrometry in metabolome analysis. *Mass spectrometry reviews*. 2005;24(5):613-46.
2. Floegel A, Stefan N, Yu Z, Mühlenbruch K, Drogan D, Joost H-G, et al. Identification of Serum Metabolites Associated With Risk of Type 2 Diabetes Using a Targeted Metabolomic Approach. *Diabetes*. 2013;62(2):639-48.
3. Zheng Y, Yu B, Alexander D, Mosley TH, Heiss G, Nettleton JA, et al. Metabolomics and Incident Hypertension Among Blacks. The Atherosclerosis Risk in Communities Study. 2013.
4. Menni C, Graham D, Kastenmüller G, Alharbi NHJ, Alsanosi SM, McBride M, et al. Metabolomic Identification of a Novel Pathway of Blood Pressure Regulation Involving Hexadecanedioate. *Hypertension*. 2015;66(2):422-9.
5. Wang TJ, Larson MG, Vasan RS, Cheng S, Rhee EP, McCabe E, et al. Metabolite profiles and the risk of developing diabetes. *Nature medicine*. 2011;17(4):448-53.
6. Shah SH, Sun J-L, Stevens RD, Bain JR, Muehlbauer MJ, Pieper KS, et al. Baseline metabolomic profiles predict cardiovascular events in patients at risk for coronary artery disease. *American Heart Journal*. 2012;163(5):844-50.e1.
7. Wang Z, Klipfell E, Bennett BJ, Koeth R, Levison BS, DuGar B, et al. Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. *Nature*. 2011;472(7341):57-63.

8. Zheng Y, Yu B, Alexander D, Manolio TA, Aguilar D, Coresh J, et al. Associations between metabolomic compounds and incident heart failure among African Americans: the ARIC Study. *American journal of epidemiology*. 2013:kwt004.
9. Tai ES, Tan MLS, Stevens RD, Low YL, Muehlbauer MJ, Goh DLM, et al. Insulin resistance is associated with a metabolic profile of altered protein metabolism in Chinese and Asian-Indian men. *Diabetologia*. 2010;53(4):757-67.
10. Rizza S, Copetti M, Rossi C, Cianfarani MA, Zucchelli M, Luzi A, et al. Metabolomics signature improves the prediction of cardiovascular events in elderly subjects. *Atherosclerosis*. 2014;232(2):260-4.
11. Ganna A, Salihovic S, Sundström J, Broeckling CD, Hedman ÅK, Magnusson PKE, et al. Large-scale Metabolomic Profiling Identifies Novel Biomarkers for Incident Coronary Heart Disease. *PLOS Genetics*. 2014;10(12):e1004801.
12. Vaarhorst AA, Verhoeven A, Weller CM, Bohringer S, Goralier S, Meissner A, et al. A metabolomic profile is associated with the risk of incident coronary heart disease. *Am Heart J*. 2014;168(1):45-52.e7.
13. Wurtz P, Havulinna AS, Soininen P, Tynkkynen T, Prieto-Merino D, Tillin T, et al. Metabolite profiling and cardiovascular event risk: a prospective study of 3 population-based cohorts. *Circulation*. 2015;131(9):774-85.
14. Paynter NP, Balasubramanian R, Giulianini F, Wang DD, Tinker LF, Gopal S, et al. Metabolic Predictors of Incident Coronary Heart Disease in Women. *Circulation*. 2018;137(8):841-53.

15. Gieger C, Geistlinger L, Altmaier E, De Angelis MH, Kronenberg F, Meitinger T, et al. Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. *PLoS Genet.* 2008;4(11):e1000282.
16. Shin S-Y, Fauman EB, Petersen A-K, Krumsiek J, Santos R, Huang J, et al. An atlas of genetic influences on human blood metabolites. *Nat Genet.* 2014;46(6):543-50.
17. Suhre K, Shin S-Y, Petersen A-K, Mohnney RP, Meredith D, Wägele B, et al. Human metabolic individuality in biomedical and pharmaceutical research. *Nature.* 2011;477(7362):10.1038/nature10354.
18. Krumsiek J, Suhre K, Evans AM, Mitchell MW, Mohnney RP, Milburn MV, et al. Mining the Unknown: A Systems Approach to Metabolite Identification Combining Genetic and Metabolic Information. *PLoS Genetics.* 2012;8(10):e1003005.
19. Draisma HHM, Pool R, Kobl M, Jansen R, Petersen A-K, Vaarhorst AAM, et al. Genome-wide association study identifies novel genetic variants contributing to variation in blood metabolite levels. *Nature communications.* 2015;6:7208-.
20. Ried JS, Shin S-Y, Krumsiek J, Illig T, Theis FJ, Spector TD, et al. Novel genetic associations with serum level metabolites identified by phenotype set enrichment analyses. *Human Molecular Genetics.* 2014;23(21):5847-57.
21. Rhee EP, Ho JE, Chen M-H, Shen D, Cheng S, Larson MG, et al. A Genome-Wide Association Study of the Human Metabolome in a Community-Based Cohort. *Cell metabolism.* 2013;18(1):130-43.

22. Rhee EP, Yang Q, Yu B, Liu X, Cheng S, Deik A, et al. An exome array study of the plasma metabolome. *Nature Communications*. 2016;7:12360.
23. Yu B, Li AH, Metcalf GA, Muzny DM, Morrison AC, White S, et al. Loss-of-function variants influence the human serum metabolome. *Science Advances*. 2016;2(8):e1600800.
24. Long T, Hicks M, Yu H-C, Biggs WH, Kirkness EF, Menni C, et al. Whole-genome sequencing identifies common-to-rare variants associated with human blood metabolites. *Nat Genet*. 2017;49(4):568-78.
25. Yu B, de Vries PS, Metcalf GA, Wang Z, Feofanova EV, Liu X, et al. Whole genome sequence analysis of serum amino acid levels. *Genome Biology*. 2016;17(1):237.
26. Mittelstrass K, Ried JS, Yu Z, Krumsiek J, Gieger C, Prehn C, et al. Discovery of Sexual Dimorphisms in Metabolic and Genetic Biomarkers. *PLOS Genetics*. 2011;7(8):e1002215.
27. Krumsiek J, Mittelstrass K, Do KT, Stücker F, Ried J, Adamski J, et al. Gender-specific pathway differences in the human serum metabolome. *Metabolomics*. 2015;11(6):1815-33.
28. Hartiala JA, Wilson Tang WH, Wang Z, Crow AL, Stewart AFR, Roberts R, et al. Genome-wide association study and targeted metabolomics identifies sex-specific association of CPS1 with coronary artery disease. *Nature Communications*. 2016;7:10558.

29. Klein MS, Connors KE, Shearer J, Vogel HJ, Hittel DS. Metabolomics Reveals the Sex-Specific Effects of the SORT1 Low-Density Lipoprotein Cholesterol Locus in Healthy Young Adults. *Journal of Proteome Research*. 2014;13(11):5063-70.
30. ARIC Investigators. The Atherosclerosis Risk in Communities (ARIC) study: design and objectives. *American journal of epidemiology*. 1989;129(4):687-702.
31. Evans AM, DeHaven CD, Barrett T, Mitchell M, Milgram E. Integrated, nontargeted ultrahigh performance liquid chromatography/electrospray ionization tandem mass spectrometry platform for the identification and relative quantification of the small-molecule complement of biological systems. *Anal Chem*. 2009;81(16):6656-67.
32. Ohta T, Masutomi N, Tsutsui N, Sakairi T, Mitchell M, Milburn MV, et al. Untargeted metabolomic profiling as an evaluative tool of fenofibrate-induced toxicology in Fischer 344 male rats. *Toxicologic pathology*. 2009;37(4):521-35.
33. Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*. 2012;28(1):112-8.
34. Breiman L. Random forests. *Machine learning*. 2001;45(1):5-32.
35. Grove ML, Yu B, Cochran BJ, Haritunians T, Bis JC, Taylor KD, et al. Best Practices and Joint Calling of the HumanExome BeadChip: The CHARGE Consortium. *PLoS ONE*. 2013;8(7):e68095.
36. Liu X, White S, Peng B, Johnson AD, Brody JA, Li AH, et al. WGS: an annotation pipeline for human genome sequencing studies. *Journal of medical genetics*. 2015;jmedgenet-2015-103423.

37. Liu X, Jian X, Boerwinkle E. dbNSFP v2. 0: a database of human non-synonymous SNVs and their functional predictions and annotations. *Human mutation*. 2013;34(9):E2393-E402.
38. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research*. 2010;38(16):e164-e.
39. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet*. 2008;83(3):311-21.
40. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*. 2006;38:904.
41. Paternoster R, Brame R, Mazerolle P, Piquero A. Using the correct statistical test for the equality of regression coefficients. *Criminology*. 1998;36(4):859-66.
42. Lin X, Lee S, Wu MC, Wang C, Chen H, Li Z, et al. Test for Rare Variants by Environment Interactions in Sequencing Association Studies. *Biometrics*. 2016;72(1):156-64.
43. Lindley ER, Pisoni RL. Demonstration of adenosine deaminase activity in human fibroblast lysosomes. *Biochem J*. 1993;290 (Pt 2):457-62.
44. Eltzschig HK, Faigle M, Knapp S, Karhausen J, Ibla J, Rosenberger P, et al. Endothelial catabolism of extracellular adenosine during hypoxia: the role of surface adenosine deaminase and CD26. *Blood*. 2006;108(5):1602-10.

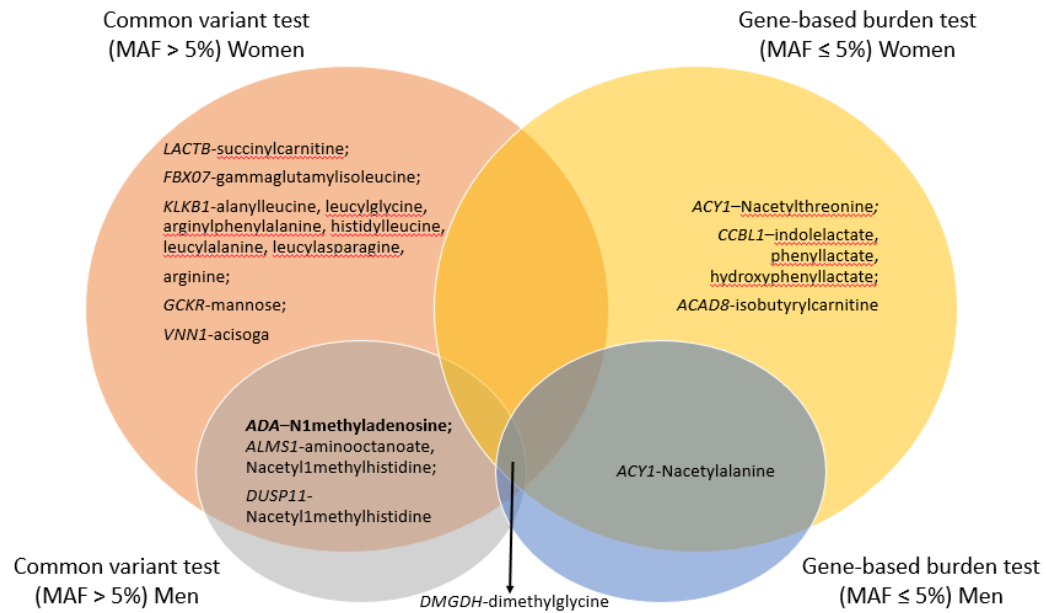
45. Dominissini D, Nachtergaele S, Moshitch-Moshkovitz S, Peer E, Kol N, Ben-Haim MS, et al. The dynamic N1-methyladenosine methylome in eukaryotic messenger RNA. *Nature*. 2016;530:441.
46. Zhang C, Jia G. Reversible RNA Modification N1-methyladenosine (m1A) in mRNA and tRNA. *Genomics, Proteomics & Bioinformatics*. 2018;16(3):155-61.
47. Sharma S, Hartmann JD, Watzinger P, Klepper A, Peifer C, Kötter P, et al. A single N1-methyladenosine on the large ribosomal subunit rRNA impacts locally its structure and the translation of key metabolic enzymes. *Scientific Reports*. 2018;8(1):11904.
48. Blackburn MR, Kellems RE. Adenosine Deaminase Deficiency: Metabolic Basis of Immune Deficiency and Pulmonary Inflammation. In: Alt FW, editor. *Advances in Immunology*. 86: Academic Press; 2005. p. 1-41.
49. Rostampour F, Biglari M, Vaisi-Raygani A, Salimi S, Tavailani H. Adenosine deaminase activity in fertile and infertile men. *Andrologia*. 2012;44 Suppl 1:586-9.
50. Fattahi A, Khodadadi I, Amiri I, Latifi Z, Ghorbani M, Tavailani H. The Role of G22 A Adenosine Deaminase 1 Gene Polymorphism and the Activities of ADA Isoenzymes in Fertile and Infertile Men. *Urology*. 2015;86(4):730-4.
51. Tishkoff SA, Williams SM. Genetic analysis of African populations: human evolution and complex disease. *Nature Reviews Genetics*. 2002;3:611.

Table II-1 Baseline characteristics of analyzed participants in the ARIC study

	Women		Men	
	AAs	EAs	AAs	EAs
N	1292	827	720	701
Age (years)	52.9	54.3	53.0	54.9
	(5.6)	(5.8)	(5.8)	(5.8)
eGFR	105.5	92.0	101.2	90.5
(mL/min/1.73 m²)	(18.6)	(14.9)	(17.8)	(14.2)

AAs: African-Americans, EAs: European-Americans, eGFR: estimated glomerular filtration rate

Figure II-1 Significantly identified and replicated gene-metabolite pairs revealed by sex-stratified exome-wide association studies.



**CHAPTER III. RARE AND LOW-FREQUENCY GENETIC VARIANT \times SEX
INTERACTIONS IDENTIFY NOVEL LOCI INFLUENCING THE SERUM
METABOLOME IN THE ATHEROSCLEROSIS RISK IN COMMUNITIES (ARIC)
STUDY**

Abstract

Metabolomic profiling and the integration of genomic data have proven to be powerful tools to investigate genetic effects underlying intermediate phenotype levels such as metabolites and may facilitate improved understanding of pathophysiologic processes of disease. However, most published studies did not consider sex as an effect modifier nor gene by sex interactions. The present study investigated rare and low-frequency genetic variants (minor allele frequency $\leq 5\%$) and sex interactions on serum metabolite levels and evaluated the joint effects of genetic main effects and gene-sex interactions. Chromatography-mass spectrometry measured metabolites and the Illumina HumanExome BeadChip genotyped exonic variants were analyzed in 2,012 African-Americans from the Atherosclerosis Risk in Communities (ARIC) study. Using gene-based rareGE and MiSTi approaches, we conducted exome-wide gene-sex interaction tests, and a joint analysis of genetic main and gene-sex interaction effects. Rare and low-frequency functional variants, (i.e. frameshift, nonsynonymous, stop/gain, stop/loss, and splicing) were aggregated by genes. Exome-wide significant genes (false discovery rate $\leq 5\%$) were evaluated for replication in an independent sample of 1,528 ARIC European-Americans. In total, we observed and replicated 14 gene-metabolite associations through the joint test, 3 of which were novel, including *PLA2G7*-arachidonate (20:4n6), *PTER*- N-acetyl-beta-alanine and *NPC2*- leucylserine. The *NPC2*-leucylserine association arose from both genetic main effects and gene-sex interaction effects, as the interaction test using rareGE for *NPC2*-sex interaction on leucylserine levels reached nominal significance level ($p = 3.79 \times 10^{-04}$). In conclusion, this study applied

emerging statistical approaches to investigate the role of rare and low-frequency genetic variants and gene-sex interactions, and successfully identified novel genes associated with metabolites.

Introduction

Metabolomics is a scientific approach that systematically evaluates small-molecule metabolites in biologic samples that reflect the state of the system or whole organism and may provide additional insights into disease pathology (1-3). Both traditional genome-wide association studies (GWAS) and sequencing analyses across the exome or whole genome have successfully identified and verified hundreds of genetic loci associated with the levels of metabolites (4-16), and many of them can be further related to complex diseases or clinically relevant risk factors of disease development (12, 17).

Sex-specific differences in metabolite patterns in healthy human have been reported in urine and plasma (18, 19), which suggests that sex should be considered further in metabolomic studies. Limited work has identified sex-specific metabolism-related genetic polymorphisms through sex-stratified GWAS and sex-specific pathway differences in the serum metabolome (20, 21). Additional systematic studies are needed to better understand the modifying effect of sex on the human metabolome and its genetic determinants with a particular focus on rare and low frequency ($MAF \leq 5\%$) variants. Rare and low frequency variants make up the vast majority of the genetic variation in the genome (22), and may account for part of the missing heritability along with gene-environmental ($G \times E$) interactions (23).

Unlike well-established G×E interaction tests for common variants (24, 25), methods development for detecting rare variant G×E interactions is challenging because of relatively low power for a single marker test with $MAF \leq 5\%$, as well as inflated type 1 error rates and biased effect estimates for conventional burden tests (26). Emerging methods have been proposed to overcome these challenges. Su et al. proposed a novel and rigorous framework, Mixed effects Score Tests for interaction (MiSTi), to derive independent score statistics for fixed effects and the variance component that is more powerful to test G×E interaction terms of rare variants (27). A joint test that allows one to simultaneously test genetic main effects and interaction effects was proposed and successfully implemented by Chen and colleagues in the R package ‘rareGE’ (28).

To date, there is no study systematically utilizing methods developed for testing rare variant G×E interactions in the setting of large-scale metabolomic data. In addition, an investigation of gene-sex interactions or sex-specific genetic variants related to metabolites has not been conducted in African-Americans (AAs). Therefore, in this study, we leveraged existing data from a large population-based multi-ancestry cohort, the Atherosclerosis Risk in Communities (ARIC) study that contains well-characterized AAs and European-Americans (EAs), to identify novel genetic loci influencing the serum metabolome that were not identified when considering the genetic main effect alone.

Methods

Study Sample

The ARIC study is a population-based prospective cohort study of 15,792 adults from four U.S. communities (Forsyth County, NC; Jackson, MS; suburbs of Minneapolis, MN; and Washington County, MD), which has been described in detail previously (29). ARIC included both EAs and AAs aged 45-64 at the baseline examination (1987-1989).

Participants completed three additional triennial follow-up examinations, a fifth exam in 2011-2013, and a sixth exam in 2016-2017. There were 3,540 participants with complete metabolite measurements and exome chip genotyped data at the baseline examination. The ARIC study has been approved by the institutional review boards at each site, and written informed consent was obtained from all participating individuals.

Measurements of Metabolites

Metabolite profiling was completed in 2010 (batch 1) and 2014 (batch 2) using fasting serum samples that had been stored at -80° since collection at the baseline examination. Batch 1 were all AAs and Batch 2 included both AAs (24.8%) and EAs (75.2%). In total, 602 metabolites were detected and semi-quantified by Metabolon (Durham, USA) using untargeted, gas- and liquid- chromatography-mass spectrometry (GC-MS and LC-MS)-based protocols (30, 31). To evaluate batch effects, a set of 97 samples were measured in both the 2010 and 2014 batches. There were 384 named metabolites that were

identified to be present in both batches and these metabolites will be used for this thesis research.

In the present study, sample-level quality control was performed to remove individuals with missing values for more than 40% of the measured metabolites (1 sample was removed from batch 2). After sample-level quality control, metabolomic profiles were available in 2,479 AAs and 1,553 EAs. Exclusion criteria for metabolites includes: 1) six-there metabolites were excluded as more than 40% of the samples have missing values or values below the detection limit (BDL) within each batch; and 2) fifty metabolites were excluded as the Pearson correlation coefficient (r) between 2010 and 2014 measurements on the same stored sample (at least 46 out of the 97 pairs) is less than 0.30. After exclusions, this study was based on an evaluation of 271 named metabolites. Metabolite levels were analyzed as continuous variables, where missing/BDL values were imputed using random forest imputation based on the remaining observed measurements (32, 33).

Genotypes

Genotyping was performed with the Illumina HumanExome BeadChip v1.0 (“exome chip”) querying 247,870 single nucleotide variants (SNVs) at the baseline examination in 11,071 EAs and 2,953 AAs in the ARIC study. The exome chip data was selected rather than exome or whole genome sequence data that is also available in ARIC because using exome chip data maximizes the available sample size for this analysis. To improve accurate calling of rare variants, genotyped data from ARIC along with 10 other studies from the Cohorts for

Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium were pulled together for joint calling, details were described elsewhere (34). A total of 8,994 variants were excluded after laboratory quality control steps, for instance call rate <95%, Hardy-Weinberg equilibrium test P value (pHWE) < 1×10^{-6} , and poorly clustering variants (34). SNVs with missing rate >5% were removed from analysis.

Exome chip variant annotation was completed using the Whole Genome Sequencing Annotation (WGS) pipeline v055 (35), including dbNSFP v2.9 (36). Functional variants and genes were determined using ANNOVAR (37) according to the reference genome GRCh37/hg19 and National Center for Biotechnology Information RefSeq.

Statistical Analysis:

Prior to the analyses, metabolite levels were winsorized (99%) within each batch. Each metabolite was investigated for its goodness-of-fit to normality. Due to right-skewed distributions of many metabolite levels, the natural log transformation was applied to most metabolites prior to analyses. For metabolites that are still not normally distributed, a rank based inverse normal transformation was used. The transformation methods applied to each metabolite were provided in Appendix B- Supplemental Table 1.

Two analyses were conducted for each metabolite within each race group: 1) a joint analysis of genetic main effects and G×E interaction effects and 2) a G×E interaction term test only. The joint analysis was conducted using rareGE (28), and the interaction term test was conducted using both rareGE and MiSTi (27, 28). The unit-of-analysis is an annotated

gene. All annotated coding variants, such as splicing, stop-gain, stop-loss, nonsynonymous, and indels within the gene were aggregated for the analysis. Additionally, genes with cumulative minor allele counts ≤ 6 in each race were excluded. Models were adjusted for age and population substructure using the first three ancestry specific principal components (PCs) (38), with additional adjustment of estimated glomerular filtration rate (eGFR) and batch effects for metabolites.

We used AAs as our discovery sample and conducted replication in EAs. Using a false discovery rate (FDR) to correct for number of genes tested while considering the 271 metabolites, we defined exome-wide significant genes as those with FDR $Q \leq 5\%$ in discovery AAs; these genes were pursued for replication analyses in EAs. Replication was defined as those genes with FDR $Q \leq 5\%$, corrected for the number of genes taken forward to evaluate in EAs. All statistical analyses were conducted in R version 3.4 (R Foundation for Statistical Computing, Vienna, Austria).

Results

There was a total of 3,540 participants with measured metabolite levels and genotyped exome chip data in the ARIC study, including 2,012 AAs for discovery and 1,528 EAs for replication. Women comprised more than half of the samples in both race groups (64.2% in AAs, 54.1% in EAs). In general, the average baseline age of AAs and EAs was comparable (AAs vs. EAs: 53.0 ± 5.7 vs. 54.6 ± 5.8 years), and AA participants tends to have slightly higher levels of eGFR (AAs vs. EAs: 104.0 ± 18.3 vs. 91.4 ± 14.6 mL/min/1.73 m²).

In AAs, we observed 48 gene-metabolite associations ($FDR\ Q \leq 5\%$) harboring rare and low-frequency variants through the joint analysis using rareGE (Appendix B-Supplemental Table 2). Distributions of QQ plots for joint test are shown in Appendix B-Supplemental Figure 1. Among them, 38 gene-metabolite associations were available in EAs, and these were taken forward for replication. In total, 14 gene-metabolite associations were successfully replicated using the joint test, 3 of which, namely *PLA2G7*- arachidonate (20:4n6), *PTER*- N-acetyl-beta-alanine and *NPC2*- leucylserine, were novel associations (**Table III-1**). The interaction term only test using rareGE for *NPC2*-sex interaction on leucylserine levels reached nominal significant ($p = 3.79 \times 10^{-04}$ -- 5.83×10^{-04} using random/fix effect interaction models (**Table III-1**), suggesting that gene-sex interaction effects contribute to the identified association between *NPC2* and leucylserine levels. We additionally tested the marginal genetic main effect of these 14 gene-metabolite associations using SKAT test. The marginal genetic main effect of *NPC2* on leucylserine levels reached nominal significance level ($p = 1.07 \times 10^{-03}$), which also suggesting that both the genetic main effects and the gene-sex interaction effects contribute to the identified joint effect. Genetic main effect for the rest of the genes showed similar p -values as results of the joint test (**Table III-1**). The rest of the identified gene-metabolite associations were mainly driven by genetic main effects (rareGE interaction test $p > 0.05$, **Table III-1**)

Tests that focused on gene-sex interaction terms alone failed to identify any genes that passed the FDR corrected significance threshold using either rareGE or MiSTi. Using $FDR\ Q < 0.2$ as a suggestive significance threshold, six gene-metabolite pairs, 4 through

MiSTi and 3 through rareGE with 1 overlapping, were identified to be suggestive (Appendix B- Supplemental Table 3-4). Half of the genes had valid interaction test results in EAs, but the results were not replicated ($p > 0.05$ in EAs, Appendix B- Supplemental Table 3-4).

Discussion

To our knowledge, this is the first study to evaluate the role of rare and low frequency variants in gene-sex interactions and joint effects of genetic main and gene-sex interaction on metabolite levels. In total, we observed and replicated 14 gene-metabolite associations through the joint test, 3 of which were novel, including *PLA2G7*- arachidonate (20:4n6), *PTER*- N-acetyl-beta-alanine and *NPC2*- leucylserine. No significant novel loci were detected via analyzing the gene-sex interactions alone.

Eleven of the 14 identified gene-metabolite pairs, comprised of 6 genes: *ALMS1*, *ACY1*, *KLKB1*, *DMGDH*, *CCBL1*, *ACAD8*, *SLC25A45*, and *HAL*, have been previously identified through either traditional GWAS or sequence-based genetic association studies that only considered genetic main effects (5, 6, 12-16). For example, rare loss-of-function variants in *HAL*, a gene that encodes histidine ammonia-lyase in the first step of histidine catabolism, was reported to be associated with increased histidine levels, and further linked to reduced incidence CHD risk (39). For these 6 known genes influencing metabolite levels, no evidence of gene-sex interactions were observed, suggesting that these gene-metabolite associations were not modified by sex.

Among the novel genes we identified through the joint test, *PLA2G7* (*Phospholipase A2 Group VII*) was observed to be associated with arachidonic acid (20:4n6) in our data. Previous studies have reported several mutations in *PLA2G7* associated with lipoprotein-associated phospholipase A2 (Lp-PLA₂) activity and mass, both positively and negatively (40-43). A large meta-analysis including 32 prospective studies by Thompson et al (44) showed that a reduction in Lp-PLA₂ activity/mass was associated with reductions in risks of coronary heart disease and ischemic stroke. In addition, Lp-PLA₂ activity has been recently approved by the FDA for routine clinical use to predict coronary heart disease events especially for black women (45). Phospholipase A2 (PLA₂) catalyzes the hydrolysis of the sn-2 position of membrane glycerophospholipids to liberate arachidonic acid (46), the metabolite we observed and known for mediating inflammation (47). Therefore, the association we observed between *PLA2G7* and arachidonic acid (20:4n6) is expected, and helps understanding the path of genetic variation in *PLA2G7* to vascular inflammation and CVD.

We also observed a novel association between *PTER* and N-acetyl-beta-alanine levels using the joint test. Previously GWAS has identified variants *near PTER* (*phosphotriesterase-related*) as a locus for obesity in European populations (48). The metabolite, N-acetyl-beta-alanine can be broken down to acetate and beta-alanine through hydrolysis, the latter of which forms carnosine (beta-alanyl-L-histidine), a dipeptide with anti-inflammatory, antioxidant, anti-glycation, and anti-ischaemic roles on cardiometabolic risk and disease (49). Current understanding of the protein encoded by *PTER* was limited to

hydrolase activity, acting on ester bonds. Our results provide insight into the molecular function or biological process that *PTER* may be involved, which may further link to cardiometabolic diseases.

Out of the three novel gene-metabolite associations we identified and replicated through the joint test, *NPC2*- leucylserine also showed a contribution from gene-sex interaction effects. *NPC2* (*NPC Intracellular Cholesterol Transporter 2*) encodes a protein that may function in regulating the transport of cholesterol through the late lysosomal system. In a recent genetic study of the human plasma proteome, common variants in *NPC2* gene have been associated with levels of a blood protein, Cathepsin H (50), which is important in the overall degradation of lysosomal proteins. The *NPC2* and *NPC2*-sex interaction associated metabolite in our data, leucylserine, is a dipeptide composed of leucine and serine. It is an incomplete breakdown product of protein digestion, which can be produced during lysosomal proteolysis. Although there is a lack of understanding the role of the *NPC2*-sex interaction on leucylserine levels, an animal study showed that in the ovary, *NPC2* was restricted to steroidogenic cells that use cholesterol to produce hormones, and reported female infertility in *NPC2* deficient mice (51). It is possible that sex hormones, for example estradiol, may be further related to lysosomal function (52) and are involved in protein catabolism that produces leucylserine.

This study has several strengths. We used an emerging statistical approach that jointly tests the genetic main effect and gene-sex interaction on the human metabolome. Previous studies have shown that inclusion of G×E interactions may be important for identifying novel

signals, particularly for rare and low-frequency variants (53). Results in the present study support this conclusion by showing that novel genes were identified through this joint approach. Our study focused on the human metabolome, an intermediate phenotype that known to have larger genetic effects than clinical end points (5, 54), which are suitable to promote novel gene discoveries in the context of rare variant G×E interactions. Another strength of the present study is the joint calling of variants in a large pooled sample of studies conducted in the same laboratory, including the ARIC study. By increasing the sample size during the calling of variants, the ability to correctly call rare variants is enhanced (55).

We successfully identified novel gene-metabolite associations, but the test that focuses on gene-sex interactions alone fail to reveal significant results for several reasons. First, sex may not modify the genetic effects on human metabolome, in other words, genetic architecture of the serum metabolome is largely consistent between men and women. Second, the study's design having discovery in one race group and replication in another may not be ideal. The discovery sample for this study was AAs, a population with high level of genetic diversity to promote novel findings (56). However, the replication sample was EAs. Rare variants aggregated in genes may differ between two races, and ancestry-specific rare variants may contribute to sex-specific effects on metabolites, which will not be consistent between races and missed in our analyses. Finally, although we applied newly developed statistical methods that were known to have improved performance in testing G×E interactions (27, 28), future studies that focus on rare and low-frequency variants to identify

novel loci and G×E interactions may require much larger sample sizes than were available in just the ARIC study.

In conclusion, this study applied emerging statistical approaches to investigate the role of rare and low-frequency variants in gene-sex interactions on the human metabolome, and successfully identified 3 novel genes associated with metabolites. Our results show promise for other larger scale studies analyzing rare variant G×E interactions to reveal novel biology.

References

1. Evans AM, DeHaven CD, Barrett T, Mitchell M, Milgram E. Integrated, nontargeted ultrahigh performance liquid chromatography/electrospray ionization tandem mass spectrometry platform for the identification and relative quantification of the small-molecule complement of biological systems. *Analytical chemistry*. 2009;81(16):6656-67.
2. Madsen R, Lundstedt T, Trygg J. Chemometrics in metabolomics—a review in human disease diagnosis. *Analitica chimica acta*. 2010;659(1-2):23-33.
3. Villas-Bôas SG, Mas S, Åkesson M, Smedsgaard J, Nielsen J. Mass spectrometry in metabolome analysis. *Mass spectrometry reviews*. 2005;24(5):613-46.
4. Gieger C, Geistlinger L, Altmaier E, De Angelis MH, Kronenberg F, Meitinger T, et al. Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. *PLoS Genet*. 2008;4(11):e1000282.
5. Shin S-Y, Fauman EB, Petersen A-K, Krumsiek J, Santos R, Huang J, et al. An atlas of genetic influences on human blood metabolites. *Nat Genet*. 2014;46(6):543-50.
6. Suhre K, Shin S-Y, Petersen A-K, Mohnhey RP, Meredith D, Wägele B, et al. Human metabolic individuality in biomedical and pharmaceutical research. *Nature*. 2011;477(7362):10.1038/nature10354.
7. Krumsiek J, Suhre K, Evans AM, Mitchell MW, Mohnhey RP, Milburn MV, et al. Mining the Unknown: A Systems Approach to Metabolite Identification Combining Genetic and Metabolic Information. *PLoS Genetics*. 2012;8(10):e1003005.

8. Draisma HHM, Pool R, Kobl M, Jansen R, Petersen A-K, Vaarhorst AAM, et al. Genome-wide association study identifies novel genetic variants contributing to variation in blood metabolite levels. *Nature communications*. 2015;6:7208-.
9. Ried JS, Shin S-Y, Krumsiek J, Illig T, Theis FJ, Spector TD, et al. Novel genetic associations with serum level metabolites identified by phenotype set enrichment analyses. *Human Molecular Genetics*. 2014;23(21):5847-57.
10. Rhee EP, Ho JE, Chen M-H, Shen D, Cheng S, Larson MG, et al. A Genome-Wide Association Study of the Human Metabolome in a Community-Based Cohort. *Cell metabolism*. 2013;18(1):130-43.
11. Rhee EP, Yang Q, Yu B, Liu X, Cheng S, Deik A, et al. An exome array study of the plasma metabolome. *Nature Communications*. 2016;7:12360.
12. Yu B, Li AH, Metcalf GA, Muzny DM, Morrison AC, White S, et al. Loss-of-function variants influence the human serum metabolome. *Science Advances*. 2016;2(8):e1600800.
13. Long T, Hicks M, Yu H-C, Biggs WH, Kirkness EF, Menni C, et al. Whole-genome sequencing identifies common-to-rare variants associated with human blood metabolites. *Nat Genet*. 2017;49(4):568-78.
14. Yu B, de Vries PS, Metcalf GA, Wang Z, Feofanova EV, Liu X, et al. Whole genome sequence analysis of serum amino acid levels. *Genome Biology*. 2016;17(1):237.

15. Yu B, Zheng Y, Alexander D, Morrison AC, Coresh J, Boerwinkle E. Genetic determinants influencing human serum metabolome among African Americans. *PLoS Genet.* 2014;10.
16. Feofanova EV, Yu B, Metcalf GA, Liu X, Muzny D, Below JE, et al. Sequence-Based Analysis of Lipid-Related Metabolites in a Multiethnic Study. *Genetics.* 2018;209(2):607-16.
17. Menni C, Graham D, Kastenmüller G, Alharbi NHJ, Alsanosi SM, McBride M, et al. Metabolomic identification of a novel pathway of blood pressure regulation involving hexadecanedioate. *Hypertension.* 2015;66(2):422-9.
18. Fan S, Yeon A, Shahid M, Anger JT, Eilber KS, Fiehn O, et al. Sex-associated differences in baseline urinary metabolites of healthy adults. *Scientific Reports.* 2018;8(1):11883.
19. Rist MJ, Roth A, Frommherz L, Weinert CH, Krüger R, Merz B, et al. Metabolite patterns predicting sex and age in participants of the Karlsruhe Metabolomics and Nutrition (KarMeN) study. *PLOS ONE.* 2017;12(8):e0183228.
20. Mittelstrass K, Ried JS, Yu Z, Krumsiek J, Gieger C, Prehn C, et al. Discovery of Sexual Dimorphisms in Metabolic and Genetic Biomarkers. *PLOS Genetics.* 2011;7(8):e1002215.
21. Krumsiek J, Mittelstrass K, Do KT, Stücker F, Ried J, Adamski J, et al. Gender-specific pathway differences in the human serum metabolome. *Metabolomics.* 2015;11(6):1815-33.

22. Gorlov IP, Gorlova OY, Frazier ML, Spitz MR, Amos CI. Evolutionary evidence of the effect of rare variants on disease etiology. *Clinical genetics*. 2011;79(3):199-206.
23. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. *Nature*. 2009;461(7265):747-53.
24. Kraft P, Yen Y-C, Stram DO, Morrison J, Gauderman WJ. Exploiting gene-environment interaction to detect genetic associations. *Human heredity*. 2007;63(2):111-9.
25. Manning AK, LaValley M, Liu C-T, Rice K, An P, Liu Y, et al. Meta-analysis of gene-environment interaction: joint estimation of SNP and SNP \times environment regression coefficients. *Genetic Epidemiology*. 2011;35(1):11-8.
26. Lin X, Lee S, Wu MC, Wang C, Chen H, Li Z, et al. Test for Rare Variants by Environment Interactions in Sequencing Association Studies. *Biometrics*. 2016;72(1):156-64.
27. Su Y-R, Di C-Z, Hsu L. A unified powerful set-based test for sequencing data analysis of GxE interactions. *Biostatistics*. 2017;18(1):119-31.
28. Chen H, Meigs JB, Dupuis J. Incorporating Gene-Environment Interaction in Testing for Association with Rare Genetic Variants. *Human Heredity*. 2014;78(2):81-90.
29. ARIC Investigators. The Atherosclerosis Risk in Communities (ARIC) study: design and objectives. *American journal of epidemiology*. 1989;129(4):687-702.

30. Evans AM, DeHaven CD, Barrett T, Mitchell M, Milgram E. Integrated, nontargeted ultrahigh performance liquid chromatography/electrospray ionization tandem mass spectrometry platform for the identification and relative quantification of the small-molecule complement of biological systems. *Anal Chem*. 2009;81(16):6656-67.
31. Ohta T, Masutomi N, Tsutsui N, Sakairi T, Mitchell M, Milburn MV, et al. Untargeted metabolomic profiling as an evaluative tool of fenofibrate-induced toxicology in Fischer 344 male rats. *Toxicologic pathology*. 2009;37(4):521-35.
32. Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*. 2012;28(1):112-8.
33. Breiman L. Random forests. *Machine learning*. 2001;45(1):5-32.
34. Grove ML, Yu B, Cochran BJ, Haritunians T, Bis JC, Taylor KD, et al. Best Practices and Joint Calling of the HumanExome BeadChip: The CHARGE Consortium. *PLoS ONE*. 2013;8(7):e68095.
35. Liu X, White S, Peng B, Johnson AD, Brody JA, Li AH, et al. WGS: an annotation pipeline for human genome sequencing studies. *Journal of medical genetics*. 2015;jmedgenet-2015-103423.
36. Liu X, Jian X, Boerwinkle E. dbNSFP v2. 0: a database of human non-synonymous SNVs and their functional predictions and annotations. *Human mutation*. 2013;34(9):E2393-E402.
37. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research*. 2010;38(16):e164-e.

38. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*. 2006;38:904.
39. Yu B, Li AH, Muzny D, Veeraraghavan N, Vries PS, Bis JC, et al. Association of rare loss-of-function alleles in HAL, serum histidine: levels and incident coronary heart disease. *Circ Cardiovasc Genet*. 2015;8.
40. Grallert H, Dupuis J, Bis JC, Dehghan A, Barbalic M, Baumert J, et al. Eight genetic loci associated with variation in lipoprotein-associated phospholipase A2 mass and activity and coronary heart disease: meta-analysis of genome-wide association studies from five community-based studies. *European heart journal*. 2012;33(2):238-51.
41. Suchindran S, Rivedal D, Guyton JR, Milledge T, Gao X, Benjamin A, et al. Genome-wide association study of Lp-PLA(2) activity and mass in the Framingham Heart Study. *PLoS Genet*. 2010;6(4):e1000928.
42. Chu AY, Guilianini F, Grallert H, Dupuis J, Ballantyne CM, Barratt BJ, et al. Genome-wide association study evaluating lipoprotein-associated phospholipase A2 mass and activity at baseline and after rosuvastatin therapy. *Circ Cardiovasc Genet*. 2012;5(6):676-85.
43. Yeo A, Li L, Warren L, Aponte J, Fraser D, King K, et al. Pharmacogenetic meta-analysis of baseline risk factors, pharmacodynamic, efficacy and tolerability endpoints from two large global cardiovascular outcomes trials for darapladib. *PLoS One*. 2017;12(7):e0182115.

44. Thompson A, Gao P, Orfei L, Watson S, Di Angelantonio E, Kaptoge S, et al. Lipoprotein-associated phospholipase A (2) and risk of coronary disease, stroke, and mortality: collaborative analysis of 32 prospective studies. Elsevier; 2010.
45. Young K. FDA clears test to help predict coronary heart disease risk NEJM Journal Watch2014 [cited 2018 11.08]. Available from:
<https://www.jwatch.org/fw109648/2014/12/17/fda-clears-test-help-predict-coronary-heart-disease-risk>.
46. Kudo I, Murakami M. Phospholipase A2 enzymes. Prostaglandins & other lipid mediators. 2002;68-69:3-58.
47. Pompeia C, Lima T, Curi R. Arachidonic acid cytotoxicity: can arachidonic acid be a physiological mediator of cell death? Cell biochemistry and function. 2003;21(2):97-104.
48. Meyre D, Delplanque J, Chèvre J-C, Lecoœur C, Lobbens S, Gallina S, et al. Genome-wide association study for early-onset and morbid adult obesity identifies three new risk loci in European populations. Nature Genetics. 2009;41:157.
49. Baye E, Ukropcova B, Ukropec J, Hipkiss A, Aldini G, de Courten B. Physiological and therapeutic effects of carnosine on cardiometabolic risk and disease. Amino Acids. 2016;48(5):1131-49.
50. Sun BB, Maranville JC, Peters JE, Stacey D, Staley JR, Blackshaw J, et al. Genomic atlas of the human plasma proteome. Nature. 2018;558(7708):73-9.

51. Busso D, Oñate-Alvarado MJ, Balboa E, Zanolungo S, Moreno RD. Female infertility due to anovulation and defective steroidogenesis in NPC2 deficient mice. *Molecular and Cellular Endocrinology*. 2010;315(1):299-307.
52. Totta P, Pesiri V, Marino M, Acconcia F. Lysosomal function is involved in 17 β -estradiol-induced estrogen receptor α degradation and cell proliferation. *PloS one*. 2014;9(4):e94880-e.
53. Sung YJ, Winkler TW, Manning AK, Aschard H, Gudnason V, Harris TB, et al. An Empirical Comparison of Joint and Stratified Frameworks for Studying $G \times E$ Interactions: Systolic Blood Pressure and Smoking in the CHARGE Gene-Lifestyle Interactions Working Group. *Genetic Epidemiology*. 2016;40(5):404-15.
54. Groves CJ, Zeggini E, Minton J, Frayling TM, Weedon MN, Rayner NW, et al. Association analysis of 6,736 U.K. subjects provides replication and confirms TCF7L2 as a type 2 diabetes susceptibility gene with a substantial effect on individual risk. *Diabetes*. 2006;55(9):2640-4.
55. Grove ML, Yu B, Cochran BJ, Haritunians T, Bis JC, Taylor KD, et al. Best Practices and Joint Calling of the HumanExome BeadChip: The CHARGE Consortium. *PLoS One*. 2013;8(7):e68095.
56. Tishkoff SA, Williams SM. Genetic analysis of African populations: human evolution and complex disease. *Nature Reviews Genetics*. 2002;3:611.

Table III-1 Genes discovered and replicated through jointly testing the genetic main effects and gene-sex interactions in the ARIC study

African-Americans										European- Americans						
Trait	Gene	Chr	nSNP	Main.p	MiSTi.p	Fix.int.p	Ran.int.p	Joint.p	FDR- <i>Q</i>	nSNP	Main.p	MiSTi.p	Fix.int.p	Ran.int.p	Joint.p	FDR- <i>Q</i>
N-acetyl-1-methylhistidine	ALMS1	2	52	4.52×10 ⁻⁰⁹	0.76	0.99	0.98	4.05×10 ⁻⁰⁹	1.05×10 ⁻⁰³	37	2.85×10 ⁻¹¹	0.46	0.46	0.45	3.32×10 ⁻¹¹	4.21×10 ⁻¹⁰
aminooctanoate	ALMS1	2	52	2.65×10 ⁻⁰⁷	0.04	0.05	0.13	2.16×10 ⁻⁰⁷	0.02	37	4.73×10 ⁻⁰³	0.01	0.05	0.03	2.15×10 ⁻⁰³	8.17×10 ⁻⁰³
N-acetyl-alanine	ACY1	3	7	2.38×10 ⁻⁵³	0.19	0.37	0.38	4.43×10 ⁻⁵²	1.50×10 ⁻⁴⁵	3	1.75×10 ⁻²³	0.55	0.47	0.47	4.07×10 ⁻²³	1.55×10 ⁻²¹
N-acetyl-threonine	ACY1	3	7	3.94×10 ⁻¹⁶	0.14	0.27	0.27	8.01×10 ⁻¹⁶	6.77×10 ⁻¹⁰	3	1.83×10 ⁻⁰⁵	0.29	0.24	0.25	2.71×10 ⁻⁰⁵	1.72×10 ⁻⁰⁴
N-acetyl-glycine	ACY1	3	7	2.06×10 ⁻⁰⁸	0.60	0.33	0.33	3.93×10 ⁻⁰⁸	6.05×10 ⁻⁰⁵	3	6.95×10 ⁻⁰⁴	0.72	0.77	0.78	1.58×10 ⁻⁰³	6.66×10 ⁻⁰³
leucylasparagine	KLKB1	4	13	4.32×10 ⁻⁰⁷	0.97	0.45	0.46	4.87×10 ⁻⁰⁷	0.04	10	2.28×10 ⁻⁰²	0.28	0.16	0.14	0.02	0.05
dimethylglycine	DMGDH	5	12	1.97×10 ⁻³³	0.17	0.08	0.06	4.57×10 ⁻³⁶	7.73×10 ⁻³⁰	8	5.13×10 ⁻⁰⁹	0.71	0.62	0.63	2.56×10 ⁻⁰⁹	1.95×10 ⁻⁰⁸
arachidonate (20:4n6)	PLA2G7	6	7	3.22×10 ⁻⁰⁷	0.66	0.65	0.67	6.19×10 ⁻⁰⁷	0.04	7	0.12	0.03	0.02	0.01	0.02	0.05
indolelactate	CCBL1	9	9	6.53×10 ⁻²³	0.62	0.58	0.62	5.15×10 ⁻²³	5.81×10 ⁻¹⁷	4	7.75×10 ⁻⁰⁵	NA	0.34	0.30	2.70×10 ⁻⁰⁴	1.47×10 ⁻⁰³
N-acetyl-betaalanine	PTER	10	7	3.05×10 ⁻⁰⁸	0.70	0.62	0.55	3.07×10 ⁻⁰⁸	0.006	4	3.28×10 ⁻¹¹	0.09	0.10	0.09	9.84×10 ⁻¹²	1.87×10 ⁻¹⁰
isobutyrylcarnitine	ACAD8	11	3	6.90×10 ⁻¹³	0.27	0.39	0.41	6.98×10 ⁻¹³	3.37×10 ⁻⁰⁷	3	3.08×10 ⁻¹⁰	0.25	0.19	0.20	2.58×10 ⁻¹⁰	2.45×10 ⁻⁰⁹
deoxycarnitine	SLC25A45	11	8	1.07×10 ⁻⁰⁷	0.50	0.65	0.66	1.70×10 ⁻⁰⁷	0.02	5	2.24×10 ⁻⁰³	0.40	0.39	0.45	4.37×10 ⁻⁰³	0.02
histidine	HAL	12	11	4.23×10 ⁻⁰⁷	0.98	0.99	0.98	7.07×10 ⁻⁰⁷	0.05	14	1.14×10 ⁻⁰³	0.22	0.17	0.16	8.69×10 ⁻⁰⁴	4.13×10 ⁻³
leucylserine	NPC2	14	4	1.07×10 ⁻³	-	5.83×10 ⁻⁰⁴	3.79×10 ⁻⁰⁴	1.93×10 ⁻⁰⁷	0.02	5	2.79×10 ⁻⁰³	0.08	0.03	0.07	6.78×10 ⁻⁰³	0.02

Main.p: *p*-value of genetic main effect from SKAT test; MiSTi.p: *p*-value of MiSTi interaction test; Fix.int.p: *p*-value of rareGE fixed effect interaction test; Ran.int.p: *p*-value of rareGE random effect interaction test; Joint.p: *p*-value of rareGE joint test; FDR-*Q*: false discovery rate *Q*-values of rareGE joint test

**CHAPTER IV. POWER OF TWO EMERGING METHODS FOR DETECTING
AND CHARACTERIZING GENE×SEX INTERACTION EFFECTS FOR RARE
VARIANT ANALYSES COMPARED TO STANDARD STRATIFIED ANALYSES**

Abstract

Although it is well known that complex diseases are influenced by both genetic and environmental factors, examples of validated gene by environment (G×E) interactions, especially for rare variants, are not common in epidemiological studies. One reason can be incomplete knowledge of the power of statistical methods used to search for rare variant G×E interactions in a given dataset. Improved understanding of the power of G×E interaction analyses may lead to better analysis and characterization of G×E interactions. We carried out a simulation study to investigate the performance of two newly developed methods, rareGE and MiSTi, that extend well-established common variant approaches in detecting rare variant gene-sex interaction effects on a quantitative phenotype. Compared with conventional burden tests, rareGE and MiSTi have superior performance in their power of identifying rare variant gene-sex interactions under a wide range of scenarios. Simulation results illustrate that an approach that jointly tests the genetic main effects and gene-sex interactions increases statistical power and has the potential to uncover novel genetic signals that have not been identified previously. In summary, use of simulated data for evaluation of the statistical power of emerging methods to detect rare variant G×E interactions shows an increase in statistical power for these newly introduced methods and justifies their use in practice.

Introduction

Traditional genome-wide association studies (GWAS) have successfully identified a large number of loci associated with complex diseases and quantitative risk factor phenotypes. However, a large proportion of the heritability of these diseases/traits remains unexplained (1). Gene-environment (G×E) interactions, defined as different effects of a genotype on disease risk between differing environmental exposures (2, 3), and rare and low-frequency genetic variants, defined as variants with minor allele frequency (MAF) $\leq 5\%$, may both account for some of the unexplained heritability of complex disease-related phenotypes (4).

Several large-scale genome-wide G×E studies have successfully identified novel loci accounting for the modifying effects of environmental exposures such as age, sex, BMI, alcohol consumption, and smoking status on cardiovascular disease (CVD) and its related intermediate traits (5-9). Studying G×E interactions involving rare variants may further extend our knowledge of the genetic architecture of complex traits and improve our understanding of the underlying mechanisms of common diseases (10-12). However, unlike well-established G×E interaction tests for common variants (13, 14), methods development for detecting rare variant G×E interactions is challenging for several reasons. First, considering typical sample sizes of most published GWAS studies, a single marker test is underpowered for rare and low frequency variants with MAF $\leq 5\%$. Second, conventional burden tests that simply summarize the total number of variants within a region and fit a model with this burden by environment interaction term, often result in inflated type 1 error rates and biased estimates when the genes and environment are not independent (i.e. G×E correlation) (15).

Recently developed novel approaches for testing rare variant G×E interaction effects (15-20) face limitations. Jiao and colleagues (18, 19) treated genetic main effects as fixed effects, which may suffer from inflated type I error when the variants are rare (21). Lin et al proposed an interaction Sequence Kernel Association Test (15) that is powerful when both positive and negative directions of G×E effects exist, yet loses power when the variants in the set have the same direction of G×E effects. Tzeng et al. (16) assumed comparable magnitude of the variance component parameters for genetic main effects and G×E interactions, which may not be powerful if this assumption is not satisfied. Emerging methods have been proposed to overcome the aforementioned limitations. Su et al. proposed a novel and rigorous framework, Mixed effects Score Tests for interaction (MiSTi), to derive independent score statistics for fixed effects and the variance component, which is more powerful to test G×E interaction terms of rare variants (21). A joint test that allows one to simultaneously test genetic main effects and interaction effects and requires no assumption about the magnitude of the variance component parameters for the genetic main effects and G×E interactions was proposed and successfully implemented by Chen and colleagues in the R package called ‘rareGE’ (22). The former interaction-only test allows detecting G×E interactions regardless of the genetic main effect, while the latter joint testing approach aims to detect associated genetic effects allowing for gene-environment interactions.

Compared to common variant analyses, rare variant analyses often require a larger sample size to attain comparable power. Interaction analyses also need larger sample sizes in comparison with main effect analysis (23, 24). Therefore, interaction analyses for rare genetic variants require extra attention, particularly related to consideration of statistical power in studies

with a fixed sample size. In this chapter, we compared the performance and power of two emerging approaches “rareGE” (22) and “MiSTi” (21) with standard stratified analyses followed by a test of the differences of the effect sizes, “Z test”, (25) in simulation studies using real genotype data from European-Americans (EAs) in the Atherosclerosis Risk in Communities (ARIC) study.

Methods- Simulation Studies

“MiSTi” and “rareGE” have been shown to maintain a correct type I error rate under the null hypothesis (no G×E interactions) (21, 22). I evaluated and compared their power with the “Z test” for detecting rare variant gene-sex interactions under different scenarios assuming the gene variants and gene-sex interactions were associated with a quantitative phenotype (metabolite levels) but with varying effect sizes and directions of effects, as well as the total sample size. I also investigated the power of the rareGE joint test that allows one to simultaneously test genetic main effects and interaction effects in the aforementioned scenarios.

Study Sample

The ARIC study is a population-based prospective cohort study of 15,792 adults from four U.S. communities (Forsyth County, NC; Jackson, MS; suburbs of Minneapolis, MN; and Washington County, MD), which has been described in detail previously (29). ARIC included both EAs and AAs aged 45-64 at the baseline examination (1987-1989). Participants completed three additional triennial follow-up examinations, a fifth exam in 2011-2013, and a sixth exam in 2016-2017. There were 11,071 participants with exome chip genotyped data at the baseline

examination. The ARIC study has been approved by the institutional review boards at each site, and written informed consent was obtained from all participating individuals.

Simulation Design

Part 1. Detecting gene-sex interaction

To evaluate the performance of the two emerging approaches and the conventional Z test in detecting gene-sex interaction, I first selected 10 genes having varying number of SNVs and pattern of linkage disequilibrium from the exome chip data genotyped in 11,071 EAs.

For each gene, the modeled metabolite was generated with 500 replicates from,

$Y = \beta_0 + \beta_E sex + \sum_{p=1}^P \beta_p^G G_p + \sum_{p=1}^P \beta_p^{GE} G_p sex + \varepsilon$, where $\varepsilon \sim N(0, 1)$ is a normal error term, β_0 and β_E was estimated from ARIC's real data. For example, to generate glycine levels, the estimated β_0 and β_E from ARIC's real data are $\beta_0 = 1.4$, $\beta_E = 0.6$, and

$$\text{Glycine} = 1.4 + 0.6 * sex + \sum_{p=1}^P \beta_p^G G_p + \sum_{p=1}^P \beta_p^{GE} G_p (sex - 0.5) + \varepsilon$$

The proportion of causal SNVs was set to 20% for each of the 10 genes. The effect size of non-causal SNVs was set to be zero. The effect size of the causal genetic main effect β_p^G was simulated under 2 settings:

Setting 1. Randomly selected causal SNVs $\beta_p^G \sim U(0,1)$, and 0 otherwise.

Setting 2. 10% randomly selected causal SNVs $\beta_p^G \sim U(0,1)$, and the betas for the other 10% will be randomly selected causal SNVs $\beta_p^G \sim U(-1,0)$, and 0 otherwise.

Settings 1 and 2 simulated two extreme cases where genetic main effects favors burden (all in the same direction) and variance component (50% positive and 50% negative), respectively.

For each genetic main effect setting considered, the size of interaction effects was controlled by a constant c , which varied from 0.5, 1, 1.5 to 2, so that the power estimated under difference methods was discernible. I simulated 2 scenarios of gene-sex interactions:

Scenario 1. 20% of randomly selected $\beta_p^{GE} = c$, and 0 otherwise

Scenario 2. 10% of randomly selected $\beta_p^{GE} = c$, and 10% of randomly selected $\beta_p^{GE} = -c$, and 0 otherwise

Scenario 1 represented a moderate interaction effect scenario with 20% of the variants having the interaction effect in the same direction. Scenario 2 represented a moderate interaction effect with 20% variants having the interaction effect in opposite directions. The constant c was varied from 0.5, 1, 1.5, to 2 to evaluate the size of the interaction effect on power. The empirical power of each method under each scenario was calculated by comparing the resulting p-value to a cut-off value declaring statistical significance, α . I then calculated the proportion of times the null hypothesis was rejected (success rate) over the 500 replicates. I considered two α levels, one for nominal significance level $p < 0.05$, the other for exome-wide false discovery rate (FDR) $< 5\%$. Because I simulated metabolite levels based on variants within one gene each time and expected only one gene to be associated with the simulated phenotype across the exome, the $\text{FDR} < 5\%$ is equivalent to a Bonferroni corrected $p < 4.95 \times 10^{-6}$.

Part 2. Effect of sample size on power

To evaluate the effect of sample size on power, I doubled the exome chip data and then randomly selected subsets of the doubled genomic dataset to vary sample size from 20,000 down to 2,000. For this power simulation, I selected two scenarios when causal markers have main

effects in opposite directions (same as setting 2 of genetic main effects in part 1) and gene-sex interaction effects in the same direction and opposite directions (same as scenario 1 and 2 of interaction effects in part 1). To be specific, again using glycine as an example, the dataset was generated with parameters settings as below

1. Glycine = $1.4 + 0.6 \cdot \text{sex} + \sum_{p=1}^P \beta_p^G G_p + \sum_{p=1}^P \beta_p^{GE} G_p (\text{sex} - 0.5) + \varepsilon$
2. 20% of the SNVs within each gene are causal SNVs
3. 10% randomly selected causal SNVs $\beta_p^G \sim U(0,1)$, and the betas for the other 10% will be randomly selected causal SNVs $\beta_p^G \sim U(-1,0)$, and 0 otherwise
4.
 - a. 10% of randomly selected $\beta_p^{GE} = 1$, and 10% of randomly selected $\beta_p^{GE} = -1$, and 0 otherwise
 - b. 20% of randomly selected $\beta_p^{GE} = 1$, and 0 otherwise

Step 3 generated 20% causal variants with genetic main effects in opposite directions. Step 4a represented a moderate interaction effect with 20% of the variants having the interaction effects in the opposite directions. 4b represented a moderate interaction effect with 20% variants having the interaction effects in the same direction. Following the same procedure as in part 1, the empirical power under different sample sizes was calculated by comparing the resulting p-value to the significance level α (0.05 or 4.95×10^{-6}) to determine success or failure and then computing the rate over 500 replicates.

Results

Part 1. Performance in detecting gene-sex interaction

The ten selected genes with varying number of SNVs and cumulative minor allele counts (cMAC) are presented in **Table IV-1**. The LD pattern for each of the selected genes is presented in supplemental figure 1. Rare and low-frequency variants aggregated in the selected genes are not in LD or in very low to moderate LD. The highest average LD observed is for the gene *KIAA1551* (average LD < 0.1).

Figure IV-1 shows the average power results across the 10 genes using the three methods with positive genetic main effects and two scenarios of interaction effects. Empirical power was calculated at the significance level of 0.05 and 4.95×10^{-6} , respectively. The data shows a clear trend of increasing power with increasing effect sizes for the interaction effects and highlights that these two newly developed methods outperform the conventional Z test under each situation investigated here. Notably, at exome-wide significance, even these newly developed methods for testing interaction effects are greatly underpowered (less than 50% power) for a sample size of 11,000, while the joint test of genetic main effects and gene-sex interaction effects has a nearly 70% power for the same sample size with a modest genetic main effect we simulated. The data presented in **Figure IV-1** shows that MiSTi is a more powerful test than rareGE when the causal markers have interaction effects in the same direction. They also suggest that the rareGE random effect interaction test has the highest power when the causal markers have interaction effects in opposite directions, which agrees with our prior expectation; the SKAT-type tests are most powerful when causal markers have interaction effects in opposite directions. Both MiSTi and

rareGE showed a substantial higher power than the Z test in each scenario as demonstrated in **Figure IV-1** .

We then examine factors that influenced the performance of each methods in each of the 10 genes, including, 1) number of SNVs and cMAC within each gene; 2) effect size of the interaction effects; 3) causal markers within a gene having gene-sex interaction effects in the same vs opposite directions; 4) genetic main effects of causal markers in the same vs opposite directions. The power results are presented at the significance levels of 0.05 and Bonferroni corrected 4.95×10^{-6} and the average power across the 10 genes are summarized in **Table IV-2** and accompanying **Figures**. **Figure IV-2** shows that power increases with increasing number of SNVs and cMAC aggregated within each gene. Comparing **Figure IV-3** to **Figure IV-2** shows that the power increases with increasing interaction effect size for each gene (**Figure IV-3** gene-sex interaction effect $c = 2$ vs. **Figure IV-2** gene-sex interaction effect $c = 1$). MiSTi appears to be slightly more powerful than the rareGE fixed effect interaction test for most genes, which is consistent with the results observed for the average power across the 10 genes (**Figure IV-1**). **Figure IV-4** shows the power when causal markers within a gene have interaction effects in opposite directions, and demonstrates that both newly developed interaction test methods are superior than the Z test when the causal markers have interaction effects in different directions. Under such a scenario, the rareGE random effect test has the highest power to test interaction effects for most genes, which again matches our expectation. Compared to **Figure 4b**, **Figure IV-5** suggests that that the power remains almost the same no matter whether the causal markers have the same or different directions of main effects except that the Z test was the least powerful

under the situation that causal markers have different directions of both genetic main effects and interaction effects.

Part 2. Effect of sample size on power

Figure IV-6 shows the average power results across 10 genes using the three methods with genetic main effects in opposite directions, and interaction effects in either the same direction or opposite directions considering sample size varying from 2000, 5000, 10000, 15000 to 20000. Empirical power was calculated at the significance levels of 0.05 and 4.95×10^{-6} , respectively. **Figure IV-6** shows a clear trend of increasing power with increasing sample size and demonstrates that the two newly developed methods, rareGE and MiSTi consistently outperforms the conventional Z test, but are still greatly underpowered (40%) to detect an exome-wide significance G×E interaction with modest effect size 1 in a sample size of 20,000. Figure 6a and 6b show the power when causal markers have interaction effects in opposite directions, and supported our results in part 1 that rareGE random effect test has the highest power to test interaction effects in such scenario, although the difference between the power of rareGE random effect test and MiSTi is small. Similarly, Figure 6c and 6d again show that MiSTi is a more powerful test than rareGE when the causal markers have interaction effects in the same direction regardless of varied sample size. For the rareGE joint test that simultaneously considers genetic main effects and interaction effects, the sample size required to detect a gene at nominal significance level with 80% power would be greater than 10,000 using the simulated effect sizes based on real data. To detect a gene at exome-wide significance level with sufficient

power, an even larger sample size ($> 20,000$) will be required for the effect sizes considered here.

Discussion

In this chapter, we compared three methods, rareGE, MiSTi and a conventional Z test, and evaluated their power in detecting gene-sex interaction effects as well as jointly testing for genetic main effects and gene-sex interaction effects. We show that, 1) both rareGE and MiSTi tests are more powerful than the conventional Z test in detecting gene-sex interaction effects, in the context of rare genetic variants analysis; 2) rareGE joint test is the most powerful when both genetic main effect and gene-sex interaction are present, and the power increases with increasing effect sizes for the interaction effects.

Compared with a conventional Z test of the interaction effects, rareGE and MiSTi tests have higher power for the simulated situations considered here, especially when causal genetic markers have different directions of gene by environment interaction effects. When causal markers have gene by environment interaction effects in the same direction, MiSTi slightly outperforms rareGE because rareGE is a SKAT-type test and suffers loss of power in such scenario (15, 22, 26). In contrast, when causal markers have gene by environment interaction effects in opposite directions, rareGE outperforms MiSTi, because in this scenario the interaction effect model favors the variance component, which is the scenario that a SKAT-type test have the greatest power (27). The rareGE joint test that simultaneously tests genetic main effects and interaction effects is generally more powerful across the simulated scenarios considered here, suggesting that a joint test is an attractive approach for testing genetic associations allowing for

G×E interactions when genetic main effects exist (28). Using this joint test, the results show that we have sufficient power to detect a gene with moderate effect size at nominal significance level with a sample size of 10000. However, no matter what methods is used, a much larger sample size is required to detect exome-wide significant genes.

We applied the rareGE and MiSTi approaches from Chapter 3 of this dissertation using real data comprising 271 measured and named metabolites and exome chip data genotyped in 3,540 African-Americans (AAs) and EAs from the ARIC study. We identified and replicated 14 gene-metabolite pairs through joint test, including 3 novel associations. There was no exome-wide significant gene-sex interaction using either rareGE or MiSTi approach. The real data results are in line with our simulation results for power: these two newly developed methods are underpowered to detect an exome-wide significance G×E interaction under the sample sizes available in the ARIC study. We successfully detected a few novel genes associated with metabolites through the joint test, likely because the genetic main effects on metabolites are normally much larger than that for disease or disease risk factor levels (29, 30).

The present simulation study has some limitations. We considered only two extreme situations, causal genetic variants with G×E interaction effects all in the same direction or completely in opposite directions. In practice, the directions of causal genetic variants with G×E interaction effects contributing to complex diseases are most likely a mixture of the two scenarios. Also, in practice, the proportion of causal variants in a gene may not be 20% as we simulated. In addition, we did not consider multiple genes simultaneously; the quantitative phenotype was simulated based on effects of genetic variants within one gene. Under a polygenic

scenario, the power of testing a particular $G \times E$ interaction may be affected by other genetic main effects or interactions.

In conclusion, we have shown in the context of rare genetic variants that utilizing emerging statistical methods for detecting $G \times E$ interactions leads to an increase in power. The approach of jointly testing the genetic main effects and $G \times E$ interactions for rare variants has the potential to detect novel genes associated with a phenotype of interest. Our simulations justify their use in practice and provide guidance on sample size needed under various scenarios.

References

1. Maher B. The case of the missing heritability. *Nature*. 2008;456(7218):18.
2. Ottman R. Gene–Environment Interaction: Definitions and Study Designs. *Preventive medicine*. 1996;25(6):764-70.
3. Mather K, Caligari PD. Genotype x environment interactions. IV. The effect of the background genotype. *Heredity*. 1976;36(1):41-8.
4. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. *Nature*. 2009;461(7265):747-53.
5. Manning AK, Hivert M-F, Scott RA, Grimsby JL, Bouatia-Naji N, Chen H, et al. A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nature genetics*. 2012;44(6):659-69.
6. Winkler TW, Justice AE, Graff M, Barata L, Feitosa MF, Chu S, et al. The Influence of Age and Sex on Genetic Associations with Adult Body Size and Shape: A Large-Scale Genome-Wide Interaction Study. *PLOS Genetics*. 2015;11(10):e1005378.
7. Simino J, Shi G, Bis Joshua C, Chasman Daniel I, Ehret Georg B, Gu X, et al. Gene-Age Interactions in Blood Pressure Regulation: A Large-Scale Investigation with the CHARGE, Global BPgen, and ICBP Consortia. *The American Journal of Human Genetics*. 2014;95(1):24-38.

8. Taylor JY, Schwander K, Kardina SLR, Arnett D, Liang J, Hunt SC, et al. A Genome-wide study of blood pressure in African Americans accounting for gene-smoking interaction. *Scientific Reports*. 2016;6:18812.
9. Justice AE, Winkler TW, Feitosa MF, Graff M, Fisher VA, Young K, et al. Genome-wide meta-analysis of 241,258 adults accounting for smoking behaviour identifies novel loci for obesity traits. *Nature Communications*. 2017;8:14977.
10. Thomas D. Gene–environment-wide association studies: emerging approaches. *Nature Reviews Genetics*. 2010;11(4):259-72.
11. Hunter DJ. Gene–environment interactions in human diseases. *Nature Reviews Genetics*. 2005;6(4):287-98.
12. Rao DC, Sung YJ, Winkler TW, Schwander K, Borecki I, Cupples LA, et al. Multiancestry Study of Gene–Lifestyle Interactions for Cardiovascular Traits in 610 475 Individuals From 124 Cohorts: Design and Rationale. *Circulation: Cardiovascular Genetics*. 2017;10(3):e001649.
13. Kraft P, Yen Y-C, Stram DO, Morrison J, Gauderman WJ. Exploiting gene-environment interaction to detect genetic associations. *Human heredity*. 2007;63(2):111-9.
14. Manning AK, LaValley M, Liu C-T, Rice K, An P, Liu Y, et al. Meta-analysis of gene-environment interaction: joint estimation of SNP and SNP \times environment regression coefficients. *Genetic Epidemiology*. 2011;35(1):11-8.
15. Lin X, Lee S, Wu MC, Wang C, Chen H, Li Z, et al. Test for Rare Variants by Environment Interactions in Sequencing Association Studies. *Biometrics*. 2016;72(1):156-64.

16. Tzeng J-Y, Zhang D, Pongpanich M, Smith C, McCarthy MI, Sale MM, et al. Studying gene and gene-environment effects of uncommon and common variants on continuous traits: a marker-set approach using gene-trait similarity regression. *The American Journal of Human Genetics*. 2011;89(2):277-88.
17. Zhao G, Marceau R, Zhang D, Tzeng J-Y. Assessing gene-environment interactions for common and rare variants with binary traits using gene-trait similarity regression. *Genetics*. 2015;199(3):695-710.
18. Jiao S, Hsu L, Bézieau S, Brenner H, Chan AT, Chang-Claude J, et al. SBERIA: Set-Based Gene-Environment Interaction Test for Rare and Common Variants in Complex Diseases. *Genetic epidemiology*. 2013;37(5):452-64.
19. Jiao S, Peters U, Berndt S, Bézieau S, Brenner H, Campbell PT, et al. Powerful set-based gene-environment interaction testing framework for complex diseases. *Genetic epidemiology*. 2015;39(8):609-18.
20. Marceau R, Lu W, Holloway S, Sale MM, Worrall BB, Williams SR, et al. A Fast Multiple-Kernel Method With Applications to Detect Gene-Environment Interaction. *Genetic epidemiology*. 2015;39(6):456-68.
21. Su Y-R, Di C-Z, Hsu L. A unified powerful set-based test for sequencing data analysis of GxE interactions. *Biostatistics*. 2017;18(1):119-31.
22. Chen H, Meigs JB, Dupuis J. Incorporating Gene-Environment Interaction in Testing for Association with Rare Genetic Variants. *Human Heredity*. 2014;78(2):81-90.
23. Hunter DJ. Gene–environment interactions in human diseases. *Nature Reviews Genetics*. 2005;6(4):287.

24. VanderWeele TJ. Sample Size and Power Calculations for Additive Interactions. *Epidemiologic methods*. 2012;1(1):159-88.
25. Paternoster R, Brame R, Mazerolle P, Piquero A. Using the correct statistical test for the equality of regression coefficients. *Criminology*. 1998;36(4):859-66.
26. Basu S, Pan W. Comparison of statistical tests for disease association with rare variants. *Genetic epidemiology*. 2011;35(7):606-19.
27. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet*. 2011;89(1):82-93.
28. Kraft P, Yen YC, Stram DO, Morrison J, Gauderman WJ. Exploiting gene-environment interaction to detect genetic associations. *Hum Hered*. 2007;63(2):111-9.
29. Groves CJ, Zeggini E, Minton J, Frayling TM, Weedon MN, Rayner NW, et al. Association analysis of 6,736 U.K. subjects provides replication and confirms TCF7L2 as a type 2 diabetes susceptibility gene with a substantial effect on individual risk. *Diabetes*. 2006;55(9):2640-4.
30. Shin S-Y, Fauman EB, Petersen A-K, Krumsiek J, Santos R, Huang J, et al. An atlas of genetic influences on human blood metabolites. *Nat Genet*. 2014;46(6):543-50.

Table IV-1 Characteristics of the 10 selected genes in European- Americans from the ARIC exome chip data

Gene	CHR	nSNV	nSNVused	cMAC
<i>SEMG1</i>	20	12	8	275
<i>LIPG</i>	18	12	11	338
<i>NEFM</i>	8	14	13	668
<i>ANKS3</i>	16	20	14	1482
<i>SLC26A4</i>	7	27	21	349
<i>SYCP2</i>	20	29	17	2899
<i>WDR17</i>	4	32	28	1106
<i>KIAA1551</i>	12	42	31	258
<i>CELSR3</i>	3	54	41	845
<i>COL6A3</i>	2	92	77	2111

Table IV-2 Power of various methods under scenarios with 2 settings of genetic main effects (setting 1 genetic main effects in the same direction and setting 2 genetic main effects in opposite directions) and two scenarios of interaction effects (GxE effects in the same direction and GxE effects in opposite directions), respectively. The significance levels α are 0.05, and bonferroni corrected 4.95×10^{-6} , respectively. 2a. the size of interaction effect $c=0.5$; 2b. the size of interaction effect $c=1$; 2c. the size of interaction effect $c=1.5$; 2d. the size of interaction effect $c=2$.

Table 2a.

Alpha levels	Methods	GxE same direction $c=0.5$		GxE opposite direction $c=\pm 0.5$	
		Setting 1	Setting 2	Setting 1	Setting 2
0.05	MiSTi	0.395	0.403	0.355	0.357
	rareGE	0.396	0.393	0.387	0.384
	Z test	0.329	0.347	0.259	0.254
	Joint	0.702	0.678	0.682	0.668
4.95×10^{-6}	MiSTi	0.158	0.165	0.137	0.137
	rareGE	0.157	0.164	0.151	0.149
	Z test	0.456	0.084	0.038	0.041
	Joint	0.077	0.434	0.434	0.427

Table 2b.

Alpha levels	Methods	GxE same direction $c=1$		GxE opposite direction $c=\pm 1$	
		Setting 1	Setting 2	Setting 1	Setting 2
0.05	MiSTi	0.578	0.579	0.507	0.513
	rareGE	0.558	0.565	0.542	0.553
	Z test	0.490	0.498	0.365	0.371
	Joint	0.785	0.768	0.762	0.759
4.95×10^{-6}	MiSTi	0.332	0.331	0.300	0.299
	rareGE	0.331	0.329	0.320	0.325
	Z test	0.245	0.250	0.187	0.186
	Joint	0.569	0.546	0.549	0.536

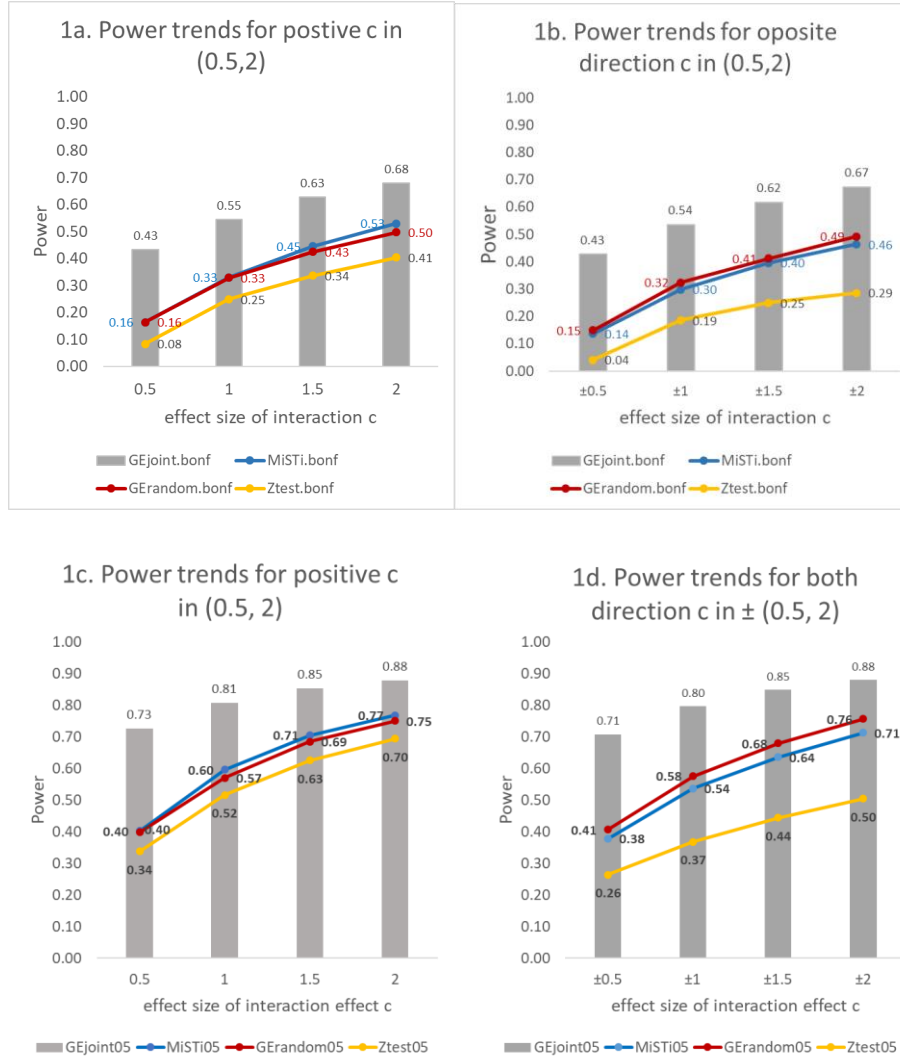
Table 2c.

Alpha levels	Methods	GxE same direction c=1.5		GxE opposite direction c=±1.5	
		Setting 1	Setting 2	Setting 1	Setting 2
0.05	MiSTi	0.693	0.690	0.604	0.618
	rareGE	0.672	0.667	0.650	0.657
	Z test	0.599	0.600	0.433	0.442
	Joint	0.838	0.822	0.819	0.814
4.95×10 ⁻⁶	MiSTi	0.442	0.447	0.392	0.396
	rareGE	0.422	0.425	0.410	0.412
	Z test	0.336	0.337	0.249	0.250
	Joint	0.642	0.629	0.623	0.618

Table 2d.

Alpha levels	Methods	GxE same direction c=2		GxE opposite direction c=±2	
		Setting 1	Setting 2	Setting 1	Setting 2
0.05	MiSTi	0.753	0.756	0.677	0.689
	rareGE	0.742	0.732	0.719	0.727
	Z test	0.673	0.678	0.505	0.494
	Joint	0.862	0.854	0.851	0.850
4.95×10 ⁻⁶	MiSTi	0.525	0.531	0.454	0.464
	rareGE	0.487	0.498	0.483	0.492
	Z test	0.397	0.405	0.297	0.286
	Joint	0.701	0.680	0.678	0.674

Figure IV-1 Power of three methods with positive genetic main effects and two scenarios of interaction effects (1a & 1c. GxE in the same direction; 1b & 1d. GxE in opposite directions). The significance threshold for 1a & 1b is Bonferroni corrected (4.95×10^{-6}), for 1c & 1d is 0.05.



GEjoint.bonf/GEjoint05: the power of rareGE joint test under Bonferroni corrected/0.05 significance threshold; MiSTi.bonf/MiSTi05: the power of MiSTi interaction test under Bonferroni corrected/0.05 significance threshold; GERandom.bonf/GERandom05: the power of rareGE random effect interaction test under Bonferroni corrected/0.05 significance threshold; Ztest.bonf/Ztest05: the power of the conventional Z test under Bonferroni corrected/0.05 significance threshold

Figure IV-2 Power of three methods with genetic main effects in both directions and gene-sex interaction effects in the same direction ($c=1$). The significance threshold for 2a is 0.05 for 2b is Bonferroni corrected 4.95×10^{-6} .

Figure 2a.

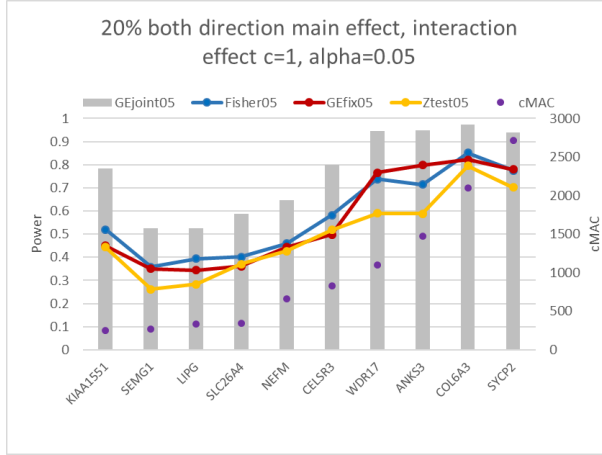
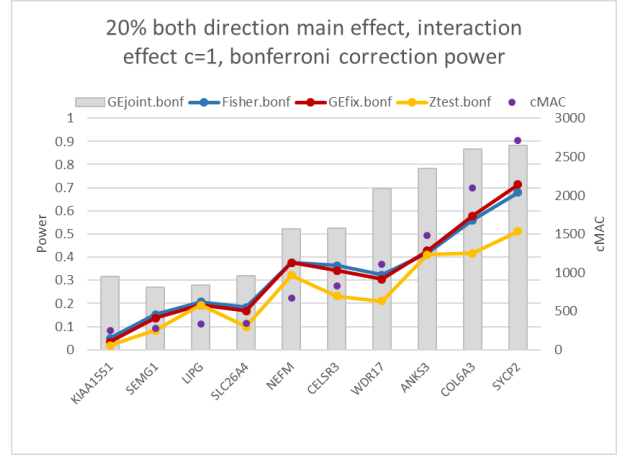


Figure 2b.



GEjoint.bonf/GEjoint05: the power of rareGE joint test under Bonferroni corrected/0.05 significance threshold; Fisher.bonf/Fisher05: the power of MiSTi interaction test under Bonferroni corrected/0.05 significance threshold; GEfix.bonf/GEfix05: the power of rareGE fix effect interaction test under Bonferroni corrected/0.05 significance threshold; Ztest.bonf/Ztest05: the power of the conventional Z test under Bonferroni corrected/0.05 significance threshold; cMAC: cumulative minor allele counts

Figure IV-3 Power of three methods with genetic main effects in both directions and gene-sex interaction effects in the same direction ($c=2$). The significance threshold for 3a is 0.05 for 3b is Bonferroni corrected 4.95×10^{-6} .

Figure 3a.

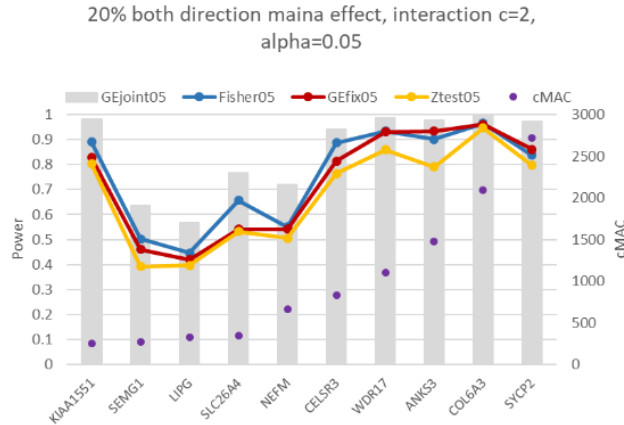
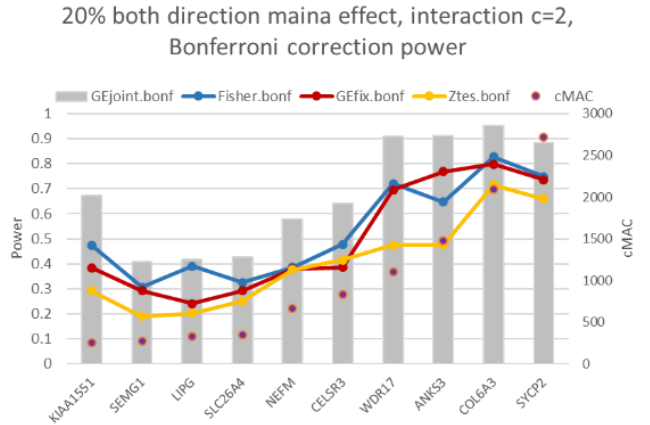


Figure 3b.



GEjoint.bonf/GEjoint05: the power of rareGE joint test under Bonferroni corrected/0.05 significance threshold; Fisher.bonf/Fisher05: the power of MiSTi interaction test under Bonferroni corrected/0.05 significance threshold; GEfix.bonf/GEfix05: the power of rareGE fix effect interaction test under Bonferroni corrected/0.05 significance threshold; Ztest.bonf/Ztest05: the power of the conventional Z test under Bonferroni corrected/0.05 significance threshold; cMAC: cumulative minor allele counts

Figure IV-4 Power of three methods with genetic main effects and gene-sex interaction effects in different directions ($c = \pm 2$). The significance threshold for 4a is 0.05 for 4b is Bonferroni corrected 4.95×10^{-6} .

Figure 4a.

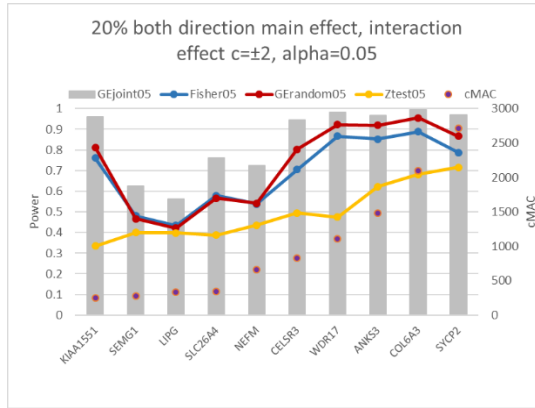
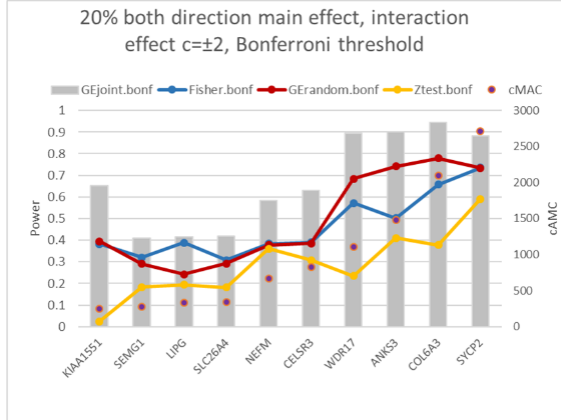
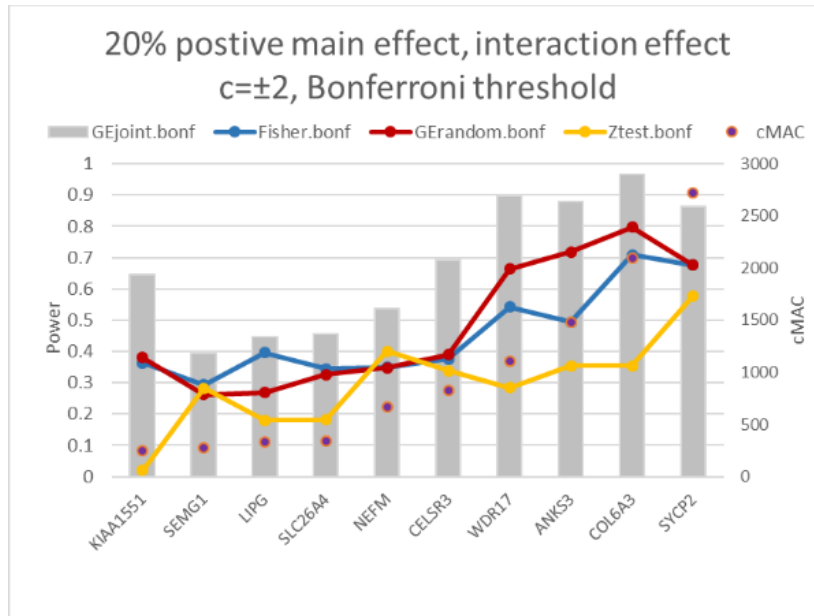


Figure 4b.



GEjoint.bonf/GEjoint05: the power of rareGE joint test under Bonferroni corrected/0.05 significance threshold; Fisher.bonf/Fisher05: the power of MiSTi interaction test under Bonferroni corrected/0.05 significance threshold; GERandom.bonf/GERandom05: the power of rareGE random effect interaction test under Bonferroni corrected/0.05 significance threshold; Ztest.bonf/Ztest05: the power of the conventional Z test under Bonferroni corrected/0.05 significance threshold; cMAC: cumulative minor allele counts

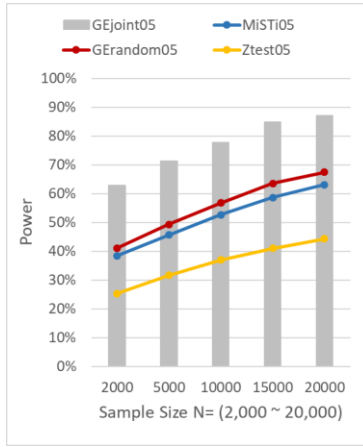
Figure IV-5 Power of three methods with genetic main effects in the same direction and gene-sex interaction effects in different directions ($c = \pm 2$). The significance threshold is Bonferroni corrected 4.95×10^{-6} .



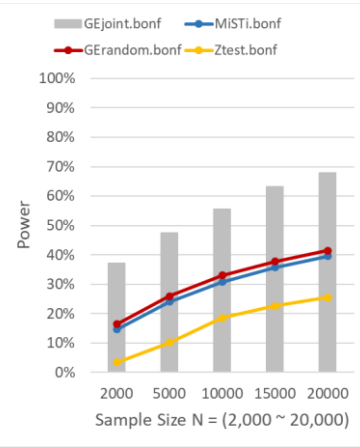
GEjoint.bonf: the power of rareGE joint test under Bonferroni corrected significance threshold; Fisher.bonf: the power of MiSTi interaction test under Bonferroni corrected significance threshold; GERandom.bonf: the power of rareGE random effect interaction test under Bonferroni corrected significance threshold; Ztest.bonf: the power of the conventional Z test under Bonferroni corrected significance threshold; cMAC: cumulative minor allele counts

Figure IV-6 Power comparisons of three methods with genetic main effects in opposite directions and two scenarios of interaction effects (6a & 6b. GxE in opposite directions $c = \pm 1$; 6c & 6d. GxE in the same direction $c = 1$) under different significance thresholds and varied sample size.

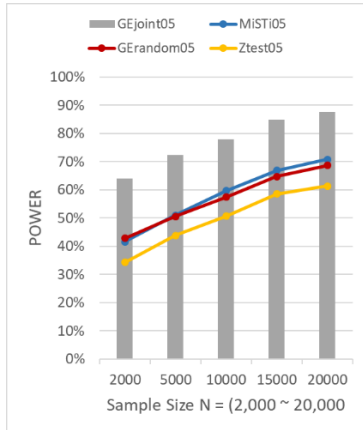
6a. $\alpha = 0.05$



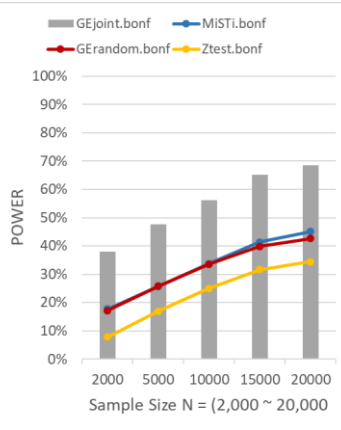
6b. $\alpha = 4.95 \times 10^{-6}$



6c. $\alpha = 0.05$



6d. $\alpha = 4.95 \times 10^{-6}$



GEjoint.bonf/GEjoint05: the power of rareGE joint test under Bonferroni corrected/0.05 significance threshold; Fisher.bonf/Fisher05: the power of MiSTi interaction test under Bonferroni corrected/0.05 significance threshold; GERandom.bonf/GERandom05: the power of rareGE random effect interaction test under Bonferroni corrected/0.05 significance threshold; Ztest.bonf/Ztest05: the power of the conventional Z test under Bonferroni corrected/0.05 significance threshold; cMAC: cumulative minor allele counts

CHAPTER V. SYNTHESIS

This dissertation utilized metabolomic profiling and exome chip data to evaluate sex-specific genetic effects and gene-sex interactions on the serum metabolome with a particular focus on rare and low-frequency (minor allele frequency $\leq 5\%$) genetic variants (Chapter 2-3). The study participants included both African-Americans (AAs) and European-Americans (EAs) belonging to the large population-based Atherosclerosis Risk in Communities (ARIC) study. A simulation study (Chapter 4) was conducted to evaluate the power of three different methods for testing rare variant gene-sex interactions, and the results of this simulation study served as a justification of the analyses performed in Chapters 2 and 3. Overall, several genetic variants, either common genetic variants or rare and low-frequency variants aggregated within a gene, were identified to be significantly associated with metabolite levels. These findings underscore challenges and opportunities for identifying gene by environment (G×E) interactions and novel genetic loci by taking into account environmental factors and may lead to better understanding of disease.

Summary of results

In Chapter 2, we performed a sex-stratified exome-wide association study for 271 GC-MS/LC-MS measured named metabolites in ARIC AAs, and pursued replication in an independent sample of ARIC EA men and women. . A novel common variant, rs11555566, in the ADA gene was associated with N1-methyladenosine levels, which was successfully identified and replicated in both men and women in both race groups. The results suggested a

larger effect of rs11555566 in men (estimated effect 0.18-0.22) as compared to women (estimated effect 0.14-0.17), but the difference was not statistically significant. In addition, we observed variants in 6 genes using common single variant tests or burden tests suggesting differing genetic effects on metabolite levels through testing for difference of the effect size estimates in sex-stratified results. However, the difference between the sexes was not shown to be consistent in an independent sample of ARIC EAs. This study suggests that sex-specific genetic effects of metabolites may exist, but the lack of consistency in testing sex differences of the genetic effects between discovery and replication samples underscores that future studies should consider sex-specific effects with improved (i.e. more powerful in the setting of rare variants) statistical methods and tools. Accordingly, we conducted an exome-wide gene-sex interaction study using emerging statistical methods, rareGE and MiSTi in Chapter 3. To our knowledge, this is the first large-scale study to evaluate the role of rare and low frequency variants in gene-sex interactions and joint effects of genetic main and gene-sex interaction on metabolite levels. In total, we observed and replicated 14 gene-metabolite associations through the joint test, 3 of which were novel, including PLA2G7- arachidonate (20:4n6), PTER- N-acetyl-beta-alanine and NPC2- leucylserine. No significant novel loci were detected via analyzing gene-sex interactions alone. Although we applied newly developed statistical methods that were known to have improved performance in testing G×E interactions (48, 49), studies that focus on rare and low-frequency variants to identify novel loci and G×E interactions may require much larger sample size than that available in this dissertation research.

In Chapter 4, we compared three methods, rareGE, MiSTi and the conventional Z test, and evaluated their power in detecting gene-sex interaction effects as well as jointly testing for genetic main effects and gene-sex interaction effects. We illustrate that: 1) compared with the conventional Z test, rareGE and MiSTi have superior performance in their power of identifying rare variant gene-sex interactions under a wide range of scenarios, and 2) the rareGE joint test is most powerful when both genetic main effect and gene-sex interaction are present, and the power increases with increasing effect sizes for the interaction effects, which demonstrates the potential to uncover novel genetic signals that have not been identified previously.

Previous studies of the association between genetic variants and metabolite levels rarely considered sex as a potential effect modifier. These studies identified numerous genetic loci associated with one or multiple metabolite levels (73-85). In contrast to previous efforts, this dissertation is devoted to evaluating possible sex-specific common genetic effects and the role of rare and low-frequency genetic variants on the serum metabolome while taking into account sex effects. My results identified several novel genetic variants influencing the human metabolome, and the rigor of these findings was established through significant discovery in AAs and successful replication in EAs from the ARIC study. Sex-specific genetic effects and gene-sex interaction effects have been shown to contribute to the observed novel gene-metabolite associations. Our results demonstrate increased power from using emerging statistical methods for detecting gene-sex interactions and show promise for other larger scale studies analyzing rare variant GxE interactions to reveal novel biology.

Strength and Limitations

This dissertation takes full advantage of available data in the large multi-ethnic ARIC study to explore sex-specific genetic effects and gene-sex interactions with a focus on rare and low-frequency genetic variants on the metabolome using multiple statistical approaches. The genetic variants of the exome chip data in the ARIC study were jointly called in a larger pooled sample of studies conducted in the same laboratory, including the ARIC study. By increasing the sample size during the calling of variants, the ability to correctly call rare variants is enhanced (94), which facilitate the identification of novel genetic variants. In addition to the conventional Z test that tests the difference of the effect size estimates from sex-stratified analyses, we applied emerging statistical approaches for rare variant gene-sex interactions and jointly tested for genetic main effects and gene-sex interactions on the human metabolome. Previous studies have shown that inclusion of G×E interactions is important in terms of identifying novel signals, particularly for rare and low-frequency variants (38). Results in the present study supported this conclusion by showing that novel genes were identified through the joint approach.

Limitations of the present dissertation warrant consideration. There are several reasons that may have caused lack of consistency in testing sex differences of the genetic effects between AAs and EAs. First, it may be that sex does not modify the genetic effects on the human metabolome. In other words, the genetic architecture of the serum metabolome is consistent between men and women. Second, the study design involved discovery in one race group and replication in another and this may not be ideal. The discovery sample for this study was AAs, a population with high level of genetic diversity to promote novel findings

(95). However, the replication sample was EAs. Rare variants aggregated in genes may differ between the two race groups, and ancestry-specific rare variants may contribute to sex-specific effects on metabolites, which will not be consistent between races and missed in our analyses. Further studies with discovery and replication samples in an ancestry-specific manner followed by a trans-ancestry meta-analysis will have the advantage of discovering ancestry-specific rare and low-frequency genetic variants and provide evidence of trans-ancestry loci. Finally, we established that there was likely lack of statistical power to detect exome-wide rare and low-frequency genetic variant gene-sex interactions in our studies. Through simulation studies and application in real data, we demonstrated that both sex-stratified analyses followed by a Z test testing for difference of the effect size estimates and current statistical methods in detecting G×E interaction effects were not powerful enough to detect small to moderate sex difference in genetic effects on the metabolome. Improved statistical methods and tools with sufficient power and flexibility for testing G×E interactions are warranted, as well as collaborations across different studies to increase sample size.

Future Directions

The results described above need follow-up studies to better understand underlying biological processes giving rise to the observed associations and to establish potential links to disease. Follow-up investigations, such as experimental animal studies, of the genes identified in the context of interactions with sex or sex-stratified analyses are likely to provide new insights into the understanding of gene functions and biochemical changes in men and women. Researchers using genetically modified mice have revealed significant sex

differences in the development of cardiovascular phenotypes. In many of the models, cardiac pathological phenotypes were developed in male, but not in female mice as summarized by Du et al (96). For example, genetic deletion of the peroxisome proliferator-activated receptor alpha (PPARalpha), a gene involved in cellular lipid utilization, caused cardiac lipid accumulation, hypoglycemia and death in all male, but only 25% of female mice (97). In another study, the generated transgenic HDAC5S/A mice overexpressed histone deacetylase (HDAC) in cardiomyocytes which caused death in male but not in female mice (98).

The genetic markers identified in this dissertation may be used in future studies for association with disease risk. Previous studies have identified different genetic variants influencing CVD risk in a sex-specific manner (12, 13, 99, 100). Using identified genetic factors to construct genetic risk scores has been demonstrated to provide powerful and robust CVD risk prediction beyond traditional risk factors (101-103). Given these findings, improved genetic risk profile from sex-specific genetic markers is expected to further facilitate disease risk prediction.

Metabolomic profiles provide significant insights into biological and pathophysiological pathways that may be altered during the development and progression of diseases. However, metabolomic profiling performed from serum may not inform us about organ-specific pathophysiological processes. For complex diseases such as CVD, molecular changes occur within the large artery wall or liver may be more informative compared to that provided by serum. Future metabolomic studies at the organ level may supplement the measurements of changes in metabolites in serum samples and shape our understanding of metabolism. On the other hand, a recent study of metabolomic signatures of Alzheimer

disease (AD) using both brain tissue and blood samples has showed a metabolite class, sphingolipids that were consistently associated with severity of AD pathology in brain and AD progression across prodromal and preclinical stages in blood.

During the course of this dissertation, I witnessed two changes to the field of human genetics First, the transition from GWAS to whole genome sequencing studies. Second, the introduction of large sample sizes in a single study, such as the Million Veteran Program (104) and the United Kingdom Biobank (105). These changes offer both opportunities and challenges to researchers doing genetic analysis. Appropriate analytical approaches are a continued concern in genetic studies of complex disease. There are multiple challenges, for examples: how to identify causal genetic variants beyond associations; how to appropriately and informatively incorporate environmental factors; and how to improve the computational speed to satisfy the need of handling big data. Utilizing metabolomic data requires additional analytical techniques to properly account for highly correlated metabolite levels. Developing novel statistical methods such as machine learning approaches for high dimensional data will benefit future studies to advance precision medicine through integrating multi-omics data.

Most genetic studies of metabolomics focus on genetic main effects. In this dissertation, the analyses were extended to G×E interactions. However, only interactions with sex were tested. This was because sexual dimorphism in metabolites has been previously observed (87, 106), sex is easy to measure, and a balanced division between sexes may lead to greater statistical power to discover novel loci. Work on rare variants and interactions with other environmental exposures (e.g., alcohol and smoking) should be done to improve understanding and uncover additional genes associated with metabolite levels, refine known

disease loci, or reveal the underlying biology mechanism influenced by both genetic variants and environmental factors. Because of limited statistical power, we have shown that most interactions of rare and low-frequency genetic variants with environment factors cannot be identified or replicated on a genome-wide scale. One way to ease this problem is to restrict to the metabolites that were previously linked to genetic variants and environment factors, which will reduce the multiple-testing burden for analyses. Another way is to pool or meta-analyze metabolomic and genomic measurements from multiple studies together to have a much larger number of individuals than represented here. It is clear that G×E interaction studies would have an increased power by increasing the size of the discovery sample (107, 108).

Conclusions

This dissertation work suggests new evidence about sex-specific genetic influences on the human metabolome and reports novel genetic variants that were not previously identified when gene-sex interaction effects were omitted in previous studies. Additionally, this dissertation provides insights into the power and desired sample size in conducting rare variant G×E interaction studies under various scenarios, which facilitate the understanding of current G×E interaction results and show promise for future large-scale studies utilizing G×E interactions to investigate the genetic and environmental factors of disease etiology.

APPENDICES

Appendix A. supplemental materials for Chapter 2

Supplemental Table 1. List of 271 metabolites and transformation methods applied

Metabolites	Super_Pathway	Sub_Pathway	Platform	Transformation
alanine	Amino Acid	Alanine and Aspartate Metabolism	LC/MS Pos	rank based inverse normal
asparagine	Amino Acid	Alanine and Aspartate Metabolism	LC/MS Pos	rank based inverse normal
aspartate	Amino Acid	Alanine and Aspartate Metabolism	LC/MS Polar	rank based inverse normal
N-acetylalanine	Amino Acid	Alanine and Aspartate Metabolism	LC/MS Neg	natural log
creatine	Amino Acid	Creatine Metabolism	LC/MS Pos	rank based inverse normal
creatinine	Amino Acid	Creatine Metabolism	LC/MS Pos	rank based inverse normal
glutamate	Amino Acid	Glutamate Metabolism	LC/MS Pos	natural log
pyroglutamine	Amino Acid	Glutamate Metabolism	LC/MS Pos	rank based inverse normal
5-oxoproline	Amino Acid	Glutathione Metabolism	LC/MS Neg	not transformed
betaine	Amino Acid	Glycine, Serine and Threonine Metabolism	LC/MS Pos	not transformed
dimethylglycine	Amino Acid	Glycine, Serine and Threonine Metabolism	LC/MS Pos	natural log
glycine	Amino Acid	Glycine, Serine and Threonine Metabolism	LC/MS Pos	natural log
N-acetyl glycine	Amino Acid	Glycine, Serine and Threonine Metabolism	LC/MS Neg	rank based inverse normal
N-acetylthreonine	Amino Acid	Glycine, Serine and Threonine Metabolism	LC/MS Neg	natural log
serine	Amino Acid	Glycine, Serine and Threonine Metabolism	LC/MS Pos	rank based inverse normal
threonine	Amino Acid	Glycine, Serine and Threonine Metabolism	LC/MS Pos	rank based inverse normal
4-guanidinobutanoate	Amino Acid	Guanidino and Acetamido Metabolism	LC/MS Pos	natural log
3-methylhistidine	Amino Acid	Histidine Metabolism	LC/MS Neg	rank based inverse normal
histidine	Amino Acid	Histidine Metabolism	LC/MS Neg	rank based inverse normal
N-acetyl-1-methylhistidine	Amino Acid	Histidine Metabolism	LC/MS Pos	natural log
trans-urocanate	Amino Acid	Histidine Metabolism	LC/MS Pos	natural log
2-methylbutyrylcarnitine (C5)	Amino Acid	Leucine, Isoleucine and Valine Metabolism	LC/MS Pos	natural log
3-hydroxyisobutyrate	Amino Acid	Leucine, Isoleucine and Valine Metabolism	LC/MS Polar	natural log
3-methyl-2-oxovalerate	Amino Acid	Leucine, Isoleucine and Valine Metabolism	LC/MS Neg	rank based inverse normal
alpha-hydroxyisovalerate	Amino Acid	Leucine, Isoleucine and Valine Metabolism	LC/MS Neg	rank based inverse normal
beta-hydroxyisovalerate	Amino Acid	Leucine, Isoleucine and Valine Metabolism	LC/MS Neg	rank based inverse normal
beta-hydroxyisovalerylcarnitine	Amino Acid	Leucine, Isoleucine and Valine Metabolism	LC/MS Pos	natural log
isobutyrylcarnitine	Amino Acid	Leucine, Isoleucine and Valine Metabolism	LC/MS Pos	natural log
isoleucine	Amino Acid	Leucine, Isoleucine and Valine Metabolism	LC/MS Pos	natural log
isovalerate	Amino Acid	Leucine, Isoleucine and Valine Metabolism	LC/MS Neg	rank based inverse normal
isovalerylcarnitine	Amino Acid	Leucine, Isoleucine and Valine Metabolism	LC/MS Pos	natural log
leucine	Amino Acid	Leucine, Isoleucine and Valine Metabolism	LC/MS Pos	natural log

tiglyl carnitine	Amino Acid	Leucine, Isoleucine and Valine Metabolism	LC/MS Pos	natural log
valine	Amino Acid	Leucine, Isoleucine and Valine Metabolism	LC/MS Pos	natural log
glutarate (pentanedioate)	Amino Acid	Lysine Metabolism	LC/MS Polar	rank based inverse normal
glutaryl carnitine (C5)	Amino Acid	Lysine Metabolism	LC/MS Pos	rank based inverse normal
lysine	Amino Acid	Lysine Metabolism	LC/MS Polar	rank based inverse normal
N6-acetyllysine	Amino Acid	Lysine Metabolism	LC/MS Pos	rank based inverse normal
pipecolate	Amino Acid	Lysine Metabolism	LC/MS Pos	rank based inverse normal
2-aminobutyrate	Amino Acid	Methionine, Cysteine, SAM and Taurine Metabolism	LC/MS Pos	rank based inverse normal
2-hydroxybutyrate (AHB)	Amino Acid	Methionine, Cysteine, SAM and Taurine Metabolism	GC/MS	natural log
methionine sulfoxide	Amino Acid	Methionine, Cysteine, SAM and Taurine Metabolism	LC/MS Polar	rank based inverse normal
S-methylcysteine	Amino Acid	Methionine, Cysteine, SAM and Taurine Metabolism	LC/MS Neg	rank based inverse normal
3-(4-hydroxyphenyl)lactate	Amino Acid	Phenylalanine and Tyrosine Metabolism	LC/MS Neg	natural log
3-methoxytyrosine	Amino Acid	Phenylalanine and Tyrosine Metabolism	LC/MS Pos	rank based inverse normal
3-phenylpropionate (hydrocinnamate)	Amino Acid	Phenylalanine and Tyrosine Metabolism	LC/MS Neg	natural log
N-acetylphenylalanine	Amino Acid	Phenylalanine and Tyrosine Metabolism	LC/MS Neg	rank based inverse normal
o-cresol sulfate	Amino Acid	Phenylalanine and Tyrosine Metabolism	LC/MS Neg	rank based inverse normal
p-cresol sulfate	Amino Acid	Phenylalanine and Tyrosine Metabolism	LC/MS Neg	rank based inverse normal
phenol sulfate	Amino Acid	Phenylalanine and Tyrosine Metabolism	LC/MS Neg	natural log
phenylacetate	Amino Acid	Phenylalanine and Tyrosine Metabolism	LC/MS Neg	natural log
phenylacetylglutamine	Amino Acid	Phenylalanine and Tyrosine Metabolism	LC/MS Pos	natural log
phenylalanine	Amino Acid	Phenylalanine and Tyrosine Metabolism	LC/MS Pos	rank based inverse normal
phenyllactate (PLA)	Amino Acid	Phenylalanine and Tyrosine Metabolism	LC/MS Neg	rank based inverse normal
tyrosine	Amino Acid	Phenylalanine and Tyrosine Metabolism	LC/MS Pos	natural log
acisoga	Amino Acid	Polyamine Metabolism	LC/MS Pos	rank based inverse normal
3-indoxyl sulfate	Amino Acid	Tryptophan Metabolism	LC/MS Neg	natural log
anthranilate	Amino Acid	Tryptophan Metabolism	LC/MS Pos	rank based inverse normal
indoleacetate	Amino Acid	Tryptophan Metabolism	LC/MS Pos	rank based inverse normal
indolelactate	Amino Acid	Tryptophan Metabolism	LC/MS Pos	natural log
indolepropionate	Amino Acid	Tryptophan Metabolism	LC/MS Pos	rank based inverse normal
kynurenine	Amino Acid	Tryptophan Metabolism	LC/MS Pos	natural log
serotonin (5HT)	Amino Acid	Tryptophan Metabolism	LC/MS Pos	rank based inverse normal
tryptophan	Amino Acid	Tryptophan Metabolism	LC/MS Pos	rank based inverse normal
tryptophan betaine	Amino Acid	Tryptophan Metabolism	LC/MS Pos	rank based inverse normal
arginine	Amino Acid	Urea cycle; Arginine and Proline Metabolism	LC/MS Pos	rank based inverse normal
citrulline	Amino Acid	Urea cycle; Arginine and Proline Metabolism	LC/MS Pos	natural log

homocitrulline	Amino Acid	Urea cycle; Arginine and Proline Metabolism	LC/MS Pos	natural log
N-methylproline	Amino Acid	Urea cycle; Arginine and Proline Metabolism	LC/MS Pos	rank based inverse normal
ornithine	Amino Acid	Urea cycle; Arginine and Proline Metabolism	LC/MS Pos	rank based inverse normal
pro-hydroxy-pro	Amino Acid	Urea cycle; Arginine and Proline Metabolism	LC/MS Pos	natural log
proline	Amino Acid	Urea cycle; Arginine and Proline Metabolism	LC/MS Pos	natural log
trans-4-hydroxyproline	Amino Acid	Urea cycle; Arginine and Proline Metabolism	LC/MS Polar	natural log
urea	Amino Acid	Urea cycle; Arginine and Proline Metabolism	LC/MS Pos	natural log
erythronate	Carbohydrate	Aminosugar Metabolism	LC/MS Polar	natural log
glucuronate	Carbohydrate	Aminosugar Metabolism	LC/MS Polar	rank based inverse normal
trehalose	Carbohydrate	Disaccharides and Oligosaccharides	GC/MS	natural log
mannitol	Carbohydrate	Fructose, Mannose and Galactose Metabolism	GC/MS	rank based inverse normal
mannose	Carbohydrate	Fructose, Mannose and Galactose Metabolism	LC/MS Polar	rank based inverse normal
1,5-anhydroglucitol (1,5-AG)	Carbohydrate	Glycolysis, Gluconeogenesis, and Pyruvate Metabolism	LC/MS Neg	rank based inverse normal
glucose	Carbohydrate	Glycolysis, Gluconeogenesis, and Pyruvate Metabolism	LC/MS Polar	rank based inverse normal
glycerate	Carbohydrate	Glycolysis, Gluconeogenesis, and Pyruvate Metabolism	LC/MS Polar	rank based inverse normal
lactate	Carbohydrate	Glycolysis, Gluconeogenesis, and Pyruvate Metabolism	LC/MS Neg	natural log
arabinose	Carbohydrate	Pentose Metabolism	GC/MS	rank based inverse normal
threitol	Carbohydrate	Pentose Metabolism	GC/MS	natural log
arabonate	Cofactors and Vitamins	Ascorbate and Aldarate Metabolism	GC/MS	rank based inverse normal
threonate	Cofactors and Vitamins	Ascorbate and Aldarate Metabolism	LC/MS Polar	not transformed
bilirubin (E,E)	Cofactors and Vitamins	Hemoglobin and Porphyrin Metabolism	LC/MS Neg	rank based inverse normal
N1-Methyl-2-pyridone-5-carboxamide	Cofactors and Vitamins	Nicotinate and Nicotinamide Metabolism	LC/MS Pos	rank based inverse normal
pantothenate	Cofactors and Vitamins	Pantothenate and CoA Metabolism	LC/MS Pos	rank based inverse normal
alpha-tocopherol	Cofactors and Vitamins	Tocopherol Metabolism	GC/MS	natural log
gamma-tocopherol	Cofactors and Vitamins	Tocopherol Metabolism	GC/MS	rank based inverse normal
pyridoxate	Cofactors and Vitamins	Vitamin B6 Metabolism	LC/MS Neg	rank based inverse normal
phosphate	Energy	Oxidative Phosphorylation	GC/MS	natural log
citrate	Energy	TCA Cycle	GC/MS	rank based inverse normal
malate	Energy	TCA Cycle	LC/MS Neg	rank based inverse normal
succinate	Energy	TCA Cycle	LC/MS Polar	rank based inverse normal
succinylcarnitine	Energy	TCA Cycle	LC/MS Pos	rank based inverse normal
carnitine	Lipid	Carnitine Metabolism	LC/MS Pos	rank based inverse normal
deoxycarnitine	Lipid	Carnitine Metabolism	LC/MS Pos	natural log
5-HETE	Lipid	Eicosanoid	LC/MS Neg	rank based inverse normal

propionylcarnitine	Lipid	Fatty Acid Metabolism (also BCAA Metabolism)	LC/MS Pos	rank based inverse normal
acetylarnitine	Lipid	Fatty Acid Metabolism(Acyl Carnitine)	LC/MS Pos	rank based inverse normal
cis-4-decenoyl carnitine	Lipid	Fatty Acid Metabolism(Acyl Carnitine)	LC/MS Pos	natural log
decanoylcarnitine	Lipid	Fatty Acid Metabolism(Acyl Carnitine)	LC/MS Pos	natural log
hexanoylcarnitine	Lipid	Fatty Acid Metabolism(Acyl Carnitine)	LC/MS Pos	rank based inverse normal
hydroxybutyrylcarnitine	Lipid	Fatty Acid Metabolism(Acyl Carnitine)	LC/MS Pos	natural log
laurylcarnitine	Lipid	Fatty Acid Metabolism(Acyl Carnitine)	LC/MS Pos	natural log
octanoylcarnitine	Lipid	Fatty Acid Metabolism(Acyl Carnitine)	LC/MS Pos	natural log
oleoylcarnitine	Lipid	Fatty Acid Metabolism(Acyl Carnitine)	LC/MS Pos	natural log
palmitoylcarnitine	Lipid	Fatty Acid Metabolism(Acyl Carnitine)	LC/MS Pos	natural log
stearoylcarnitine	Lipid	Fatty Acid Metabolism(Acyl Carnitine)	LC/MS Pos	natural log
2-aminoheptanoate	Lipid	Fatty Acid, Amino	LC/MS Pos	rank based inverse normal
2-aminooctanoate	Lipid	Fatty Acid, Amino	LC/MS Pos	rank based inverse normal
2-hydroxyglutarate	Lipid	Fatty Acid, Dicarboxylate	GC/MS	rank based inverse normal
3-carboxy-4-methyl-5-propyl-2-furanpropanoate (CMPF)	Lipid	Fatty Acid, Dicarboxylate	LC/MS Neg	rank based inverse normal
adipate	Lipid	Fatty Acid, Dicarboxylate	GC/MS	rank based inverse normal
azelate (nonanedioate)	Lipid	Fatty Acid, Dicarboxylate	LC/MS Polar	rank based inverse normal
dodecanedioate	Lipid	Fatty Acid, Dicarboxylate	LC/MS Neg	rank based inverse normal
eicosanodioate	Lipid	Fatty Acid, Dicarboxylate	LC/MS Neg	natural log
hexadecanedioate	Lipid	Fatty Acid, Dicarboxylate	LC/MS Neg	natural log
octadecanedioate	Lipid	Fatty Acid, Dicarboxylate	LC/MS Neg	rank based inverse normal
sebacate (decanedioate)	Lipid	Fatty Acid, Dicarboxylate	LC/MS Neg	rank based inverse normal
suberate (octanedioate)	Lipid	Fatty Acid, Dicarboxylate	LC/MS Neg	rank based inverse normal
tetradecanedioate	Lipid	Fatty Acid, Dicarboxylate	LC/MS Neg	rank based inverse normal
undecanedioate	Lipid	Fatty Acid, Dicarboxylate	LC/MS Neg	rank based inverse normal
13-HODE + 9-HODE	Lipid	Fatty Acid, Monohydroxy	LC/MS Neg	rank based inverse normal
2-hydroxypalmitate	Lipid	Fatty Acid, Monohydroxy	LC/MS Neg	rank based inverse normal
2-hydroxystearate	Lipid	Fatty Acid, Monohydroxy	LC/MS Neg	natural log
glycerol	Lipid	Glycerolipid Metabolism	LC/MS Neg	rank based inverse normal
glycerol 3-phosphate (G3P)	Lipid	Glycerolipid Metabolism	GC/MS	rank based inverse normal
inositol 1-phosphate (I1P)	Lipid	Inositol Metabolism	GC/MS	natural log
myo-inositol	Lipid	Inositol Metabolism	LC/MS Polar	natural log
scyllo-inositol	Lipid	Inositol Metabolism	GC/MS	natural log
3-hydroxybutyrate (BHBA)	Lipid	Ketone Bodies	LC/MS Polar	rank based inverse normal
10-heptadecenoate (17:1n7)	Lipid	Long Chain Fatty Acid	LC/MS Neg	natural log
10-nonadecenoate (19:1n9)	Lipid	Long Chain Fatty Acid	LC/MS Neg	natural log
eicosenoate (20:1n9 or 11)	Lipid	Long Chain Fatty Acid	LC/MS Neg	natural log

margarate (17:0)	Lipid	Long Chain Fatty Acid	LC/MS Neg	natural log
myristate (14:0)	Lipid	Long Chain Fatty Acid	LC/MS Neg	natural log
myristoleate (14:1n5)	Lipid	Long Chain Fatty Acid	LC/MS Neg	natural log
nonadecanoate (19:0)	Lipid	Long Chain Fatty Acid	LC/MS Neg	rank based inverse normal
oleate (18:1n9)	Lipid	Long Chain Fatty Acid	LC/MS Neg	not transformed
palmitate (16:0)	Lipid	Long Chain Fatty Acid	LC/MS Neg	not transformed
palmitoleate (16:1n7)	Lipid	Long Chain Fatty Acid	LC/MS Neg	natural log
stearate (18:0)	Lipid	Long Chain Fatty Acid	LC/MS Neg	rank based inverse normal
1-arachidonoylglycerophosphocholine (20:4n6)	Lipid	Lysolipid	LC/MS Pos	rank based inverse normal
1-arachidonoylglycerophosphoethanolamine	Lipid	Lysolipid	LC/MS Neg	natural log
1-arachidonoylglycerophosphoinositol	Lipid	Lysolipid	LC/MS Neg	natural log
1-docosaheptaenoylglycerophosphocholine (22:6n3)	Lipid	Lysolipid	LC/MS Pos	natural log
1-docosaheptaenoylglycerophosphoethanolamine	Lipid	Lysolipid	LC/MS Neg	rank based inverse normal
1-docosapentaenoylglycerophosphocholine (22:5n3)	Lipid	Lysolipid	LC/MS Pos	natural log
1-linoleoylglycerophosphoethanolamine	Lipid	Lysolipid	LC/MS Neg	natural log
1-oleoylglycerophosphocholine (18:1)	Lipid	Lysolipid	LC/MS Pos	rank based inverse normal
1-palmitoleoylglycerophosphocholine (16:1)	Lipid	Lysolipid	LC/MS Pos	natural log
1-palmitoylglycerophosphoinositol	Lipid	Lysolipid	LC/MS Neg	natural log
1-palmitoylplasmenylethanolamine	Lipid	Lysolipid	LC/MS Neg	natural log
1-stearoylglycerophosphoethanolamine	Lipid	Lysolipid	LC/MS Neg	natural log
2-arachidonoylglycerophosphocholine	Lipid	Lysolipid	LC/MS Neg	rank based inverse normal
2-linoleoylglycerophosphoethanolamine	Lipid	Lysolipid	LC/MS Neg	rank based inverse normal
2-myristoylglycerophosphocholine	Lipid	Lysolipid	LC/MS Pos	natural log
5-dodecenoate (12:1n7)	Lipid	Medium Chain Fatty Acid	LC/MS Neg	natural log
caprate (10:0)	Lipid	Medium Chain Fatty Acid	LC/MS Neg	rank based inverse normal
caproate (6:0)	Lipid	Medium Chain Fatty Acid	LC/MS Neg	rank based inverse normal
heptanoate (7:0)	Lipid	Medium Chain Fatty Acid	LC/MS Neg	rank based inverse normal
laurate (12:0)	Lipid	Medium Chain Fatty Acid	LC/MS Neg	rank based inverse normal
pelargonate (9:0)	Lipid	Medium Chain Fatty Acid	LC/MS Neg	rank based inverse normal
1-oleoylglycerol (1-monoolein)	Lipid	Monoacylglycerol	LC/MS Neg	rank based inverse normal
1-stearoylglycerol (1-monostearin)	Lipid	Monoacylglycerol	GC/MS	rank based inverse normal
choline	Lipid	Phospholipid Metabolism	LC/MS Pos	rank based inverse normal

glycerophosphorylcholine (GPC)	Lipid	Phospholipid Metabolism	LC/MS Pos	natural log
adrenate (22:4n6)	Lipid	Polyunsaturated Fatty Acid (n3 and n6)	LC/MS Neg	rank based inverse normal
arachidonate (20:4n6)	Lipid	Polyunsaturated Fatty Acid (n3 and n6)	LC/MS Neg	rank based inverse normal
dihomo-linoleate (20:2n6)	Lipid	Polyunsaturated Fatty Acid (n3 and n6)	LC/MS Neg	natural log
dihomo-linolenate (20:3n3 or n6)	Lipid	Polyunsaturated Fatty Acid (n3 and n6)	LC/MS Neg	natural log
docosahexaenoate (DHA; 22:6n3)	Lipid	Polyunsaturated Fatty Acid (n3 and n6)	LC/MS Neg	rank based inverse normal
docosapentaenoate (n3 DPA; 22:5n3)	Lipid	Polyunsaturated Fatty Acid (n3 and n6)	LC/MS Neg	natural log
docosapentaenoate (n6 DPA; 22:5n6)	Lipid	Polyunsaturated Fatty Acid (n3 and n6)	LC/MS Neg	rank based inverse normal
eicosapentaenoate (EPA; 20:5n3)	Lipid	Polyunsaturated Fatty Acid (n3 and n6)	LC/MS Neg	natural log
linoleate (18:2n6)	Lipid	Polyunsaturated Fatty Acid (n3 and n6)	LC/MS Neg	natural log
linolenate [alpha or gamma; (18:3n3 or 6)]	Lipid	Polyunsaturated Fatty Acid (n3 and n6)	LC/MS Neg	natural log
stearidonate (18:4n3)	Lipid	Polyunsaturated Fatty Acid (n3 and n6)	LC/MS Neg	rank based inverse normal
glycochenodeoxycholate	Lipid	Primary Bile Acid Metabolism	LC/MS Pos	natural log
glycocholate	Lipid	Primary Bile Acid Metabolism	LC/MS Neg	natural log
taurochenodeoxycholate	Lipid	Primary Bile Acid Metabolism	LC/MS Neg	rank based inverse normal
glycochenolate sulfate	Lipid	Secondary Bile Acid Metabolism	LC/MS Neg	rank based inverse normal
glycodeoxycholate	Lipid	Secondary Bile Acid Metabolism	LC/MS Pos	natural log
glycolithocholate sulfate	Lipid	Secondary Bile Acid Metabolism	LC/MS Neg	rank based inverse normal
glycoursodeoxycholate	Lipid	Secondary Bile Acid Metabolism	LC/MS Neg	natural log
taurochenolate sulfate	Lipid	Secondary Bile Acid Metabolism	LC/MS Neg	natural log
tauroolithocholate 3-sulfate	Lipid	Secondary Bile Acid Metabolism	LC/MS Neg	rank based inverse normal
21-hydroxypregnenolone disulfate	Lipid	Steroid	LC/MS Neg	rank based inverse normal
4-androsten-3beta,17beta-diol disulfate (1)	Lipid	Steroid	LC/MS Neg	rank based inverse normal
4-androsten-3beta,17beta-diol disulfate (2)	Lipid	Steroid	LC/MS Neg	rank based inverse normal
5alpha-androstan-3beta,17beta-diol disulfate	Lipid	Steroid	LC/MS Neg	rank based inverse normal
5alpha-pregnan-3beta,20alpha-diol disulfate	Lipid	Steroid	LC/MS Neg	natural log
androsterone sulfate	Lipid	Steroid	LC/MS Neg	rank based inverse normal
cortisol	Lipid	Steroid	LC/MS Pos	natural log
cortisone	Lipid	Steroid	LC/MS Pos	natural log
dehydroisoandrosterone sulfate (DHEA-S)	Lipid	Steroid	LC/MS Neg	natural log
epiandrosterone sulfate	Lipid	Steroid	LC/MS Neg	rank based inverse normal
pregn steroid monosulfate	Lipid	Steroid	LC/MS Neg	natural log
pregnen-diol disulfate	Lipid	Steroid	LC/MS Neg	natural log
7-alpha-hydroxy-3-oxo-4-cholestenoate (7-Hoca)	Lipid	Sterol	LC/MS Neg	rank based inverse normal
cholesterol	Lipid	Sterol	GC/MS	rank based inverse normal
hypoxanthine	Nucleotide	Purine Metabolism, (Hypo)Xanthine/Inosine containing	LC/MS Neg	rank based inverse normal

urate	Nucleotide	Purine Metabolism, (Hypo)Xanthine/Inosine containing	LC/MS Neg	rank based inverse normal
xanthine	Nucleotide	Purine Metabolism, (Hypo)Xanthine/Inosine containing	LC/MS Pos	rank based inverse normal
N1-methyladenosine	Nucleotide	Purine Metabolism, Adenine containing	LC/MS Pos	rank based inverse normal
guanosine	Nucleotide	Purine Metabolism, Guanine containing	LC/MS Pos	rank based inverse normal
5-methyluridine (ribothymidine)	Nucleotide	Pyrimidine Metabolism, Uracil containing	LC/MS Neg	rank based inverse normal
N-acetyl-beta-alanine	Nucleotide	Pyrimidine Metabolism, Uracil containing	LC/MS Pos	natural log
pseudouridine	Nucleotide	Pyrimidine Metabolism, Uracil containing	LC/MS Pos	rank based inverse normal
uridine	Nucleotide	Pyrimidine Metabolism, Uracil containing	LC/MS Neg	rank based inverse normal
alanylleucine	Peptide	Dipeptide	LC/MS Pos	rank based inverse normal
arginylleucine	Peptide	Dipeptide	LC/MS Pos	natural log
arginylphenylalanine	Peptide	Dipeptide	LC/MS Pos	rank based inverse normal
glycylleucine	Peptide	Dipeptide	LC/MS Pos	rank based inverse normal
glycylphenylalanine	Peptide	Dipeptide	LC/MS Pos	natural log
glycylvaline	Peptide	Dipeptide	LC/MS Pos	rank based inverse normal
histidylleucine	Peptide	Dipeptide	LC/MS Pos	natural log
leucylalanine	Peptide	Dipeptide	LC/MS Pos	rank based inverse normal
leucylasparagine	Peptide	Dipeptide	LC/MS Pos	rank based inverse normal
leucylglycine	Peptide	Dipeptide	LC/MS Pos	natural log
leucylleucine	Peptide	Dipeptide	LC/MS Pos	rank based inverse normal
leucylserine	Peptide	Dipeptide	LC/MS Pos	natural log
phenylalanylglutamate	Peptide	Dipeptide	LC/MS Pos	rank based inverse normal
phenylalanylleucine	Peptide	Dipeptide	LC/MS Pos	rank based inverse normal
phenylalanylphenylalanine	Peptide	Dipeptide	LC/MS Pos	rank based inverse normal
phenylalanylserine	Peptide	Dipeptide	LC/MS Pos	rank based inverse normal
pyroglutamylglycine	Peptide	Dipeptide	LC/MS Neg	natural log
serylleucine	Peptide	Dipeptide	LC/MS Pos	rank based inverse normal
N-acetylcarnosine	Peptide	Dipeptide Derivative	LC/MS Pos	natural log
DSGEGDFXAEGGGVR	Peptide	Fibrinogen Cleavage Peptide	LC/MS Pos	rank based inverse normal
gamma-glutamylalanine	Peptide	Gamma-glutamyl Amino Acid	LC/MS Pos	rank based inverse normal
gamma-glutamylglutamate	Peptide	Gamma-glutamyl Amino Acid	LC/MS Pos	rank based inverse normal
gamma-glutamylisoleucine	Peptide	Gamma-glutamyl Amino Acid	LC/MS Pos	rank based inverse normal
gamma-glutamylleucine	Peptide	Gamma-glutamyl Amino Acid	LC/MS Pos	rank based inverse normal
gamma-glutamylphenylalanine	Peptide	Gamma-glutamyl Amino Acid	LC/MS Pos	rank based inverse normal
gamma-glutamylthreonine	Peptide	Gamma-glutamyl Amino Acid	LC/MS Pos	rank based inverse normal
gamma-glutamyltyrosine	Peptide	Gamma-glutamyl Amino Acid	LC/MS Pos	rank based inverse normal
gamma-glutamylvaline	Peptide	Gamma-glutamyl Amino Acid	LC/MS Pos	rank based inverse normal
HWESASLLR	Peptide	Polypeptide	LC/MS Pos	rank based inverse normal

2-hydroxyhippurate (salicylurate)	Xenobiotics	Benzoate Metabolism	LC/MS Neg	rank based inverse normal
3-methyl catechol sulfate (1)	Xenobiotics	Benzoate Metabolism	LC/MS Neg	natural log
4-hydroxyhippurate	Xenobiotics	Benzoate Metabolism	LC/MS Neg	natural log
4-methylcatechol sulfate	Xenobiotics	Benzoate Metabolism	LC/MS Neg	rank based inverse normal
4-vinylphenol sulfate	Xenobiotics	Benzoate Metabolism	LC/MS Neg	rank based inverse normal
catechol sulfate	Xenobiotics	Benzoate Metabolism	LC/MS Neg	rank based inverse normal
hippurate	Xenobiotics	Benzoate Metabolism	LC/MS Neg	rank based inverse normal
methyl-4-hydroxybenzoate sulfate	Xenobiotics	Benzoate Metabolism	LC/MS Neg	rank based inverse normal
O-methylcatechol sulfate	Xenobiotics	Benzoate Metabolism	LC/MS Neg	rank based inverse normal
propyl 4-hydroxybenzoate sulfate	Xenobiotics	Benzoate Metabolism	LC/MS Neg	rank based inverse normal
1,2-propanediol	Xenobiotics	Chemical	GC/MS	rank based inverse normal
2-hydroxyisobutyrate	Xenobiotics	Chemical	GC/MS	natural log
3-hydroxypyridine sulfate	Xenobiotics	Chemical	LC/MS Neg	rank based inverse normal
O-sulfo-L-tyrosine	Xenobiotics	Chemical	LC/MS Neg	natural log
salicylate	Xenobiotics	Drug	LC/MS Neg	rank based inverse normal
erythritol	Xenobiotics	Food Component/Plant	GC/MS	rank based inverse normal
gluconate	Xenobiotics	Food Component/Plant	LC/MS Polar	rank based inverse normal
homostachydrine	Xenobiotics	Food Component/Plant	LC/MS Pos	natural log
piperine	Xenobiotics	Food Component/Plant	LC/MS Pos	rank based inverse normal
stachydrine	Xenobiotics	Food Component/Plant	LC/MS Pos	rank based inverse normal
thymol sulfate	Xenobiotics	Food Component/Plant	LC/MS Neg	rank based inverse normal
1,7-dimethylurate	Xenobiotics	Xanthine Metabolism	LC/MS Neg	rank based inverse normal
1-methylurate	Xenobiotics	Xanthine Metabolism	LC/MS Pos	rank based inverse normal
5-acetylamino-6-amino-3-methyluracil	Xenobiotics	Xanthine Metabolism	LC/MS Neg	rank based inverse normal
caffeine	Xenobiotics	Xanthine Metabolism	LC/MS Pos	rank based inverse normal
paraxanthine	Xenobiotics	Xanthine Metabolism	LC/MS Pos	rank based inverse normal
theobromine	Xenobiotics	Xanthine Metabolism	LC/MS Pos	rank based inverse normal
theophylline	Xenobiotics	Xanthine Metabolism	LC/MS Pos	rank based inverse normal

Supplemental Table 2. Female common variants

traits	Name	Gene	maf.EA	beta.EA	se.EA	p.EA	pval.BH.EA	maf.AA	beta.AA	se.AA	p.AA	pval.BH.AA
acisoga	rs2294757	VNN1	0.38	-0.21	0.03	3.18E-15	5.72E-15	0.16	-0.17	0.03	9.33E-11	1.85E-05
alanylleucine	rs3733402	KLKB1	0.49	-0.24	0.03	1.82E-13	3.12E-13	0.25	-0.12	0.02	3.23E-07	3.13E-02
2-aminooctanoate	rs3813227	ALMS1	0.26	-0.32	0.03	2.48E-25	6.87E-25	0.24	-0.20	0.03	2.21E-13	6.27E-08
2-aminooctanoate	rs6546837	ALMS1	0.26	-0.33	0.03	4.25E-26	1.27E-25	0.24	-0.20	0.03	2.21E-13	6.27E-08
2-aminooctanoate	rs6546838	ALMS1	0.26	-0.33	0.03	6.34E-27	2.85E-26	0.25	-0.20	0.03	2.57E-14	8.51E-09
2-aminooctanoate	rs6546839	ALMS1	0.26	-0.33	0.03	1.38E-26	5.54E-26	0.24	-0.20	0.03	3.44E-13	8.80E-08
2-aminooctanoate	rs2056486	ALMS1	0.25	-0.33	0.03	2.48E-26	8.94E-26	0.25	-0.20	0.03	2.40E-14	8.51E-09
2-aminooctanoate	rs10193972	ALMS1	0.25	-0.33	0.03	3.19E-26	1.04E-25	0.25	-0.20	0.03	1.92E-14	7.61E-09
2-aminooctanoate	rs1052161	ALMS1	0.39	-0.23	0.03	1.21E-16	2.42E-16	0.33	-0.24	0.02	6.79E-23	2.70E-16
arginine	rs3733402	KLKB1	0.49	-0.06	0.01	2.73E-10	4.10E-10	0.25	-0.07	0.01	1.03E-07	1.24E-02
arginylphenylalanine	rs3733402	KLKB1	0.49	-0.17	0.03	1.81E-09	2.61E-09	0.25	-0.13	0.02	8.45E-09	1.20E-03
dimethylglycine	rs1805073	DMGDH	0.28	-0.07	0.03	0.007489	0.008987	0.48	-0.09	0.02	1.74E-07	1.87E-02
gamma-glutamylisoleucine	rs11107	FBXO7	0.37	-0.06	0.02	0.009588	0.011134	0.39	0.09	0.02	2.79E-07	2.84E-02
histidylleucine	rs3733402	KLKB1	0.49	-0.31	0.03	7.84E-20	1.66E-19	0.25	-0.17	0.02	3.55E-13	8.80E-08
leucylalanine	rs3733402	KLKB1	0.49	-0.17	0.03	5.77E-12	9.02E-12	0.25	-0.19	0.02	2.55E-15	1.12E-09
leucylalanine	rs4253301	KLKB1	0.13	-0.14	0.04	0.000245	0.000327	0.06	-0.22	0.04	9.78E-08	1.21E-02
leucylasparagine	rs3733402	KLKB1	0.49	-0.30	0.04	2.88E-12	4.72E-12	0.25	-0.21	0.03	3.39E-10	6.11E-05
leucylglycine	rs3733402	KLKB1	0.49	-0.26	0.03	6.54E-16	1.24E-15	0.25	-0.20	0.03	1.91E-12	4.21E-07
mannose	rs1260326	GCKR	0.42	-0.13	0.01	2.30E-21	5.51E-21	0.13	-0.12	0.02	1.18E-08	1.61E-03
N1-methyladenosine	rs11555566	ADA	0.07	0.17	0.04	7.55E-05	0.000105	0.08	0.14	0.02	2.24E-11	4.67E-06
N-acetyl-1-methylhistidine	rs3813227	ALMS1	0.26	0.44	0.03	5.11E-49	3.07E-48	0.24	0.18	0.02	1.95E-20	1.11E-14
N-acetyl-1-methylhistidine	rs2037814	ALMS1	0.13	-0.10	0.04	0.006627	0.008227	0.11	-0.15	0.03	3.30E-09	5.04E-04
N-acetyl-1-methylhistidine	rs6546837	ALMS1	0.26	0.44	0.03	5.12E-50	6.14E-49	0.24	0.18	0.02	1.95E-20	1.11E-14
N-acetyl-1-methylhistidine	rs6546838	ALMS1	0.26	0.44	0.03	2.43E-49	1.75E-48	0.25	0.18	0.02	4.07E-21	7.65E-15

N-acetyl-1-methylhistidine	rs6546839	ALMS1	0.26	0.44	0.03	2.08E-49	1.75E-48	0.24	0.18	0.02	9.64E-21	7.65E-15
N-acetyl-1-methylhistidine	rs3820700	ALMS1	0.14	-0.12	0.04	0.001535	0.001973	0.11	-0.16	0.03	1.43E-09	2.27E-04
N-acetyl-1-methylhistidine	rs2056486	ALMS1	0.25	0.44	0.03	3.84E-50	6.14E-49	0.25	0.18	0.02	8.27E-21	7.65E-15
N-acetyl-1-methylhistidine	rs10193972	ALMS1	0.25	0.44	0.03	2.49E-50	6.14E-49	0.25	0.18	0.02	7.26E-21	7.65E-15
N-acetyl-1-methylhistidine	rs1052161	ALMS1	0.39	0.29	0.03	5.11E-27	2.63E-26	0.33	0.15	0.02	2.72E-18	1.35E-12
N-acetyl-1-methylhistidine	rs2272051	DUSP11	0.35	0.27	0.03	5.62E-23	1.44E-22	0.47	0.12	0.02	1.24E-12	2.89E-07
succinylcarnitine	rs2729835	LACTB	0.32	0.15	0.02	5.93E-21	1.33E-20	0.41	0.07	0.01	1.23E-07	1.39E-02

Supplemental Table 3. Male common variants

traits	Name	Gene	maf.EA	beta.EA	se.EA	p.EA	pval.BH.EA	maf.AA	beta.AA	se.AA	p.AA	pval.BH.AA
aminooctanoate	rs3813227	ALMS1	0.24	-0.24	0.03	3.22E-13	5.10E-13	0.25	-0.19	0.03	4.09E-08	1.07E-02
aminooctanoate	rs6546837	ALMS1	0.24	-0.24	0.03	3.22E-13	5.10E-13	0.25	-0.19	0.03	4.83E-08	1.14E-02
aminooctanoate	rs6546838	ALMS1	0.24	-0.24	0.03	7.13E-13	9.68E-13	0.27	-0.20	0.03	8.07E-09	2.48E-03
aminooctanoate	rs6546839	ALMS1	0.24	-0.24	0.03	7.13E-13	9.68E-13	0.25	-0.19	0.03	4.17E-08	1.07E-02
aminooctanoate	rs2056486	ALMS1	0.24	-0.24	0.03	3.22E-13	5.10E-13	0.27	-0.20	0.03	8.07E-09	2.48E-03
aminooctanoate	rs10193972	ALMS1	0.24	-0.24	0.03	3.22E-13	5.10E-13	0.26	-0.20	0.03	8.06E-09	2.48E-03
aminooctanoate	rs1052161	ALMS1	0.38	-0.19	0.03	4.59E-10	5.81E-10	0.34	-0.25	0.03	1.33E-15	5.33E-09
aminooctanoate	rs1805074	DMGDH	0.30	-0.11	0.03	8.63E-05	9.11E-05	0.48	-0.12	0.02	6.65E-08	1.48E-02
aminooctanoate	rs1805073	DMGDH	0.30	-0.11	0.03	8.63E-05	9.11E-05	0.49	-0.12	0.02	4.30E-08	1.07E-02
N1-methyladenosine	rs11555566	ADA	0.06	0.22	0.05	9.71E-06	1.15E-05	0.07	0.18	0.03	2.09E-07	3.97E-02
N-acetyl-1-methylhistidine	rs3813227	ALMS1	0.24	0.49	0.03	7.86E-52	3.73E-51	0.25	0.17	0.02	2.57E-11	1.56E-05
N-acetyl-1-methylhistidine	rs6546837	ALMS1	0.24	0.49	0.03	7.86E-52	3.73E-51	0.25	0.17	0.02	3.12E-11	1.56E-05
N-acetyl-1-methylhistidine	rs6546838	ALMS1	0.24	0.48	0.03	2.32E-51	7.35E-51	0.27	0.17	0.02	3.49E-12	2.79E-06
N-acetyl-1-methylhistidine	rs6546839	ALMS1	0.24	0.48	0.03	2.32E-51	7.35E-51	0.25	0.17	0.03	3.01E-11	1.56E-05
N-acetyl-1-methylhistidine	rs2056486	ALMS1	0.24	0.49	0.03	7.86E-52	3.73E-51	0.27	0.17	0.02	3.49E-12	2.79E-06
N-acetyl-1-methylhistidine	rs10193972	ALMS1	0.24	0.49	0.03	7.86E-52	3.73E-51	0.26	0.17	0.02	3.35E-12	2.79E-06
N-acetyl-1-methylhistidine	rs1052161	ALMS1	0.38	0.28	0.03	6.47E-23	1.76E-22	0.34	0.17	0.02	3.67E-14	7.33E-08
N-acetyl-1-methylhistidine	rs2272051	DUSP11	0.35	0.24	0.03	4.30E-17	1.02E-16	0.49	0.12	0.02	5.88E-09	2.35E-03

Supplemental Table 4. Sex-specific burden test

				EA							AA						
sex	traits	gene	chr	N	nsnp	MAC	beta	se	p	Fdr-q	N	nsnp	MAC	beta	se	p	Fdr-q
male	dimethylglycine	DMGDH	5	701	6	26	0.64	0.10	1.48E-11	4.59E-10	720	10	26	0.44	0.07	1.73E-09	1.21E-03
male	N-acetylalanine	ACY1	3	701	3	15	0.23	0.04	6.00E-08	9.30E-07	720	5	81	0.15	0.02	1.63E-13	5.68E-07
female	N-acetylalanine	ACY1	3	827	3	17	0.23	0.04	8.22E-10	1.64E-08	1292	7	151	0.17	0.02	1.06E-27	2.03E-21
female	isobutyrylcarnitine	ACAD8	11	827	1	39	0.46	0.09	1.44E-07	1.44E-06	1292	3	90	0.32	0.05	3.25E-10	2.07E-04
female	indolelactate	CCBL1	9	827	3	17	0.29	0.07	2.41E-05	1.61E-04	1292	9	36	0.36	0.05	1.65E-12	2.10E-06
female	dimethylglycine	DMGDH	5	827	6	39	0.28	0.08	0.000389	1.94E-03	1292	11	66	0.58	0.05	2.36E-29	9.02E-23
female	N-acetylthreonine	ACY1	3	827	3	17	0.23	0.08	0.001936	6.74E-03	1292	7	151	0.13	0.02	2.77E-08	8.83E-03
female	phenyllactate (PLA)	CCBL1	9	827	3	17	0.23	0.07	0.002022	6.74E-03	1292	9	36	0.29	0.05	1.28E-08	5.46E-03
female	3-(4-hydroxyphenyl)lactate	CCBL1	9	827	3	17	0.23	0.08	0.002634	7.52E-03	1292	9	36	0.29	0.05	2.33E-08	8.09E-03

Supplemental Table 5. Z test for sex-specific burden test

			EA				AA				
traits	Gene	chr	cmafUsed.male	cmafUsed.female	Z.tst	p.val	cmafUsed.male.	cmafUsed.female	Z.tst	p.val	FDR-q
leucylleucine	NMT2	10	0.009	0.007	-0.66	0.51	0.003	0.004	-5.45	4.97E-08	0.03
10-nonadecenoate (19:1n9)	AGER	6	0.059	0.053	-1.50	0.13	0.051	0.040	5.66	1.56E-08	0.03
phenylalanylphenylalanine	BAG4	8	0.008	0.007	-1.22	0.22	0.003	0.004	-5.51	3.59E-08	0.03
phenylalanylphenylalanine	SPATA17	1	0.027	0.026	-0.61	0.54	0.019	0.018	-5.60	2.13E-08	0.03
S-methylcysteine	IKZF4	12	0.010	0.008	0.10	0.92	0.003	0.008	-5.57	2.54E-08	0.03

Appendix B. supplemental materials for Chapter 3

Supplemental Table 1. List of 271 metabolites and transformation methods applied

Metabolites	Super_Pathway	Sub_Pathway	Platform	Transformation
alanine	Amino Acid	Alanine and Aspartate Metabolism	LC/MS Pos	rank based inverse normal
asparagine	Amino Acid	Alanine and Aspartate Metabolism	LC/MS Pos	rank based inverse normal
aspartate	Amino Acid	Alanine and Aspartate Metabolism	LC/MS Polar	rank based inverse normal
N-acetylalanine	Amino Acid	Alanine and Aspartate Metabolism	LC/MS Neg	natural log
creatine	Amino Acid	Creatine Metabolism	LC/MS Pos	rank based inverse normal
creatinine	Amino Acid	Creatine Metabolism	LC/MS Pos	rank based inverse normal
glutamate	Amino Acid	Glutamate Metabolism	LC/MS Pos	natural log
pyroglutamine	Amino Acid	Glutamate Metabolism	LC/MS Pos	rank based inverse normal
5-oxoproline	Amino Acid	Glutathione Metabolism	LC/MS Neg	not transformed
betaine	Amino Acid	Glycine, Serine and Threonine Metabolism	LC/MS Pos	not transformed
dimethylglycine	Amino Acid	Glycine, Serine and Threonine Metabolism	LC/MS Pos	natural log
glycine	Amino Acid	Glycine, Serine and Threonine Metabolism	LC/MS Pos	natural log
N-acetyl glycine	Amino Acid	Glycine, Serine and Threonine Metabolism	LC/MS Neg	rank based inverse normal
N-acetylthreonine	Amino Acid	Glycine, Serine and Threonine Metabolism	LC/MS Neg	natural log
serine	Amino Acid	Glycine, Serine and Threonine Metabolism	LC/MS Pos	rank based inverse normal
threonine	Amino Acid	Glycine, Serine and Threonine Metabolism	LC/MS Pos	rank based inverse normal
4-guanidinobutanoate	Amino Acid	Guanidino and Acetamido Metabolism	LC/MS Pos	natural log
3-methylhistidine	Amino Acid	Histidine Metabolism	LC/MS Neg	rank based inverse normal
histidine	Amino Acid	Histidine Metabolism	LC/MS Neg	rank based inverse normal
N-acetyl-1-methylhistidine	Amino Acid	Histidine Metabolism	LC/MS Pos	natural log
trans-uocanate	Amino Acid	Histidine Metabolism	LC/MS Pos	natural log
2-methylbutyrylcarnitine (C5)	Amino Acid	Leucine, Isoleucine and Valine Metabolism	LC/MS Pos	natural log
3-hydroxyisobutyrate	Amino Acid	Leucine, Isoleucine and Valine Metabolism	LC/MS Polar	natural log
3-methyl-2-oxovalerate	Amino Acid	Leucine, Isoleucine and Valine Metabolism	LC/MS Neg	rank based inverse normal
alpha-hydroxyisovalerate	Amino Acid	Leucine, Isoleucine and Valine Metabolism	LC/MS Neg	rank based inverse normal
beta-hydroxyisovalerate	Amino Acid	Leucine, Isoleucine and Valine Metabolism	LC/MS Neg	rank based inverse normal
beta-hydroxyisovaleroylcarnitine	Amino Acid	Leucine, Isoleucine and Valine Metabolism	LC/MS Pos	natural log
isobutyrylcarnitine	Amino Acid	Leucine, Isoleucine and Valine Metabolism	LC/MS Pos	natural log
isoleucine	Amino Acid	Leucine, Isoleucine and Valine Metabolism	LC/MS Pos	natural log
isovalerate	Amino Acid	Leucine, Isoleucine and Valine Metabolism	LC/MS Neg	rank based inverse normal
isovalerylcarnitine	Amino Acid	Leucine, Isoleucine and Valine Metabolism	LC/MS Pos	natural log
leucine	Amino Acid	Leucine, Isoleucine and Valine Metabolism	LC/MS Pos	natural log
tiglyl carnitine	Amino Acid	Leucine, Isoleucine and Valine Metabolism	LC/MS Pos	natural log
valine	Amino Acid	Leucine, Isoleucine and Valine Metabolism	LC/MS Pos	natural log

glutarate (pentanedioate)	Amino Acid	Lysine Metabolism	LC/MS Polar	rank based inverse normal
glutaryl carnitine (C5)	Amino Acid	Lysine Metabolism	LC/MS Pos	rank based inverse normal
lysine	Amino Acid	Lysine Metabolism	LC/MS Polar	rank based inverse normal
N6-acetyllysine	Amino Acid	Lysine Metabolism	LC/MS Pos	rank based inverse normal
pipecolate	Amino Acid	Lysine Metabolism	LC/MS Pos	rank based inverse normal
2-aminobutyrate	Amino Acid	Methionine, Cysteine, SAM and Taurine Metabolism	LC/MS Pos	rank based inverse normal
2-hydroxybutyrate (AHB)	Amino Acid	Methionine, Cysteine, SAM and Taurine Metabolism	GC/MS	natural log
methionine sulfoxide	Amino Acid	Methionine, Cysteine, SAM and Taurine Metabolism	LC/MS Polar	rank based inverse normal
S-methylcysteine	Amino Acid	Methionine, Cysteine, SAM and Taurine Metabolism	LC/MS Neg	rank based inverse normal
3-(4-hydroxyphenyl)lactate	Amino Acid	Phenylalanine and Tyrosine Metabolism	LC/MS Neg	natural log
3-methoxytyrosine	Amino Acid	Phenylalanine and Tyrosine Metabolism	LC/MS Pos	rank based inverse normal
3-phenylpropionate (hydrocinnamate)	Amino Acid	Phenylalanine and Tyrosine Metabolism	LC/MS Neg	natural log
N-acetylphenylalanine	Amino Acid	Phenylalanine and Tyrosine Metabolism	LC/MS Neg	rank based inverse normal
o-cresol sulfate	Amino Acid	Phenylalanine and Tyrosine Metabolism	LC/MS Neg	rank based inverse normal
p-cresol sulfate	Amino Acid	Phenylalanine and Tyrosine Metabolism	LC/MS Neg	rank based inverse normal
phenol sulfate	Amino Acid	Phenylalanine and Tyrosine Metabolism	LC/MS Neg	natural log
phenylacetate	Amino Acid	Phenylalanine and Tyrosine Metabolism	LC/MS Neg	natural log
phenylacetylglutamine	Amino Acid	Phenylalanine and Tyrosine Metabolism	LC/MS Pos	natural log
phenylalanine	Amino Acid	Phenylalanine and Tyrosine Metabolism	LC/MS Pos	rank based inverse normal
phenyllactate (PLA)	Amino Acid	Phenylalanine and Tyrosine Metabolism	LC/MS Neg	rank based inverse normal
tyrosine	Amino Acid	Phenylalanine and Tyrosine Metabolism	LC/MS Pos	natural log
acisoga	Amino Acid	Polyamine Metabolism	LC/MS Pos	rank based inverse normal
3-indoxyl sulfate	Amino Acid	Tryptophan Metabolism	LC/MS Neg	natural log
anthranilate	Amino Acid	Tryptophan Metabolism	LC/MS Pos	rank based inverse normal
indoleacetate	Amino Acid	Tryptophan Metabolism	LC/MS Pos	rank based inverse normal
indolelactate	Amino Acid	Tryptophan Metabolism	LC/MS Pos	natural log
indolepropionate	Amino Acid	Tryptophan Metabolism	LC/MS Pos	rank based inverse normal
kynurenine	Amino Acid	Tryptophan Metabolism	LC/MS Pos	natural log
serotonin (5HT)	Amino Acid	Tryptophan Metabolism	LC/MS Pos	rank based inverse normal
tryptophan	Amino Acid	Tryptophan Metabolism	LC/MS Pos	rank based inverse normal
tryptophan betaine	Amino Acid	Tryptophan Metabolism	LC/MS Pos	rank based inverse normal
arginine	Amino Acid	Urea cycle; Arginine and Proline Metabolism	LC/MS Pos	rank based inverse normal
citrulline	Amino Acid	Urea cycle; Arginine and Proline Metabolism	LC/MS Pos	natural log
homocitrulline	Amino Acid	Urea cycle; Arginine and Proline Metabolism	LC/MS Pos	natural log
N-methylproline	Amino Acid	Urea cycle; Arginine and Proline Metabolism	LC/MS Pos	rank based inverse normal

ornithine	Amino Acid	Urea cycle; Arginine and Proline Metabolism	LC/MS Pos	rank based inverse normal
pro-hydroxy-pro	Amino Acid	Urea cycle; Arginine and Proline Metabolism	LC/MS Pos	natural log
proline	Amino Acid	Urea cycle; Arginine and Proline Metabolism	LC/MS Pos	natural log
trans-4-hydroxyproline	Amino Acid	Urea cycle; Arginine and Proline Metabolism	LC/MS Polar	natural log
urea	Amino Acid	Urea cycle; Arginine and Proline Metabolism	LC/MS Pos	natural log
erythronate	Carbohydrate	Aminosugar Metabolism	LC/MS Polar	natural log
glucuronate	Carbohydrate	Aminosugar Metabolism	LC/MS Polar	rank based inverse normal
trehalose	Carbohydrate	Disaccharides and Oligosaccharides	GC/MS	natural log
mannitol	Carbohydrate	Fructose, Mannose and Galactose Metabolism	GC/MS	rank based inverse normal
mannose	Carbohydrate	Fructose, Mannose and Galactose Metabolism	LC/MS Polar	rank based inverse normal
1,5-anhydroglucitol (1,5-AG)	Carbohydrate	Glycolysis, Gluconeogenesis, and Pyruvate Metabolism	LC/MS Neg	rank based inverse normal
glucose	Carbohydrate	Glycolysis, Gluconeogenesis, and Pyruvate Metabolism	LC/MS Polar	rank based inverse normal
glycerate	Carbohydrate	Glycolysis, Gluconeogenesis, and Pyruvate Metabolism	LC/MS Polar	rank based inverse normal
lactate	Carbohydrate	Glycolysis, Gluconeogenesis, and Pyruvate Metabolism	LC/MS Neg	natural log
arabinose	Carbohydrate	Pentose Metabolism	GC/MS	rank based inverse normal
threitol	Carbohydrate	Pentose Metabolism	GC/MS	natural log
arabonate	Cofactors and Vitamins	Ascorbate and Aldarate Metabolism	GC/MS	rank based inverse normal
threonate	Cofactors and Vitamins	Ascorbate and Aldarate Metabolism	LC/MS Polar	not transformed
bilirubin (E,E)	Cofactors and Vitamins	Hemoglobin and Porphyrin Metabolism	LC/MS Neg	rank based inverse normal
N1-Methyl-2-pyridone-5-carboxamide	Cofactors and Vitamins	Nicotinate and Nicotinamide Metabolism	LC/MS Pos	rank based inverse normal
pantothenate	Cofactors and Vitamins	Pantothenate and CoA Metabolism	LC/MS Pos	rank based inverse normal
alpha-tocopherol	Cofactors and Vitamins	Tocopherol Metabolism	GC/MS	natural log
gamma-tocopherol	Cofactors and Vitamins	Tocopherol Metabolism	GC/MS	rank based inverse normal
pyridoxate	Cofactors and Vitamins	Vitamin B6 Metabolism	LC/MS Neg	rank based inverse normal
phosphate	Energy	Oxidative Phosphorylation	GC/MS	natural log
citrate	Energy	TCA Cycle	GC/MS	rank based inverse normal
malate	Energy	TCA Cycle	LC/MS Neg	rank based inverse normal
succinate	Energy	TCA Cycle	LC/MS Polar	rank based inverse normal
succinylcarnitine	Energy	TCA Cycle	LC/MS Pos	rank based inverse normal
carnitine	Lipid	Carnitine Metabolism	LC/MS Pos	rank based inverse normal
deoxycarnitine	Lipid	Carnitine Metabolism	LC/MS Pos	natural log
5-HETE	Lipid	Eicosanoid	LC/MS Neg	rank based inverse normal
propionylcarnitine	Lipid	Fatty Acid Metabolism (also BCAA Metabolism)	LC/MS Pos	rank based inverse normal
acetylcarnitine	Lipid	Fatty Acid Metabolism(Acyl Carnitine)	LC/MS Pos	rank based inverse normal

cis-4-decenoyl carnitine	Lipid	Fatty Acid Metabolism(Acyl Carnitine)	LC/MS Pos	natural log
decanoylcarnitine	Lipid	Fatty Acid Metabolism(Acyl Carnitine)	LC/MS Pos	natural log
hexanoylcarnitine	Lipid	Fatty Acid Metabolism(Acyl Carnitine)	LC/MS Pos	rank based inverse normal
hydroxybutyrylcarnitine	Lipid	Fatty Acid Metabolism(Acyl Carnitine)	LC/MS Pos	natural log
laurylcarnitine	Lipid	Fatty Acid Metabolism(Acyl Carnitine)	LC/MS Pos	natural log
octanoylcarnitine	Lipid	Fatty Acid Metabolism(Acyl Carnitine)	LC/MS Pos	natural log
oleoylcarnitine	Lipid	Fatty Acid Metabolism(Acyl Carnitine)	LC/MS Pos	natural log
palmitoylcarnitine	Lipid	Fatty Acid Metabolism(Acyl Carnitine)	LC/MS Pos	natural log
stearoylcarnitine	Lipid	Fatty Acid Metabolism(Acyl Carnitine)	LC/MS Pos	natural log
2-aminoheptanoate	Lipid	Fatty Acid, Amino	LC/MS Pos	rank based inverse normal
2-aminooctanoate	Lipid	Fatty Acid, Amino	LC/MS Pos	rank based inverse normal
2-hydroxyglutarate	Lipid	Fatty Acid, Dicarboxylate	GC/MS	rank based inverse normal
3-carboxy-4-methyl-5-propyl-2-furanpropanoate (CMPF)	Lipid	Fatty Acid, Dicarboxylate	LC/MS Neg	rank based inverse normal
adipate	Lipid	Fatty Acid, Dicarboxylate	GC/MS	rank based inverse normal
azelate (nonanedioate)	Lipid	Fatty Acid, Dicarboxylate	LC/MS Polar	rank based inverse normal
dodecanedioate	Lipid	Fatty Acid, Dicarboxylate	LC/MS Neg	rank based inverse normal
eicosanodioate	Lipid	Fatty Acid, Dicarboxylate	LC/MS Neg	natural log
hexadecanedioate	Lipid	Fatty Acid, Dicarboxylate	LC/MS Neg	natural log
octadecanedioate	Lipid	Fatty Acid, Dicarboxylate	LC/MS Neg	rank based inverse normal
sebacate (decanedioate)	Lipid	Fatty Acid, Dicarboxylate	LC/MS Neg	rank based inverse normal
suberate (octanedioate)	Lipid	Fatty Acid, Dicarboxylate	LC/MS Neg	rank based inverse normal
tetradecanedioate	Lipid	Fatty Acid, Dicarboxylate	LC/MS Neg	rank based inverse normal
undecanedioate	Lipid	Fatty Acid, Dicarboxylate	LC/MS Neg	rank based inverse normal
13-HODE + 9-HODE	Lipid	Fatty Acid, Monohydroxy	LC/MS Neg	rank based inverse normal
2-hydroxypalmitate	Lipid	Fatty Acid, Monohydroxy	LC/MS Neg	rank based inverse normal
2-hydroxystearate	Lipid	Fatty Acid, Monohydroxy	LC/MS Neg	natural log
glycerol	Lipid	Glycerolipid Metabolism	LC/MS Neg	rank based inverse normal
glycerol 3-phosphate (G3P)	Lipid	Glycerolipid Metabolism	GC/MS	rank based inverse normal
inositol 1-phosphate (I1P)	Lipid	Inositol Metabolism	GC/MS	natural log
myo-inositol	Lipid	Inositol Metabolism	LC/MS Polar	natural log
scyllo-inositol	Lipid	Inositol Metabolism	GC/MS	natural log
3-hydroxybutyrate (BHBA)	Lipid	Ketone Bodies	LC/MS Polar	rank based inverse normal
10-heptadecenoate (17:1n7)	Lipid	Long Chain Fatty Acid	LC/MS Neg	natural log
10-nonadecenoate (19:1n9)	Lipid	Long Chain Fatty Acid	LC/MS Neg	natural log
eicosenoate (20:1n9 or 11)	Lipid	Long Chain Fatty Acid	LC/MS Neg	natural log
margarate (17:0)	Lipid	Long Chain Fatty Acid	LC/MS Neg	natural log
myristate (14:0)	Lipid	Long Chain Fatty Acid	LC/MS Neg	natural log
myristoleate (14:1n5)	Lipid	Long Chain Fatty Acid	LC/MS Neg	natural log

nonadecanoate (19:0)	Lipid	Long Chain Fatty Acid	LC/MS Neg	rank based inverse normal
oleate (18:1n9)	Lipid	Long Chain Fatty Acid	LC/MS Neg	not transformed
palmitate (16:0)	Lipid	Long Chain Fatty Acid	LC/MS Neg	not transformed
palmitoleate (16:1n7)	Lipid	Long Chain Fatty Acid	LC/MS Neg	natural log
stearate (18:0)	Lipid	Long Chain Fatty Acid	LC/MS Neg	rank based inverse normal
1-arachidonoylglycerophosphocholine (20:4n6)	Lipid	Lysolipid	LC/MS Pos	rank based inverse normal
1-arachidonoylglycerophosphoethanolamine	Lipid	Lysolipid	LC/MS Neg	natural log
1-arachidonoylglycerophosphoinositol	Lipid	Lysolipid	LC/MS Neg	natural log
1-docosaheptaenoylglycerophosphocholine (22:6n3)	Lipid	Lysolipid	LC/MS Pos	natural log
1-docosaheptaenoylglycerophosphoethanolamine	Lipid	Lysolipid	LC/MS Neg	rank based inverse normal
1-docosapentaenoylglycerophosphocholine (22:5n3)	Lipid	Lysolipid	LC/MS Pos	natural log
1-linoleoylglycerophosphoethanolamine	Lipid	Lysolipid	LC/MS Neg	natural log
1-oleoylglycerophosphocholine (18:1)	Lipid	Lysolipid	LC/MS Pos	rank based inverse normal
1-palmitoleoylglycerophosphocholine (16:1)	Lipid	Lysolipid	LC/MS Pos	natural log
1-palmitoylglycerophosphoinositol	Lipid	Lysolipid	LC/MS Neg	natural log
1-palmitoylplasmeneylethanolamine	Lipid	Lysolipid	LC/MS Neg	natural log
1-stearoylglycerophosphoethanolamine	Lipid	Lysolipid	LC/MS Neg	natural log
2-arachidonoylglycerophosphocholine	Lipid	Lysolipid	LC/MS Neg	rank based inverse normal
2-linoleoylglycerophosphoethanolamine	Lipid	Lysolipid	LC/MS Neg	rank based inverse normal
2-myristoylglycerophosphocholine	Lipid	Lysolipid	LC/MS Pos	natural log
5-dodecenoate (12:1n7)	Lipid	Medium Chain Fatty Acid	LC/MS Neg	natural log
caprate (10:0)	Lipid	Medium Chain Fatty Acid	LC/MS Neg	rank based inverse normal
caproate (6:0)	Lipid	Medium Chain Fatty Acid	LC/MS Neg	rank based inverse normal
heptanoate (7:0)	Lipid	Medium Chain Fatty Acid	LC/MS Neg	rank based inverse normal
laurate (12:0)	Lipid	Medium Chain Fatty Acid	LC/MS Neg	rank based inverse normal
pelargonate (9:0)	Lipid	Medium Chain Fatty Acid	LC/MS Neg	rank based inverse normal
1-oleoylglycerol (1-monoolein)	Lipid	Monoacylglycerol	LC/MS Neg	rank based inverse normal
1-stearoylglycerol (1-monostearin)	Lipid	Monoacylglycerol	GC/MS	rank based inverse normal
choline	Lipid	Phospholipid Metabolism	LC/MS Pos	rank based inverse normal
glycerophosphorylcholine (GPC)	Lipid	Phospholipid Metabolism	LC/MS Pos	natural log
adrenate (22:4n6)	Lipid	Polyunsaturated Fatty Acid (n3 and n6)	LC/MS Neg	rank based inverse normal
arachidonate (20:4n6)	Lipid	Polyunsaturated Fatty Acid (n3 and n6)	LC/MS Neg	rank based inverse normal

dihomo-linoleate (20:2n6)	Lipid	Polyunsaturated Fatty Acid (n3 and n6)	LC/MS Neg	natural log
dihomo-linolenate (20:3n3 or n6)	Lipid	Polyunsaturated Fatty Acid (n3 and n6)	LC/MS Neg	natural log
docosaheptaenoate (DHA; 22:6n3)	Lipid	Polyunsaturated Fatty Acid (n3 and n6)	LC/MS Neg	rank based inverse normal
docosapentaenoate (n3 DPA; 22:5n3)	Lipid	Polyunsaturated Fatty Acid (n3 and n6)	LC/MS Neg	natural log
docosapentaenoate (n6 DPA; 22:5n6)	Lipid	Polyunsaturated Fatty Acid (n3 and n6)	LC/MS Neg	rank based inverse normal
eicosapentaenoate (EPA; 20:5n3)	Lipid	Polyunsaturated Fatty Acid (n3 and n6)	LC/MS Neg	natural log
linoleate (18:2n6)	Lipid	Polyunsaturated Fatty Acid (n3 and n6)	LC/MS Neg	natural log
linolenate [alpha or gamma; (18:3n3 or 6)]	Lipid	Polyunsaturated Fatty Acid (n3 and n6)	LC/MS Neg	natural log
stearidonate (18:4n3)	Lipid	Polyunsaturated Fatty Acid (n3 and n6)	LC/MS Neg	rank based inverse normal
glycochenodeoxycholate	Lipid	Primary Bile Acid Metabolism	LC/MS Pos	natural log
glycocholate	Lipid	Primary Bile Acid Metabolism	LC/MS Neg	natural log
taurochenodeoxycholate	Lipid	Primary Bile Acid Metabolism	LC/MS Neg	rank based inverse normal
glycochenolate sulfate	Lipid	Secondary Bile Acid Metabolism	LC/MS Neg	rank based inverse normal
glycodeoxycholate	Lipid	Secondary Bile Acid Metabolism	LC/MS Pos	natural log
glycolithocholate sulfate	Lipid	Secondary Bile Acid Metabolism	LC/MS Neg	rank based inverse normal
glycoursodeoxycholate	Lipid	Secondary Bile Acid Metabolism	LC/MS Neg	natural log
taurocholenate sulfate	Lipid	Secondary Bile Acid Metabolism	LC/MS Neg	natural log
tauroolithocholate 3-sulfate	Lipid	Secondary Bile Acid Metabolism	LC/MS Neg	rank based inverse normal
21-hydroxypregnenolone disulfate	Lipid	Steroid	LC/MS Neg	rank based inverse normal
4-androsten-3beta,17beta-diol disulfate (1)	Lipid	Steroid	LC/MS Neg	rank based inverse normal
4-androsten-3beta,17beta-diol disulfate (2)	Lipid	Steroid	LC/MS Neg	rank based inverse normal
5alpha-androstan-3beta,17beta-diol disulfate	Lipid	Steroid	LC/MS Neg	rank based inverse normal
5alpha-pregnan-3beta,20alpha-diol disulfate	Lipid	Steroid	LC/MS Neg	natural log
androsterone sulfate	Lipid	Steroid	LC/MS Neg	rank based inverse normal
cortisol	Lipid	Steroid	LC/MS Pos	natural log
cortisone	Lipid	Steroid	LC/MS Pos	natural log
dehydroisoandrosterone sulfate (DHEA-S)	Lipid	Steroid	LC/MS Neg	natural log
epiandrosterone sulfate	Lipid	Steroid	LC/MS Neg	rank based inverse normal
pregn steroid monosulfate	Lipid	Steroid	LC/MS Neg	natural log
pregnen-diol disulfate	Lipid	Steroid	LC/MS Neg	natural log
7-alpha-hydroxy-3-oxo-4-cholestenoate (7-Hoca)	Lipid	Sterol	LC/MS Neg	rank based inverse normal
cholesterol	Lipid	Sterol	GC/MS	rank based inverse normal
hypoxanthine	Nucleotide	Purine Metabolism, (Hypo)Xanthine/Inosine containing	LC/MS Neg	rank based inverse normal
urate	Nucleotide	Purine Metabolism, (Hypo)Xanthine/Inosine containing	LC/MS Neg	rank based inverse normal

xanthine	Nucleotide	Purine Metabolism, (Hypo)Xanthine/Inosine containing	LC/MS Pos	rank based inverse normal
N1-methyladenosine	Nucleotide	Purine Metabolism, Adenine containing	LC/MS Pos	rank based inverse normal
guanosine	Nucleotide	Purine Metabolism, Guanine containing	LC/MS Pos	rank based inverse normal
5-methyluridine (ribothymidine)	Nucleotide	Pyrimidine Metabolism, Uracil containing	LC/MS Neg	rank based inverse normal
N-acetyl-beta-alanine	Nucleotide	Pyrimidine Metabolism, Uracil containing	LC/MS Pos	natural log
pseudouridine	Nucleotide	Pyrimidine Metabolism, Uracil containing	LC/MS Pos	rank based inverse normal
uridine	Nucleotide	Pyrimidine Metabolism, Uracil containing	LC/MS Neg	rank based inverse normal
alanylleucine	Peptide	Dipeptide	LC/MS Pos	rank based inverse normal
arginylleucine	Peptide	Dipeptide	LC/MS Pos	natural log
arginylphenylalanine	Peptide	Dipeptide	LC/MS Pos	rank based inverse normal
glycylleucine	Peptide	Dipeptide	LC/MS Pos	rank based inverse normal
glycylphenylalanine	Peptide	Dipeptide	LC/MS Pos	natural log
glycylvaline	Peptide	Dipeptide	LC/MS Pos	rank based inverse normal
histidylleucine	Peptide	Dipeptide	LC/MS Pos	natural log
leucylalanine	Peptide	Dipeptide	LC/MS Pos	rank based inverse normal
leucylasparagine	Peptide	Dipeptide	LC/MS Pos	rank based inverse normal
leucylglycine	Peptide	Dipeptide	LC/MS Pos	natural log
leucylleucine	Peptide	Dipeptide	LC/MS Pos	rank based inverse normal
leucylserine	Peptide	Dipeptide	LC/MS Pos	natural log
phenylalanylglutamate	Peptide	Dipeptide	LC/MS Pos	rank based inverse normal
phenylalanylleucine	Peptide	Dipeptide	LC/MS Pos	rank based inverse normal
phenylalanylphenylalanine	Peptide	Dipeptide	LC/MS Pos	rank based inverse normal
phenylalanylserine	Peptide	Dipeptide	LC/MS Pos	rank based inverse normal
pyroglutamylglycine	Peptide	Dipeptide	LC/MS Neg	natural log
serylleucine	Peptide	Dipeptide	LC/MS Pos	rank based inverse normal
N-acetylcarnosine	Peptide	Dipeptide Derivative	LC/MS Pos	natural log
DSGEGDFXAEGGGVR	Peptide	Fibrinogen Cleavage Peptide	LC/MS Pos	rank based inverse normal
gamma-glutamylalanine	Peptide	Gamma-glutamyl Amino Acid	LC/MS Pos	rank based inverse normal
gamma-glutamylglutamate	Peptide	Gamma-glutamyl Amino Acid	LC/MS Pos	rank based inverse normal
gamma-glutamylisoleucine	Peptide	Gamma-glutamyl Amino Acid	LC/MS Pos	rank based inverse normal
gamma-glutamylleucine	Peptide	Gamma-glutamyl Amino Acid	LC/MS Pos	rank based inverse normal
gamma-glutamylphenylalanine	Peptide	Gamma-glutamyl Amino Acid	LC/MS Pos	rank based inverse normal
gamma-glutamylthreonine	Peptide	Gamma-glutamyl Amino Acid	LC/MS Pos	rank based inverse normal
gamma-glutamyltyrosine	Peptide	Gamma-glutamyl Amino Acid	LC/MS Pos	rank based inverse normal
gamma-glutamylvaline	Peptide	Gamma-glutamyl Amino Acid	LC/MS Pos	rank based inverse normal
HWESASLLR	Peptide	Polypeptide	LC/MS Pos	rank based inverse normal
2-hydroxyhippurate (salicylurate)	Xenobiotics	Benzoate Metabolism	LC/MS Neg	rank based inverse normal
3-methyl catechol sulfate (1)	Xenobiotics	Benzoate Metabolism	LC/MS Neg	natural log

4-hydroxyhippurate	Xenobiotics	Benzoate Metabolism	LC/MS Neg	natural log
4-methylcatechol sulfate	Xenobiotics	Benzoate Metabolism	LC/MS Neg	rank based inverse normal
4-vinylphenol sulfate	Xenobiotics	Benzoate Metabolism	LC/MS Neg	rank based inverse normal
catechol sulfate	Xenobiotics	Benzoate Metabolism	LC/MS Neg	rank based inverse normal
hippurate	Xenobiotics	Benzoate Metabolism	LC/MS Neg	rank based inverse normal
methyl-4-hydroxybenzoate sulfate	Xenobiotics	Benzoate Metabolism	LC/MS Neg	rank based inverse normal
O-methylcatechol sulfate	Xenobiotics	Benzoate Metabolism	LC/MS Neg	rank based inverse normal
propyl 4-hydroxybenzoate sulfate	Xenobiotics	Benzoate Metabolism	LC/MS Neg	rank based inverse normal
1,2-propanediol	Xenobiotics	Chemical	GC/MS	rank based inverse normal
2-hydroxyisobutyrate	Xenobiotics	Chemical	GC/MS	natural log
3-hydroxypyridine sulfate	Xenobiotics	Chemical	LC/MS Neg	rank based inverse normal
O-sulfo-L-tyrosine	Xenobiotics	Chemical	LC/MS Neg	natural log
salicylate	Xenobiotics	Drug	LC/MS Neg	rank based inverse normal
erythritol	Xenobiotics	Food Component/Plant	GC/MS	rank based inverse normal
gluconate	Xenobiotics	Food Component/Plant	LC/MS Polar	rank based inverse normal
homostachydrine	Xenobiotics	Food Component/Plant	LC/MS Pos	natural log
piperine	Xenobiotics	Food Component/Plant	LC/MS Pos	rank based inverse normal
stachydrine	Xenobiotics	Food Component/Plant	LC/MS Pos	rank based inverse normal
thymol sulfate	Xenobiotics	Food Component/Plant	LC/MS Neg	rank based inverse normal
1,7-dimethylurate	Xenobiotics	Xanthine Metabolism	LC/MS Neg	rank based inverse normal
1-methylurate	Xenobiotics	Xanthine Metabolism	LC/MS Pos	rank based inverse normal
5-acetylamino-6-amino-3-methyluracil	Xenobiotics	Xanthine Metabolism	LC/MS Neg	rank based inverse normal
caffeine	Xenobiotics	Xanthine Metabolism	LC/MS Pos	rank based inverse normal
paraxanthine	Xenobiotics	Xanthine Metabolism	LC/MS Pos	rank based inverse normal
theobromine	Xenobiotics	Xanthine Metabolism	LC/MS Pos	rank based inverse normal
theophylline	Xenobiotics	Xanthine Metabolism	LC/MS Pos	rank based inverse normal

Supplemental Table 2. Genes discovered through jointly testing the genetic main effects and gene-sex interactions

			AA					EA				
Metabolites	Gene	Chr	N_S NP	Fix.pvalu e	Random. pvalue	Joint.pva lue	Fdr.p.joi nt	N_S NP	Fix.pvalu e	Random. pvalue	Joint.pva lue	Postfdr.p .joint
2-aminooctanoate	ALMS1	2	52	5.24E-02	1.26E-01	2.16E-07	2.28E-02	37	5.24E-02	3.02E-02	2.15E-03	8.17E-03
2-aminooctanoate	ACY1	3	7	2.64E-02	2.89E-02	2.19E-08	4.63E-03	3	4.93E-02	6.23E-02	1.29E-01	1.96E-01
arachidonate (20:4n6)	PLA2G7	6	7	6.50E-01	6.67E-01	6.19E-07	4.62E-02	7	1.68E-02	1.25E-02	1.86E-02	5.06E-02
l- arachidonoylglycerophosphoethanolamine	MAP10	1	13	4.11E-03	4.22E-03	1.72E-07	1.94E-02	5	6.11E-01	7.62E-01	6.88E-01	7.68E-01
arginylleucine	NPC2	14	4	9.95E-02	8.02E-02	5.26E-07	4.34E-02	5	2.22E-01	4.75E-01	1.79E-01	2.61E-01
arginylleucine	NDRG3	20	4	2.79E-01	2.26E-01	4.55E-08	6.69E-03	NA	NA	NA	NA	NA
cholesterol	PCSK9	1	18	9.52E-01	9.59E-01	6.90E-10	2.12E-04	6	6.84E-01	6.04E-01	6.72E-02	1.14E-01
deoxycarnitine	SLC25A45	11	8	6.48E-01	6.62E-01	1.70E-07	1.94E-02	5	3.89E-01	4.51E-01	4.37E-03	1.51E-02
deoxycarnitine	C1QTNF1	17	5	9.57E-03	7.16E-03	6.29E-07	4.62E-02	5	6.66E-01	7.47E-01	2.45E-01	3.21E-01
dimethylglycine	DMGDH	5	12	8.01E-02	6.50E-02	4.57E-36	7.73E-30	8	6.23E-01	6.33E-01	2.56E-09	1.95E-08
DSGEGDFXAEGGGV R	CPN1	10	3	9.52E-01	9.14E-01	5.25E-10	1.77E-04	NA	NA	NA	NA	NA
eicosanodioate	ALDH7A1	5	8	8.55E-07	5.64E-07	5.62E-07	4.52E-02	4	1.80E-01	1.70E-01	3.12E-01	3.82E-01
gamma- glutamylphenylalanine	COPE	19	3	7.02E-08	7.53E-08	9.65E-08	1.31E-02	3	5.60E-01	6.91E-01	8.47E-01	8.70E-01
glutaryl carnitine (C5)	RNASEH2A	19	5	5.05E-01	4.98E-01	2.88E-08	5.72E-03	3	9.32E-01	9.20E-01	2.65E-01	3.35E-01
glycochenodeoxycholate	SLC10A1	14	8	2.11E-01	2.96E-01	5.75E-07	4.52E-02	NA	NA	NA	NA	NA
glycocholate	LRP4	11	18	7.30E-06	5.14E-06	3.98E-07	3.54E-02	12	9.37E-01	9.48E-01	9.89E-01	9.89E-01
glycocholate	SLC10A1	14	8	1.81E-01	2.44E-01	2.01E-10	8.52E-05	NA	NA	NA	NA	NA
glycochenolate sulfate	CYP27A1	2	12	9.12E-01	9.08E-01	1.34E-07	1.74E-02	7	3.26E-01	3.57E-01	6.22E-02	1.13E-01
glycochenolate sulfate	HRASLS5	11	6	7.41E-01	7.93E-01	3.24E-08	5.78E-03	7	9.12E-02	1.29E-01	2.26E-01	3.07E-01

glycocholate sulfate	SLCO1B1	12	9	9.60E-01	9.42E-01	1.26E-08	2.84E-03	NA	NA	NA	NA	NA
hexadecanedioate	SLCO1B1	12	9	2.28E-01	2.39E-01	5.25E-10	1.77E-04	NA	NA	NA	NA	NA
histidine	HAL	12	11	9.90E-01	9.84E-01	7.07E-07	4.98E-02	14	1.66E-01	1.61E-01	8.69E-04	4.13E-03
histidylleucine	NPC2	14	4	1.70E-03	1.25E-03	3.74E-08	6.02E-03	5	1.35E-01	2.73E-01	4.83E-02	9.67E-02
13-HODE + 9-HODE	PLA2G7	6	7	7.24E-02	7.82E-02	3.44E-08	5.81E-03	7	3.57E-01	4.03E-01	5.85E-01	6.74E-01
3-(4-hydroxyphenyl)lactate	CCBL1	9	9	8.97E-01	9.40E-01	4.57E-14	2.57E-08	4	7.91E-01	7.87E-01	2.51E-02	6.11E-02
indolelactate	CCBL1	9	9	5.77E-01	6.23E-01	5.15E-23	5.81E-17	4	3.38E-01	2.99E-01	2.70E-04	1.47E-03
isobutyrylcarnitine	ACAD8	11	3	3.87E-01	4.08E-01	6.98E-13	3.37E-07	3	1.91E-01	1.99E-01	2.58E-10	2.45E-09
kynurenine	ZNF827	4	8	1.66E-02	2.52E-02	2.50E-07	2.48E-02	7	4.31E-01	3.74E-01	3.71E-02	8.28E-02
kynurenine	IDO1	8	6	8.57E-01	8.95E-01	3.99E-09	1.05E-03	NA	NA	NA	NA	NA
leucylasparagine	KLKB1	4	13	4.52E-01	4.65E-01	4.87E-07	4.12E-02	10	1.61E-01	1.37E-01	1.75E-02	5.06E-02
leucylglycine	NDRG3	20	4	7.57E-01	6.56E-01	7.84E-08	1.10E-02	NA	NA	NA	NA	NA
leucylserine	NPC2	14	4	5.83E-04	3.79E-04	1.93E-07	2.10E-02	5	2.93E-02	6.59E-02	6.78E-03	2.15E-02
linolenate [alpha or gamma; (18:3n3 or 6)]	COL3A1	2	13	2.05E-01	2.37E-01	1.41E-07	1.76E-02	7	1.04E-01	9.31E-02	4.01E-02	8.47E-02
4-methylcatechol sulfate	PSME4	2	12	8.52E-01	8.58E-01	6.18E-09	1.49E-03	9	5.27E-01	5.06E-01	7.36E-01	7.93E-01
N1-methyladenosine	ADA	20	5	5.64E-02	5.59E-02	2.69E-07	2.60E-02	NA	NA	NA	NA	NA
N-acetyl-1-methylhistidine	ALMS1	2	52	9.88E-01	9.77E-01	4.05E-09	1.05E-03	37	4.60E-01	4.45E-01	3.32E-11	4.21E-10
N-acetyl-1-methylhistidine	DDTL	22	2	3.65E-08	1.18E-07	2.24E-07	2.29E-02	NA	NA	NA	NA	NA
N-acetylalanine	ACY1	3	7	3.76E-01	3.82E-01	4.43E-52	1.50E-45	3	4.65E-01	4.74E-01	4.07E-23	1.55E-21
N-acetylalanine	ADCY5	3	7	1.06E-06	1.71E-06	7.02E-07	4.98E-02	9	2.72E-01	2.96E-01	6.90E-02	1.14E-01
N-acetyl-beta-alanine	CRYGA	2	4	6.98E-01	8.47E-01	4.38E-07	3.80E-02	4	1.99E-01	2.06E-01	3.50E-01	4.16E-01
N-acetyl-beta-alanine	PTER	10	7	6.20E-01	5.55E-01	3.07E-08	5.77E-03	4	9.58E-02	8.68E-02	9.84E-12	1.87E-10
N-acetylcarnosine	SEPT-9	17	3	4.60E-02	8.83E-02	3.31E-07	3.11E-02	6	1.56E-01	1.85E-01	2.23E-01	3.07E-01

N-acetyl glycine	ACY1	3	7	3.30E-01	3.28E-01	3.93E-08	6.05E-03	3	7.68E-01	7.81E-01	1.58E-03	6.66E-03
N-acetylthreonine	ACY1	3	7	2.68E-01	2.72E-01	8.01E-16	6.77E-10	3	2.39E-01	2.53E-01	2.71E-05	1.72E-04
10-nonadecenoate (19:1n9)	AGER	6	7	4.32E-07	2.16E-07	3.95E-07	3.54E-02	9	4.06E-01	3.25E-01	5.74E-02	1.09E-01
phenyllactate (PLA)	CCBL1	9	9	4.13E-01	4.66E-01	2.73E-14	1.85E-08	4	4.43E-01	3.76E-01	2.57E-02	6.11E-02
taurocholate sulfate	OR2C3	1	7	6.14E-07	4.38E-07	6.26E-07	4.62E-02	3	5.91E-01	5.65E-01	7.51E-01	7.93E-01
tiglyl carnitine	DARS	2	6	7.52E-06	2.75E-06	1.53E-07	1.85E-02	2	1.47E-01	1.16E-01	7.61E-02	1.21E-01

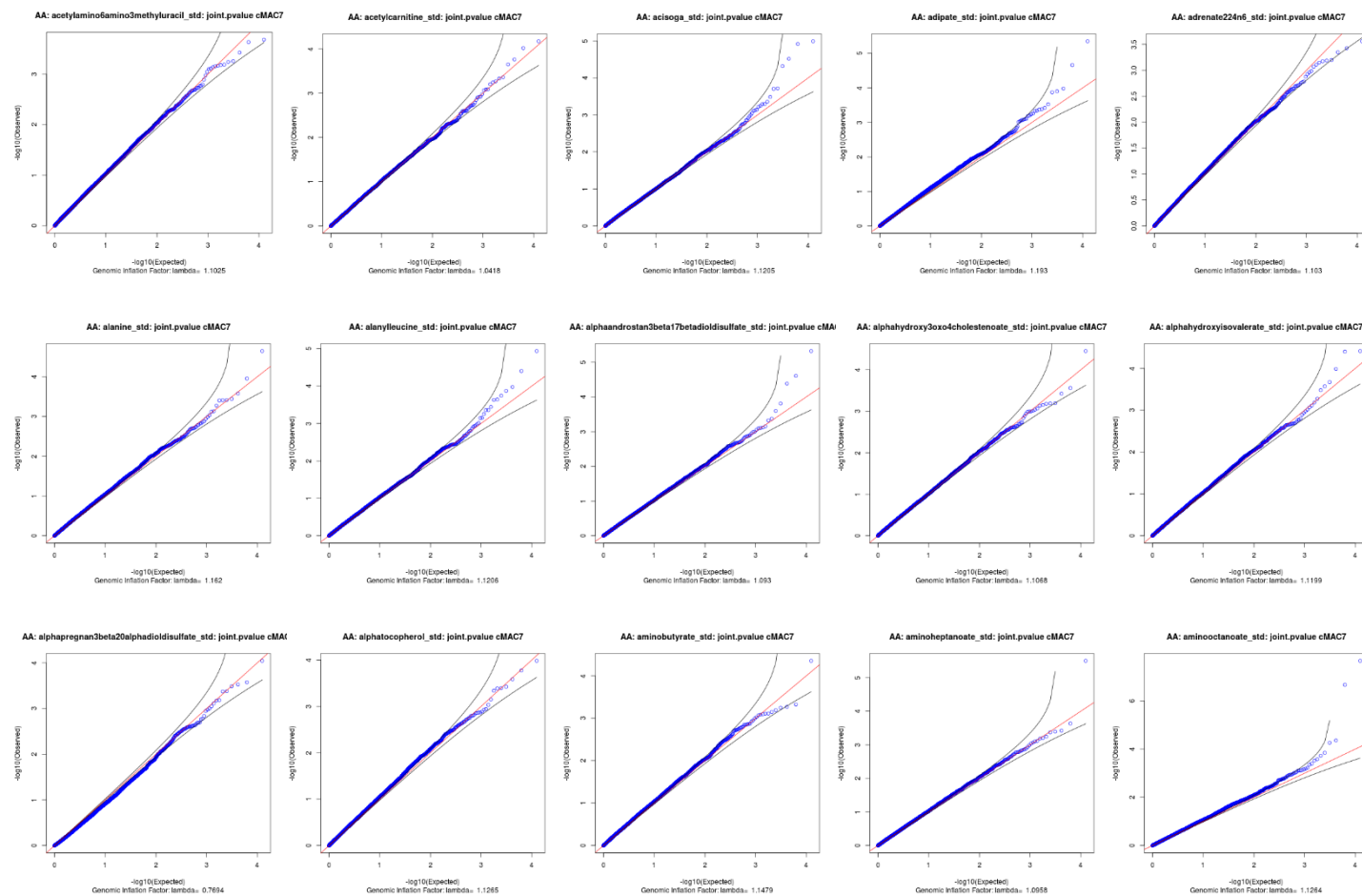
Supplemental Table 3 Genes revealed as suggestively significant (false discovery rate <20%) through rareGE fix/random effect interaction test

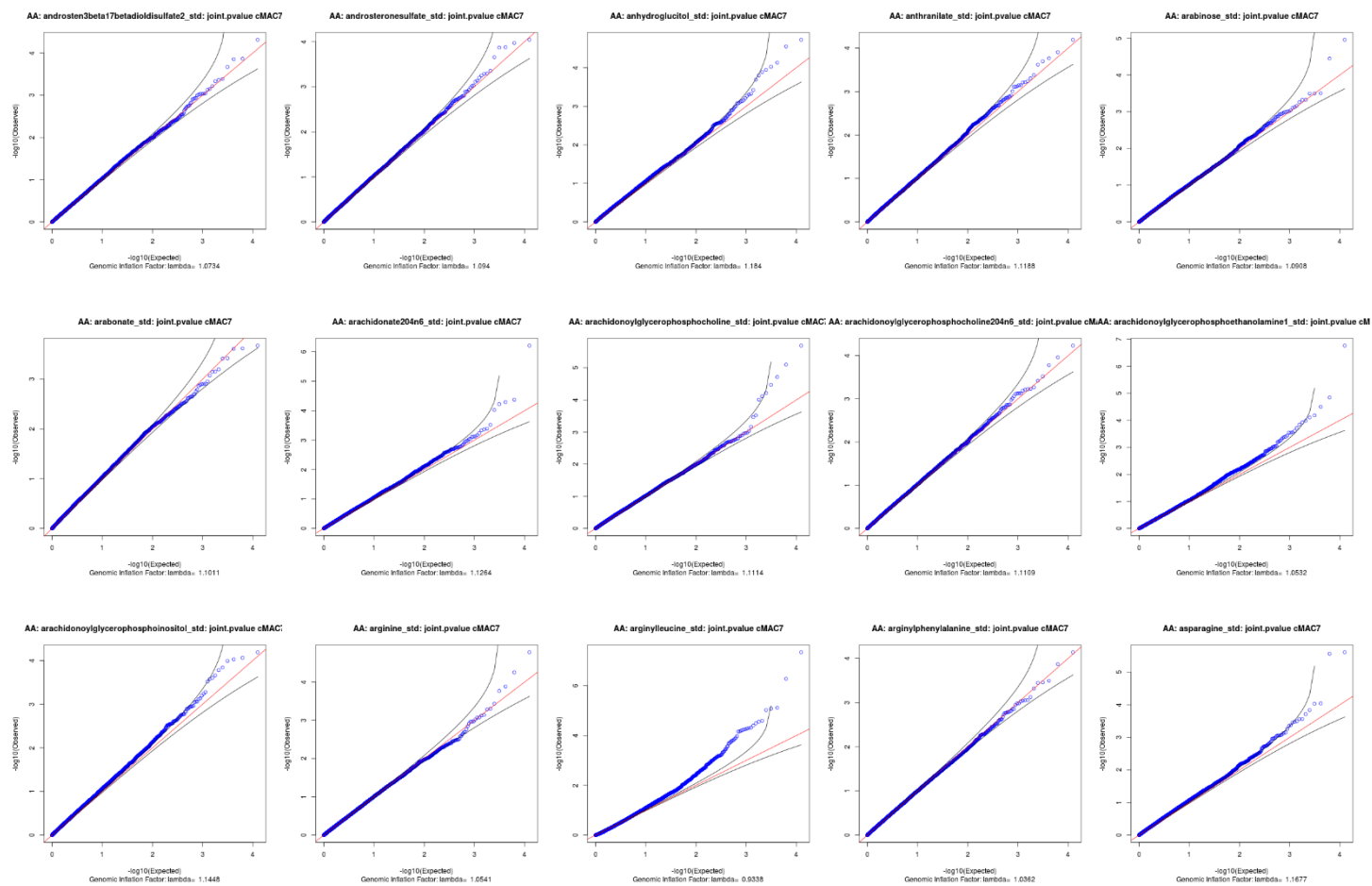
AA								EA		
Metabolites	Gene	Chr	N_S NP	Fix.pvalue	Random.pvalue	Fdr.fix. p	Fdr.random. p	N_SN P	Fix.pvalue	Random.pvalue
gamma-glutamylphenylalanine	COPE	19	3	7.02E-08	7.53E-08	0.12	0.20	3	0.56	0.69
N-acetyl-1-methylhistidine	DDTL	22	2	3.65E-08	1.18E-07	0.12	0.20	NA	NA	NA
arginylleucine	SPACA4	19	2	1.33E-07	2.77E-05	0.15	0.73	NA	NA	NA

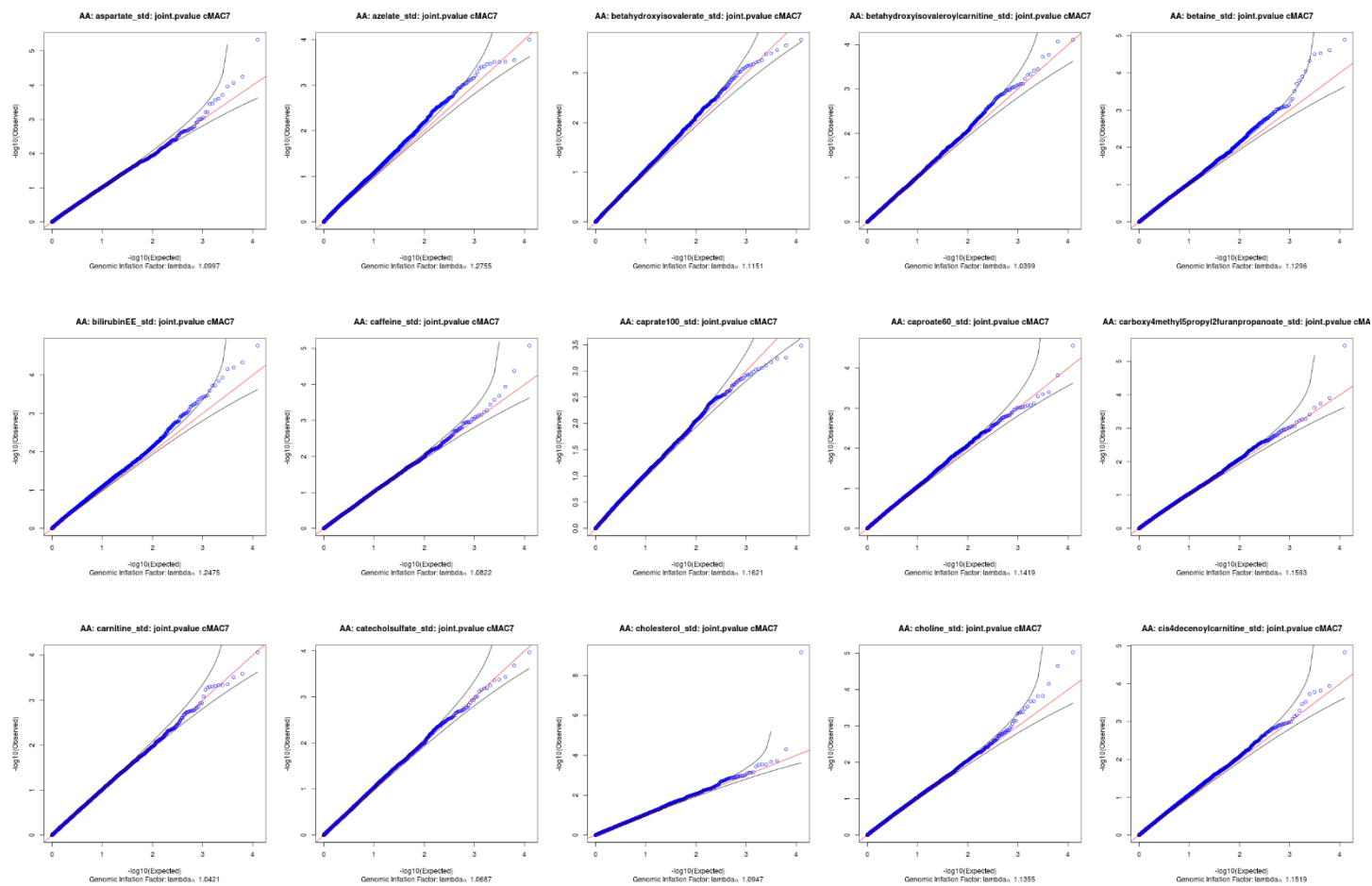
Supplemental Table 4 Genes revealed as suggestively significant (false discovery rate <20%) through MiSTi interaction test

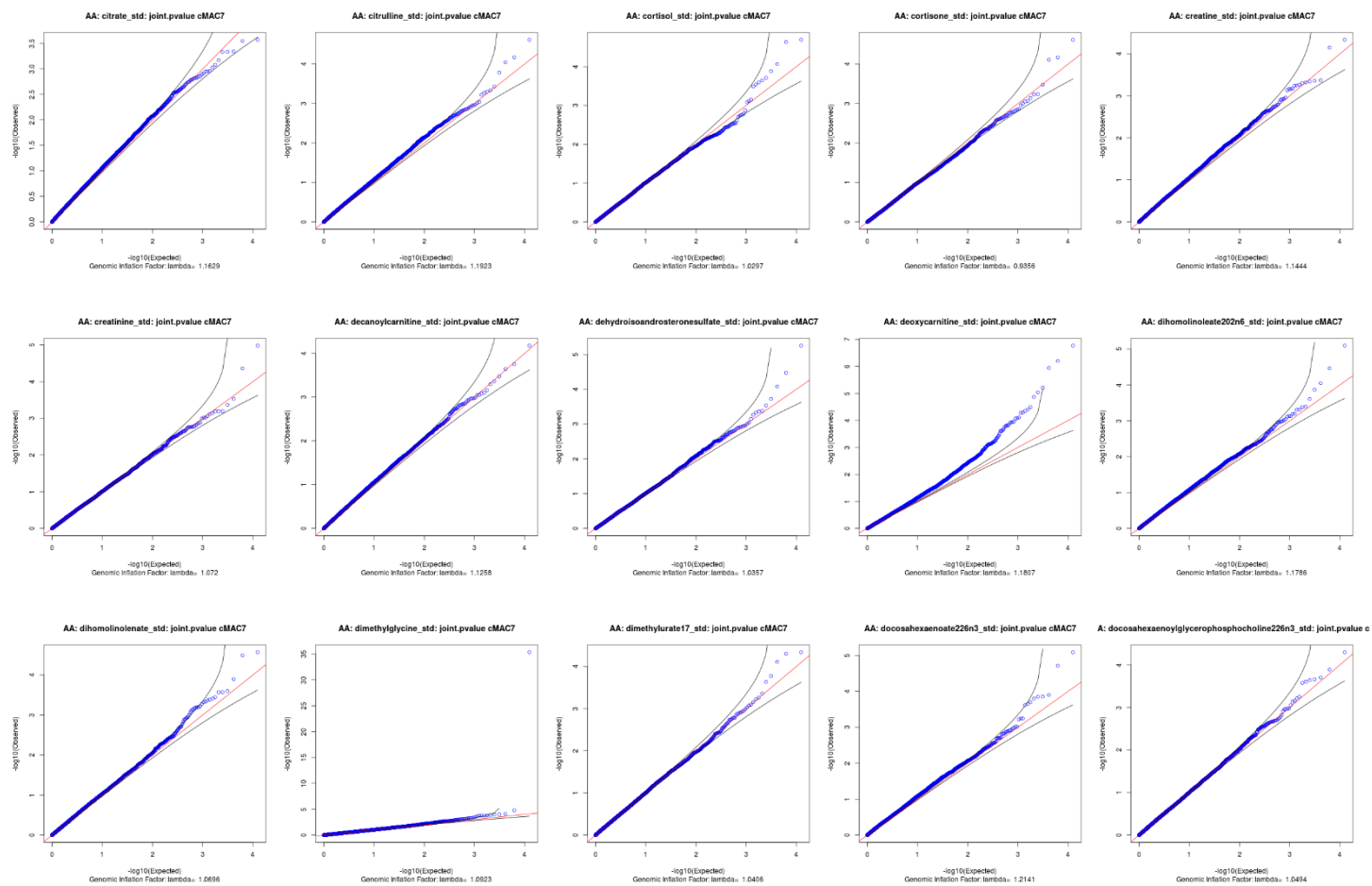
AA								EA			
Metabolites	Gene	Chr	N_S NP	Fisher.p value	BurdenComp. pvalue	VarComp.p value	Fdr.Fisher.p	N_SN P	Fisher. pvalue	BurdenComp. .pvalue	VarComp.p value
bilirubin (E,E)	OLH			5 .18E-08	2.14E -05	1.16 E-04	0 .09		0 .57	0 0.46	0 0.51
N-acetyl- 1-methylhistidine	DTL	2		5 .41E-08	8.64E -04	3.01 E-06	0 .09	A	A	NA	NA
arginylleu cine	NL1			1 .21E-07	1.15E -02	5.30 E-07	0 .13		0 .49	0 0.87	0 0.21
arginylleu cine	TPN5	1		1 .86E-07	5.11E -04	1.87 E-05	0 .15	A	A	NA	NA

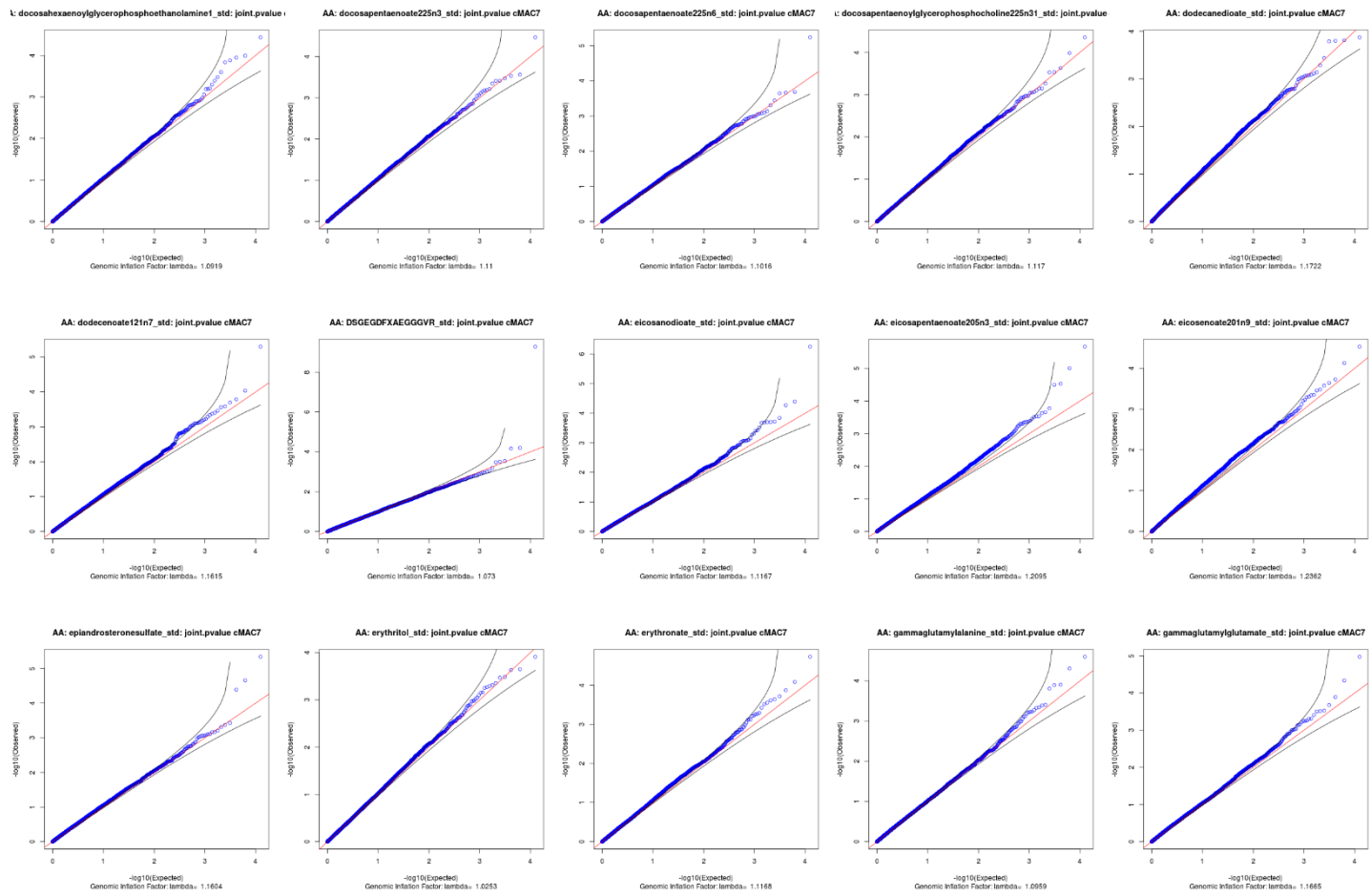
Supplemental Figure 1. QQ plots of joint test results for each metabolite

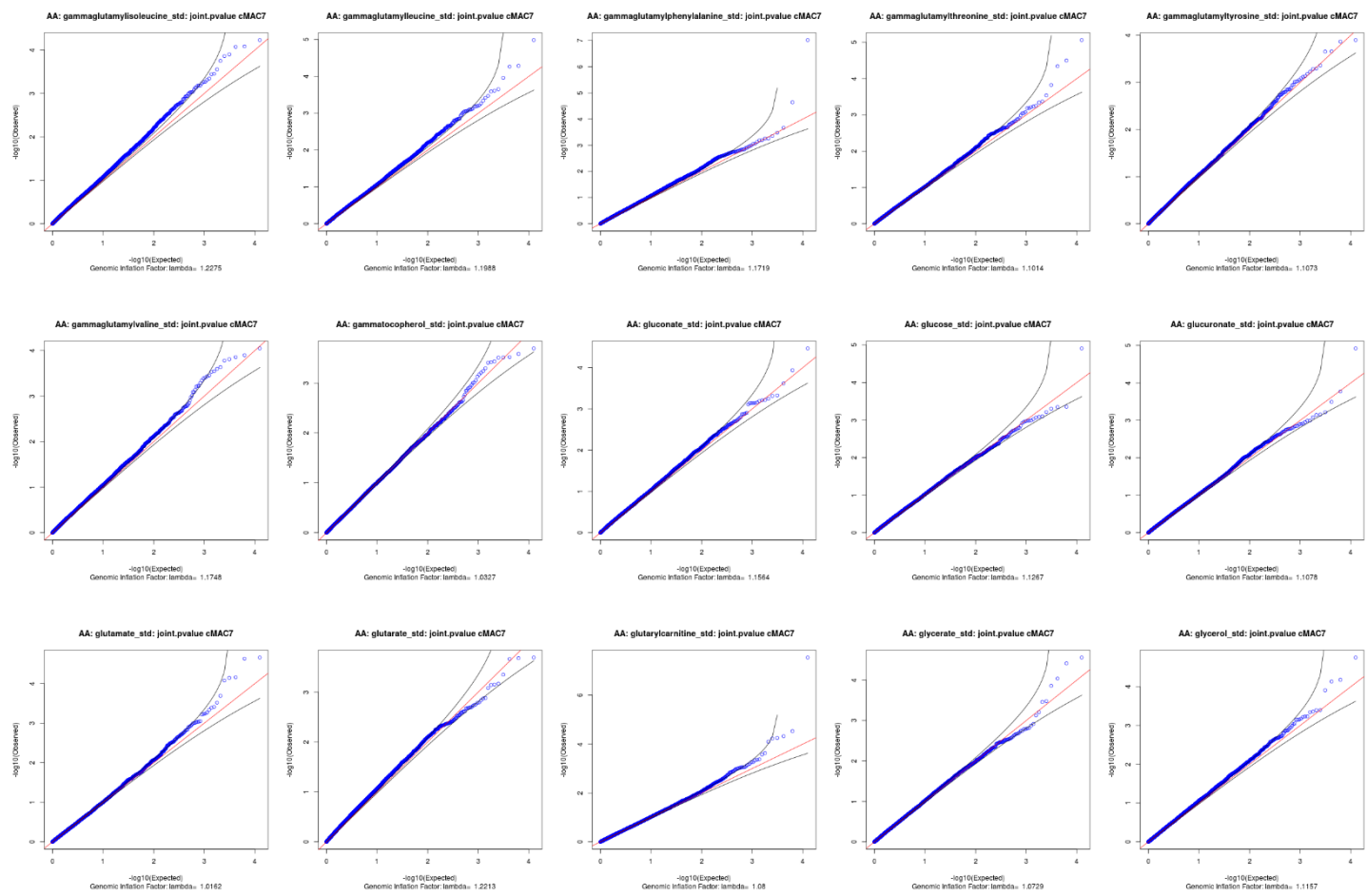


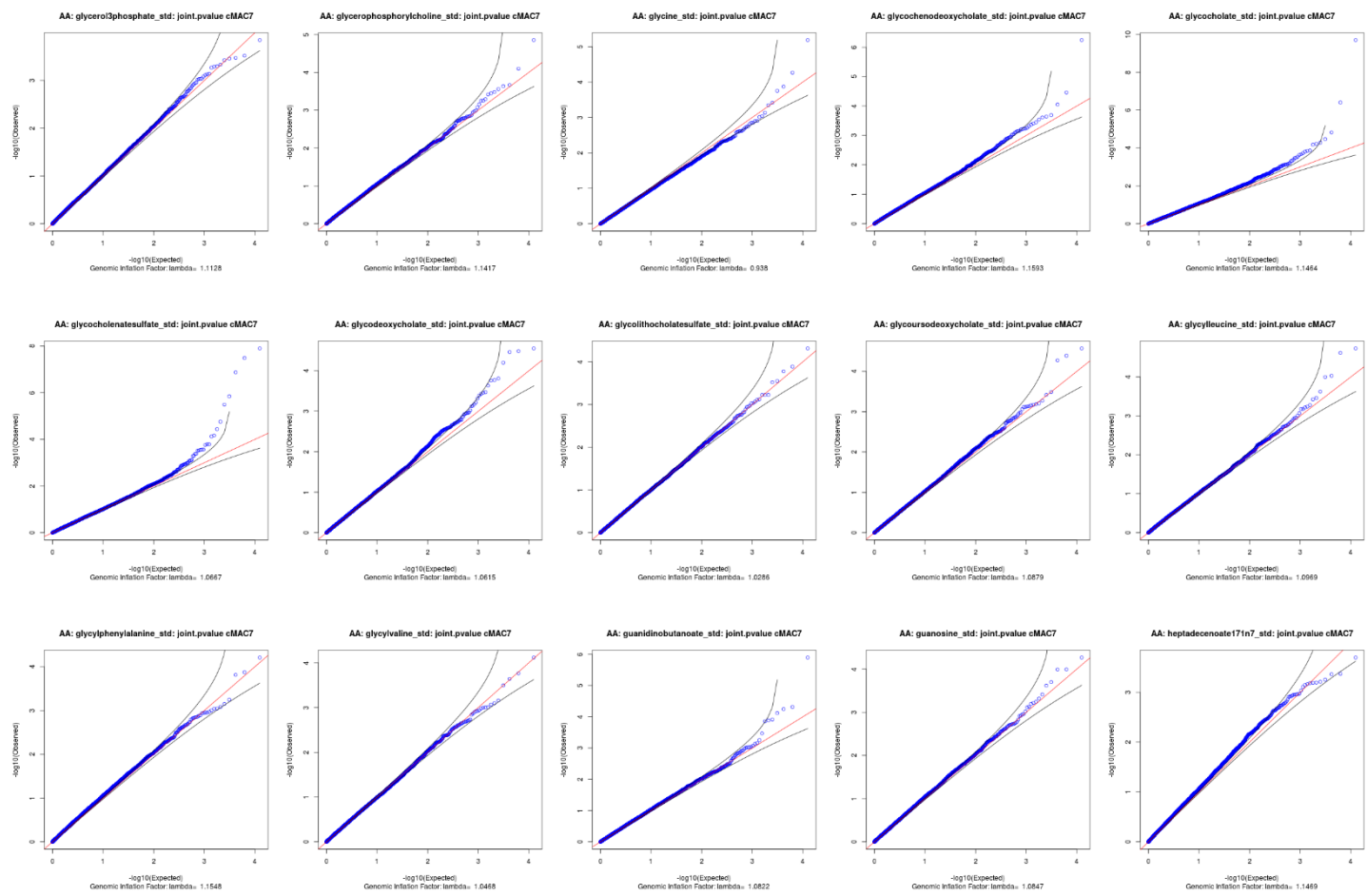


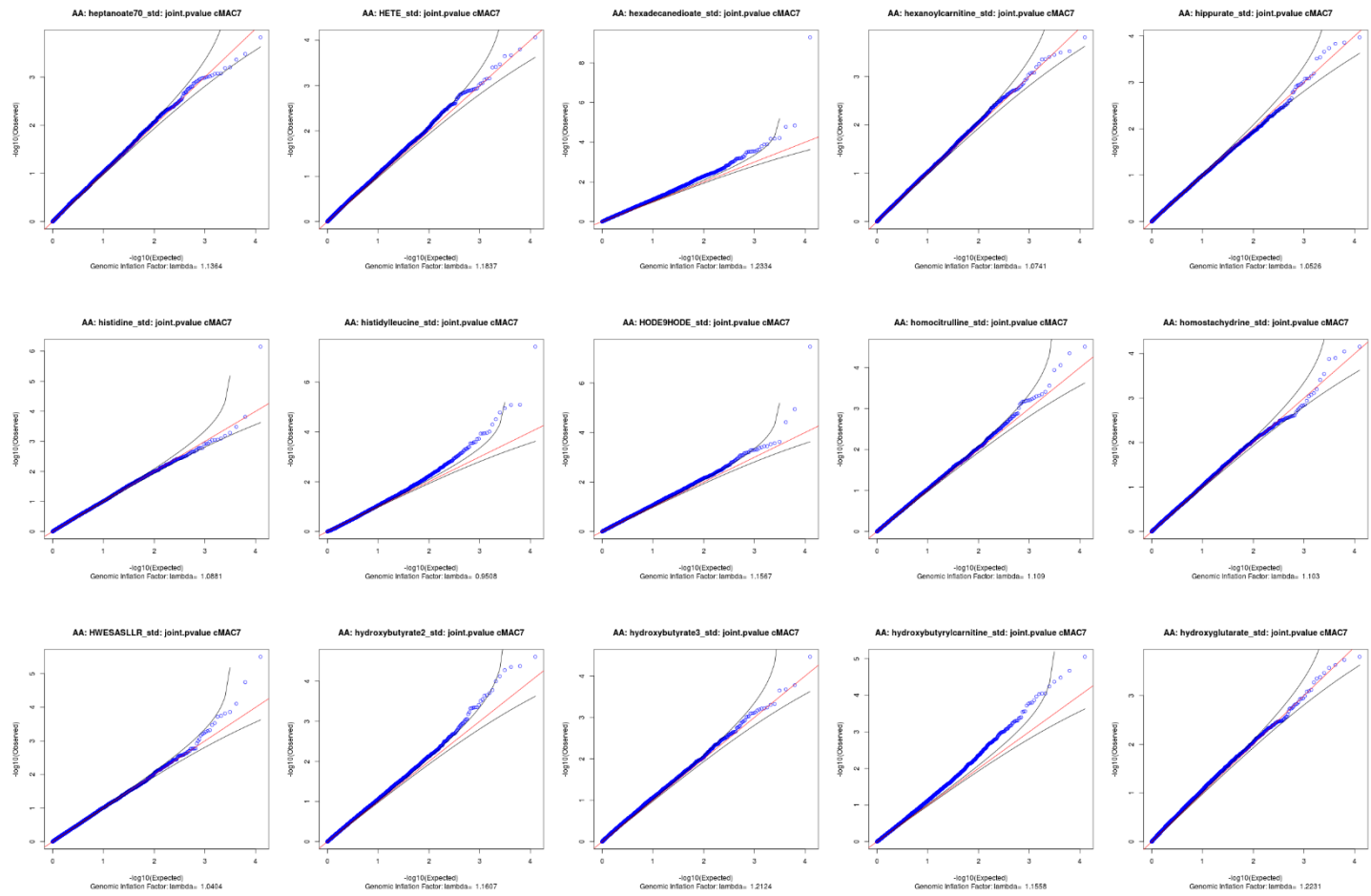


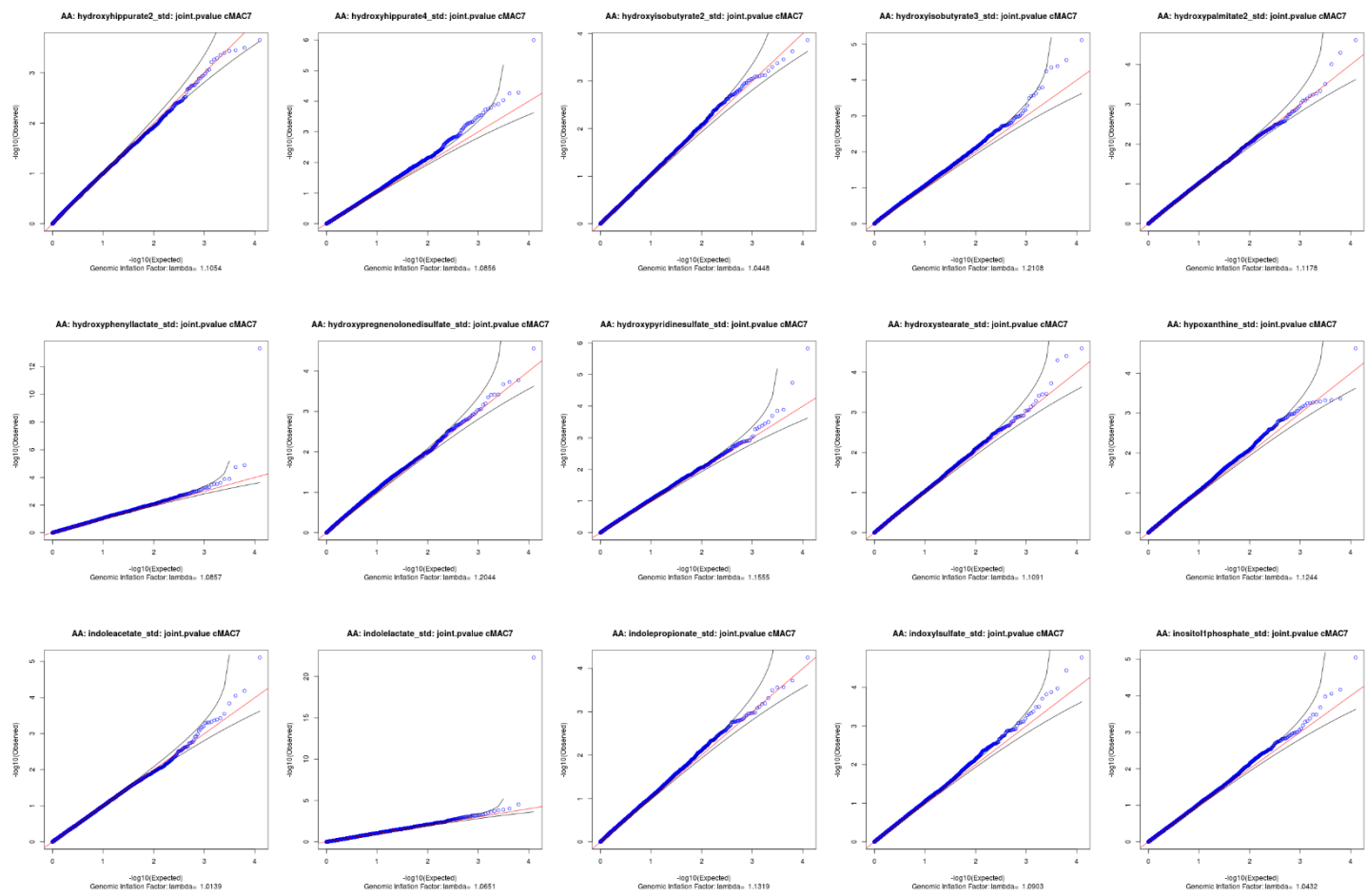


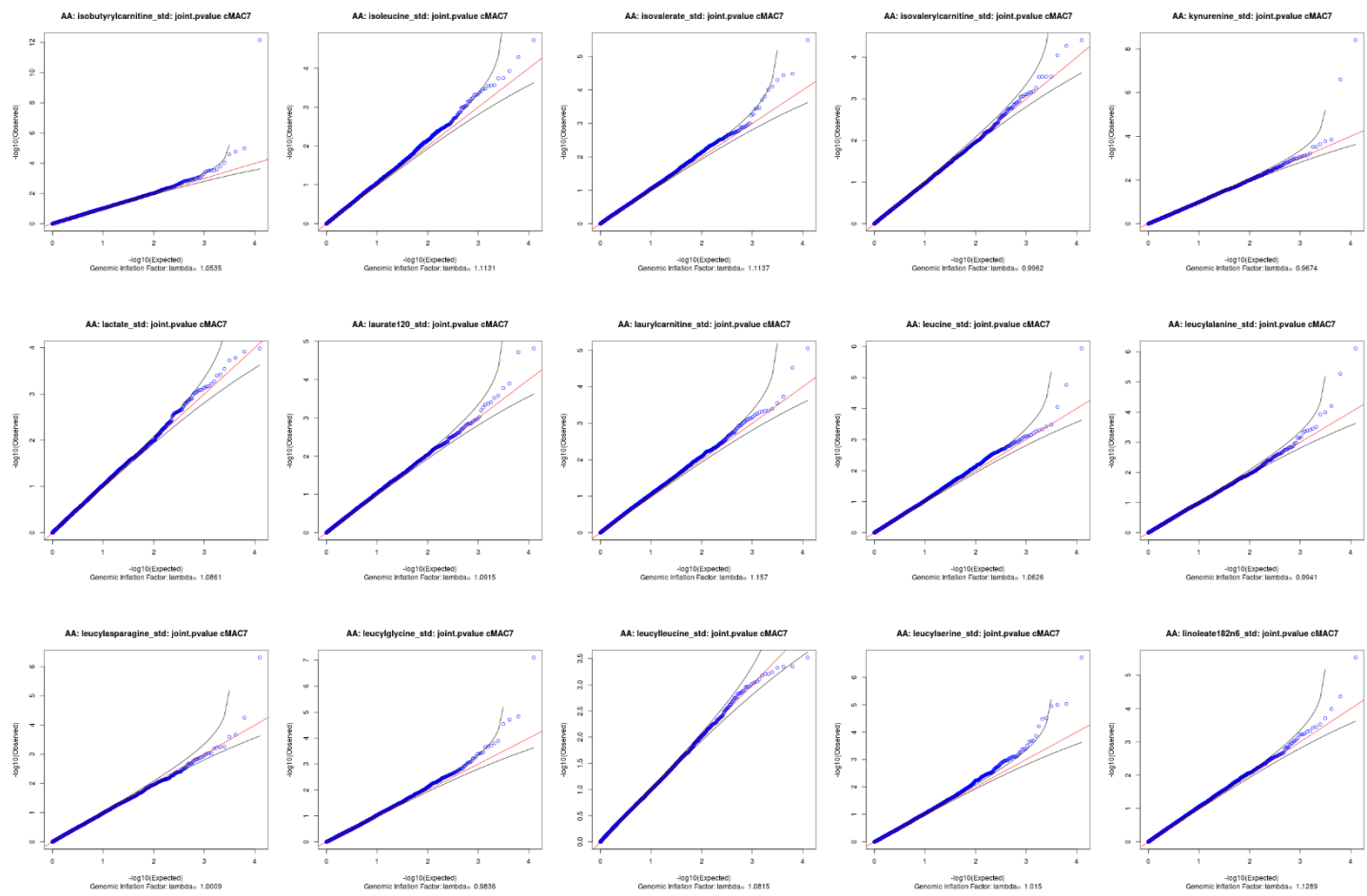


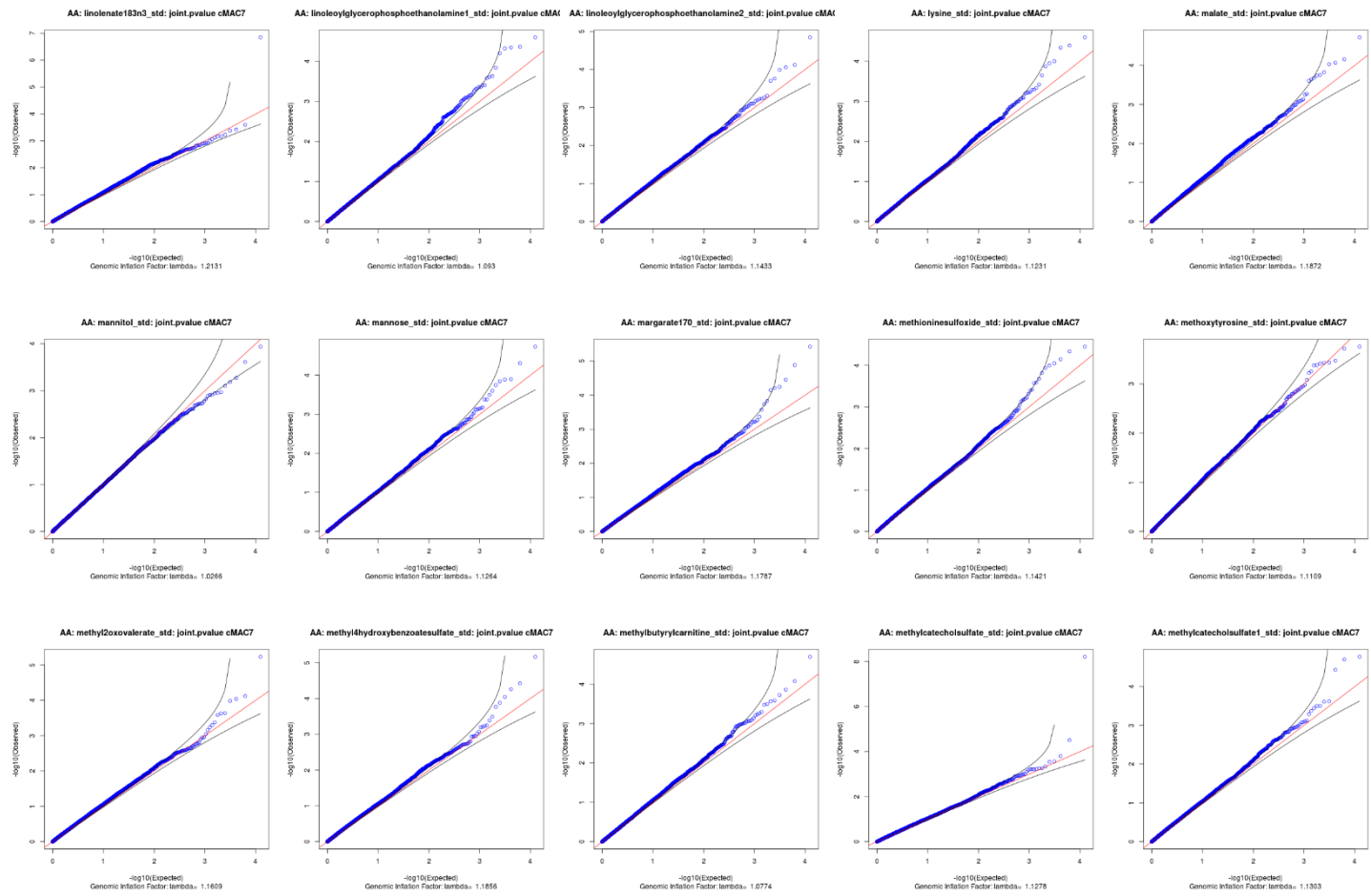


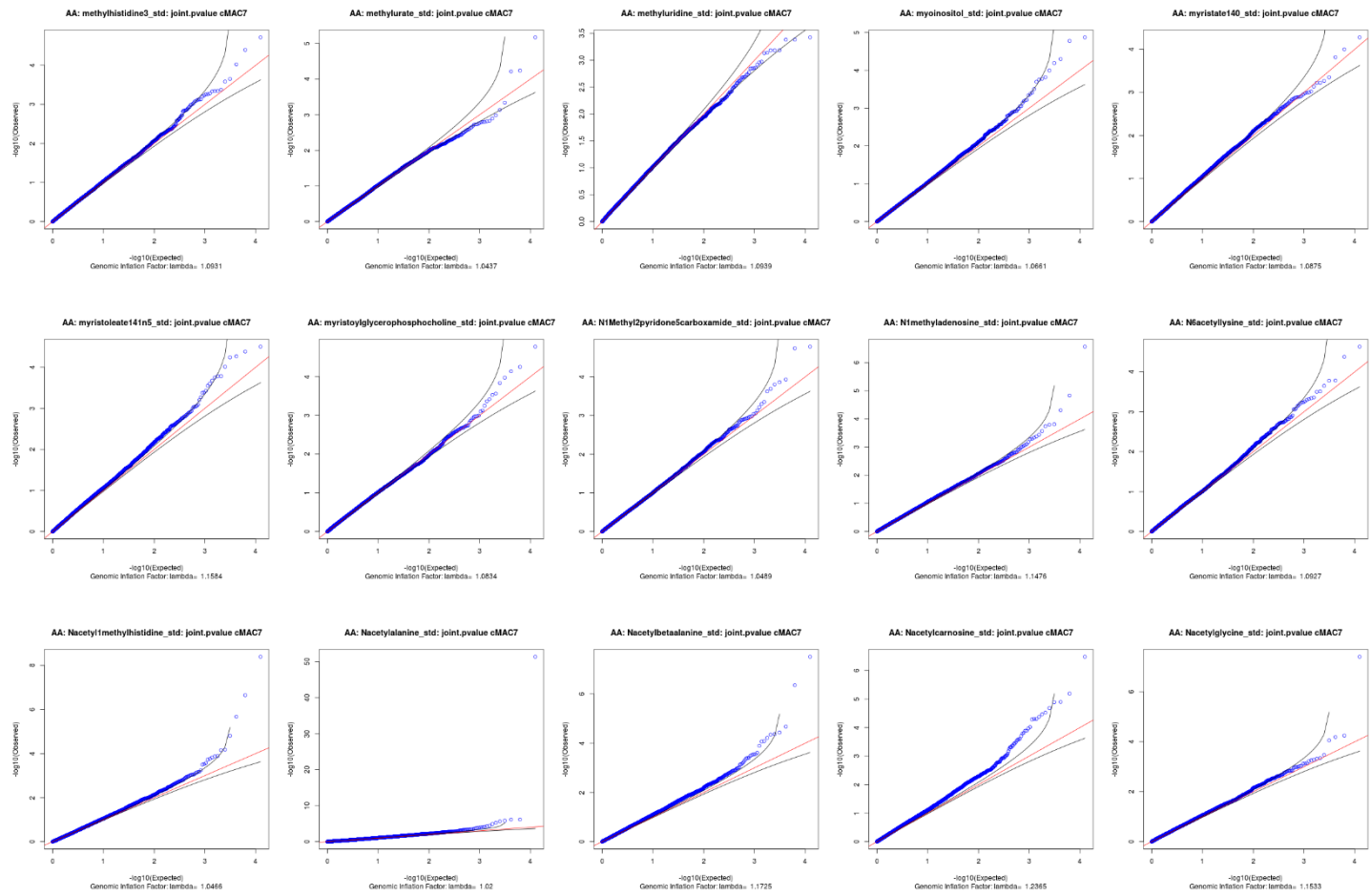


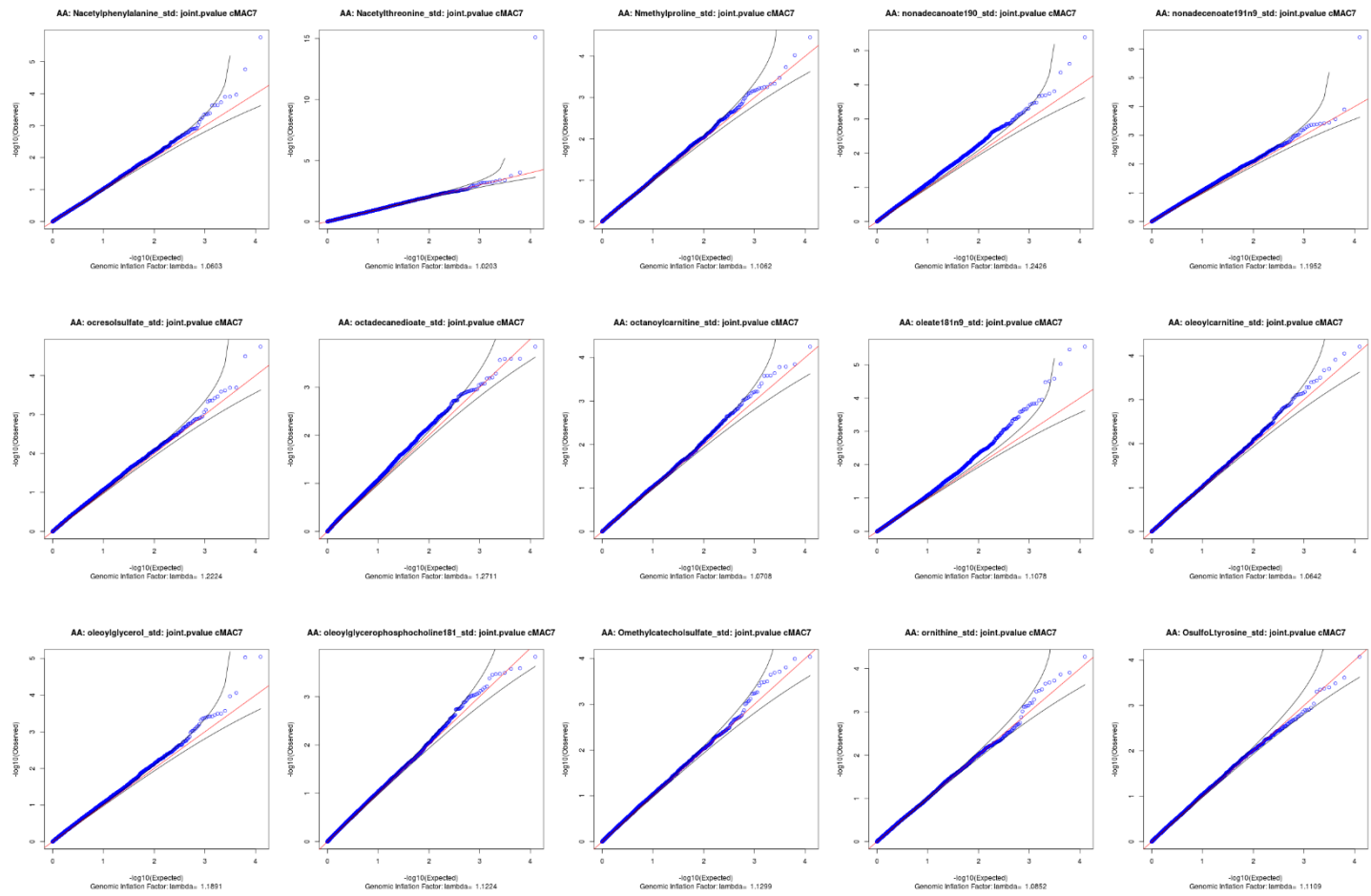


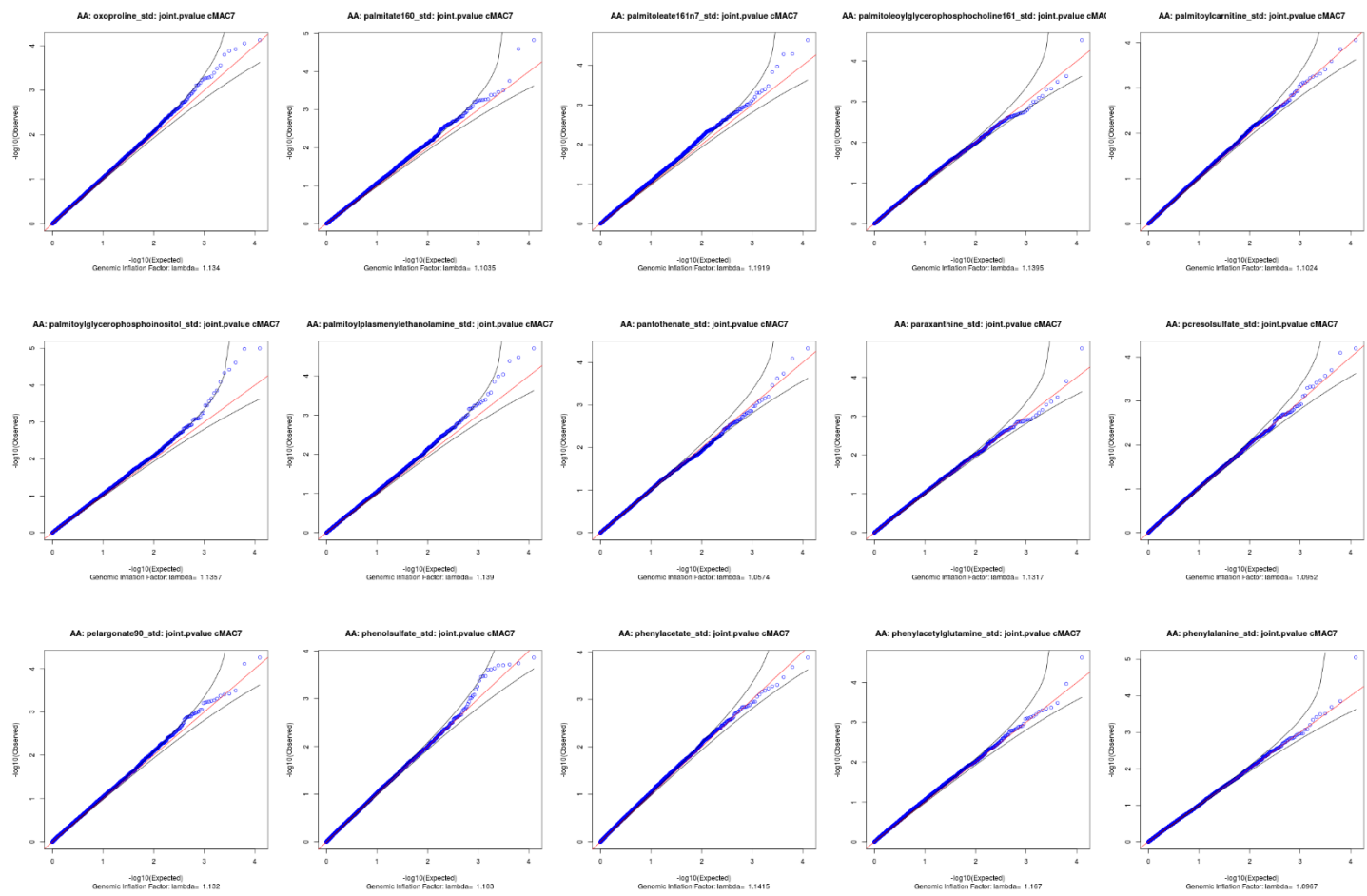


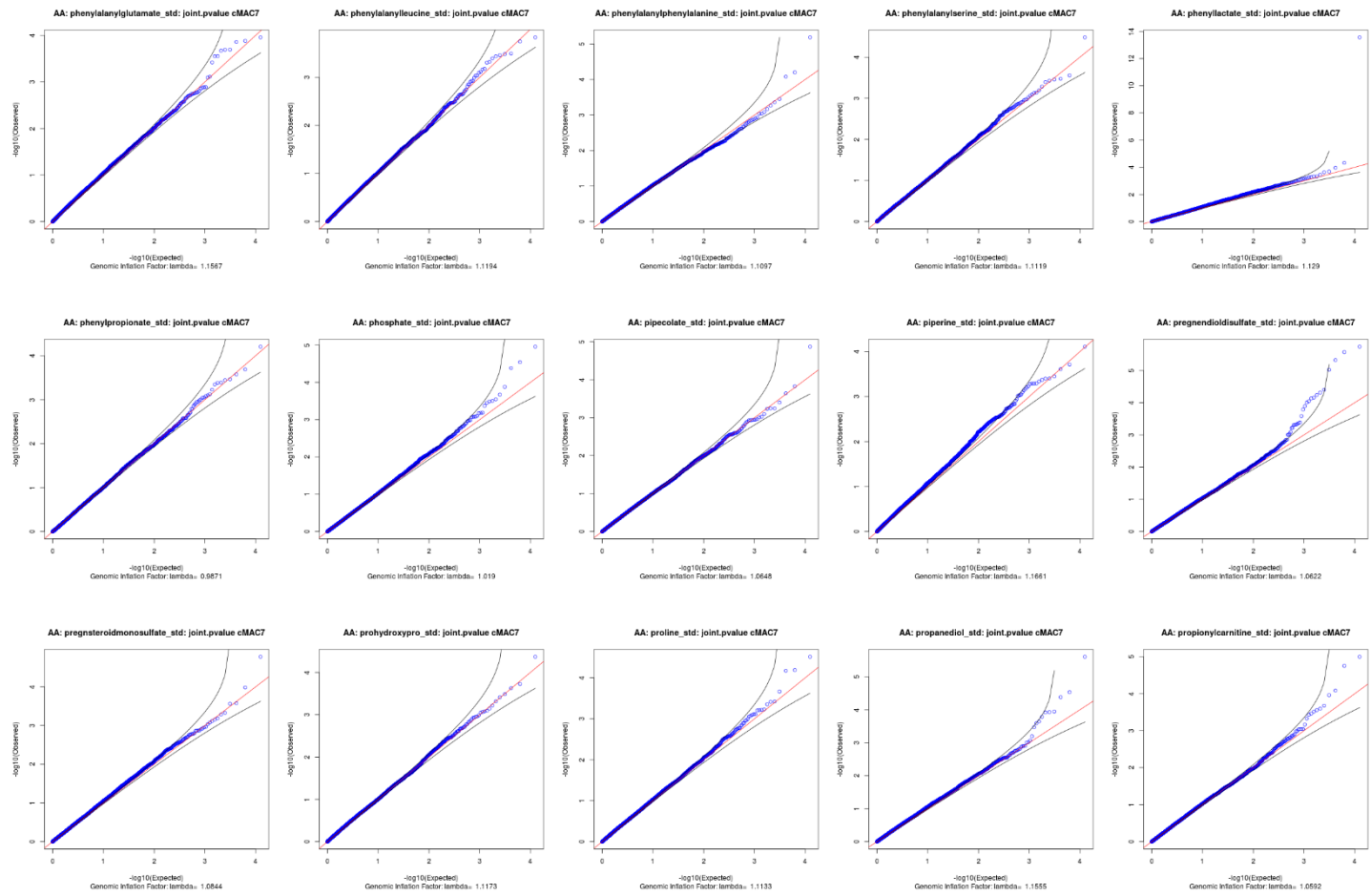


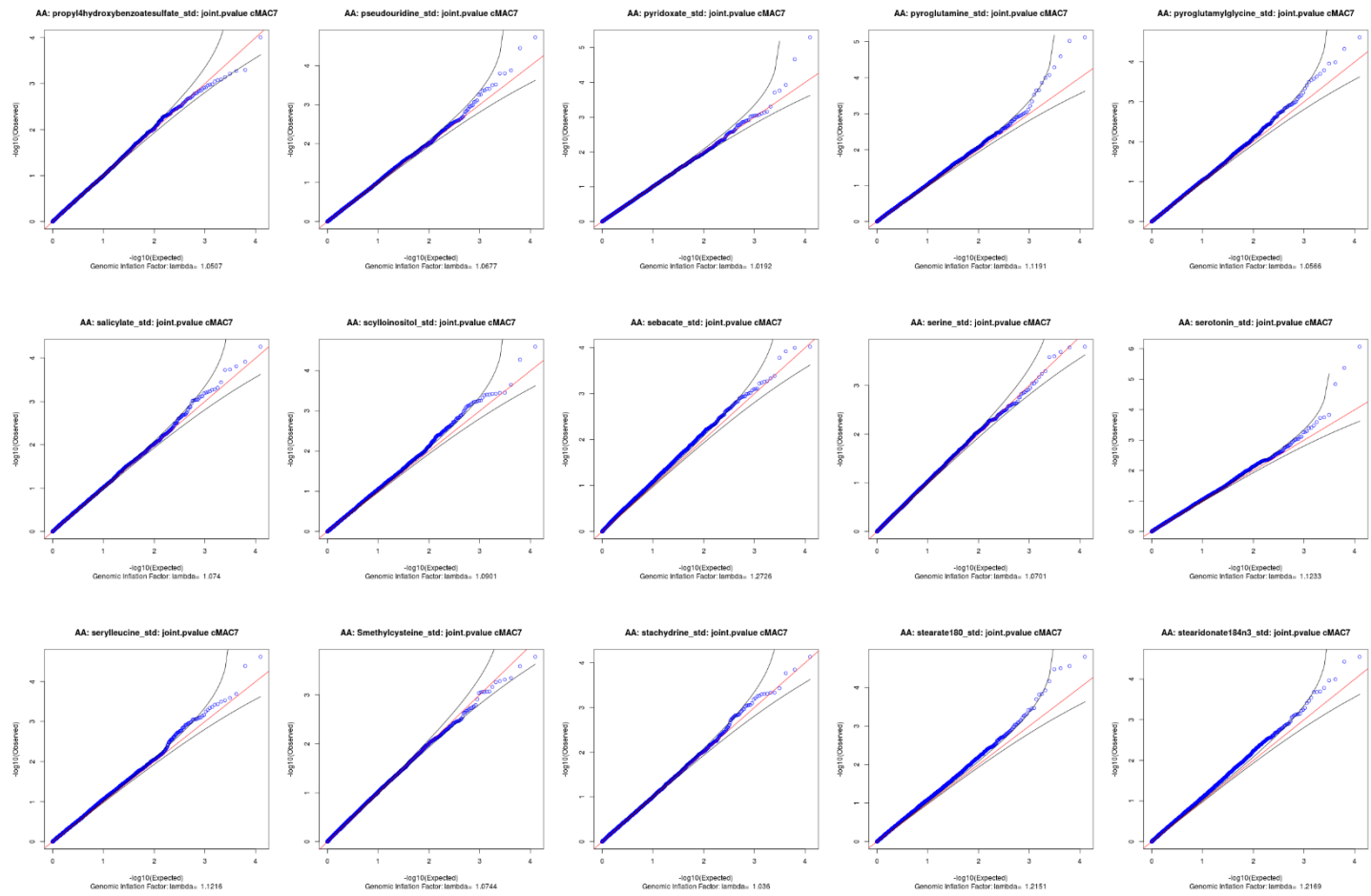


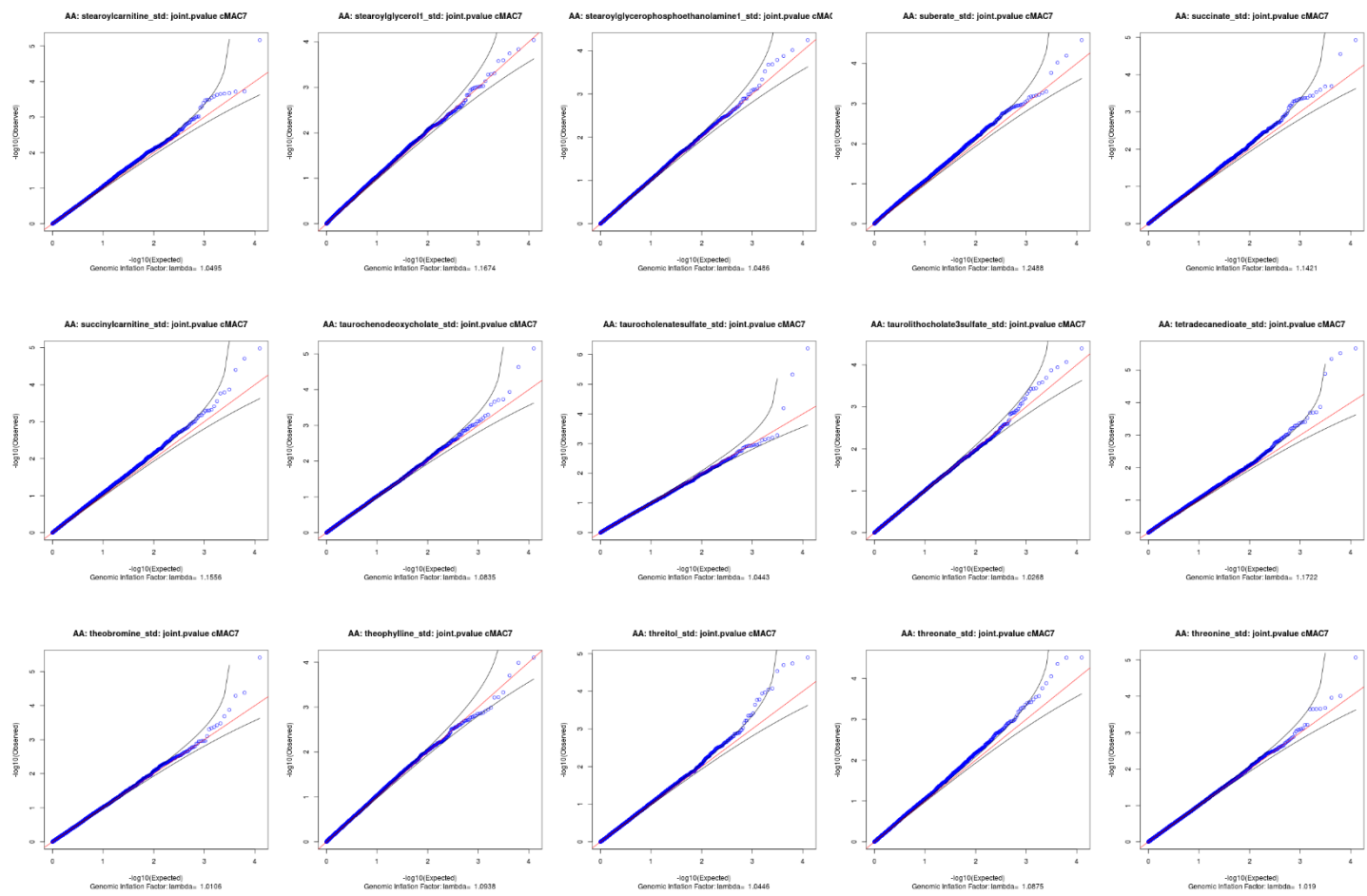


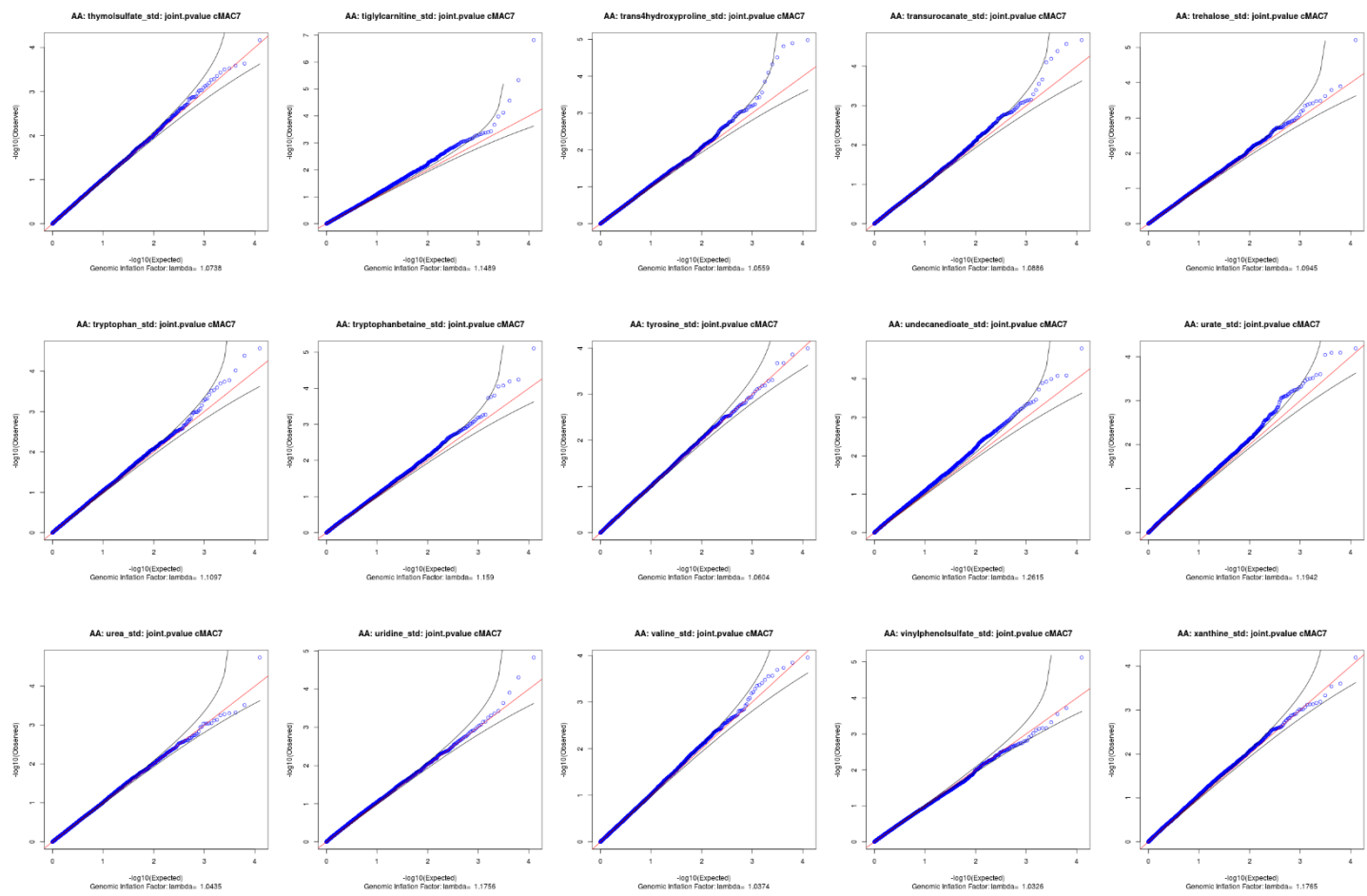












REFERENCES

1. National Center for Health Statistics. Health, United States, 2015: with special feature on racial and ethnic health disparities. 2016.
2. Lam CS, McEntegart M, Claggett B, Liu J, Skali H, Lewis E, et al. Sex differences in clinical characteristics and outcomes after myocardial infarction: insights from the Valsartan in Acute Myocardial Infarction Trial (VALIANT). *European journal of heart failure*. 2015;17(3):301-12.
3. Vaccarino V, Parsons L, Every NR, Barron HV, Krumholz HM. Sex-based differences in early mortality after myocardial infarction. *National Registry of Myocardial Infarction 2 Participants*. *The New England journal of medicine*. 1999;341(4):217-25.
4. Canto JG, Goldberg RJ, Hand MM, et al. Symptom presentation of women with acute coronary syndromes: Myth vs reality. *Archives of Internal Medicine*. 2007;167(22):2405-13.
5. Benjamin EJ, Blaha MJ, Chiuve SE, Cushman M, Das SR, Deo R, et al. Heart disease and stroke statistics—2017 update: a report from the American Heart Association. *Circulation*. 2017;135(10):e146-e603.
6. Barrett-Connor E. Sex Differences in Coronary Heart Disease: Why Are Women So Superior? The 1995 Ancel Keys Lecture. *Circulation*. 1997;95(1):252-64.

7. Pilote L, Dasgupta K, Guru V, Humphries KH, McGrath J, Norris C, et al. A comprehensive view of sex-specific issues related to cardiovascular disease. *Canadian Medical Association Journal*. 2007;176(6):S1-S44.
8. Ventura-Clapier R, Dworatzek E, Seeland U, Kararigas G, Arnal J-F, Brunelleschi S, et al. Sex in basic research: concepts in the cardiovascular field. *Cardiovascular Research*. 2017;113(7):711-24.
9. Vitale C, Mendelsohn ME, Rosano GMC. Gender differences in the cardiovascular effect of sex hormones. *Nat Rev Cardiol*. 2009;6(8):532-42.
10. Yamada Y, Izawa H, Ichihara S, Takatsu F, Ishihara H, Hirayama H, et al. Prediction of the risk of myocardial infarction from polymorphisms in candidate genes. *New England Journal of Medicine*. 2002;347(24):1916-23.
11. McCarthy JJ, Meyer J, Moliterno DJ, Newby LK, Rogers WJ, Topol EJ. Evidence for substantial effect modification by gender in a large-scale genetic association study of the metabolic syndrome among coronary heart disease patients. *Human genetics*. 2003;114(1):87-98.
12. Silander K, Alanne M, Kristiansson K, Saarela O, Ripatti S, Auro K, et al. Gender Differences in Genetic Risk Profiles for Cardiovascular Disease. *PLOS ONE*. 2008;3(10):e3615.
13. Shungin D, Winkler TW, Croteau-Chonka DC, Ferreira T, Locke AE, Mägi R, et al. New genetic loci link adipose and insulin biology to body fat distribution. *Nature*. 2015;518:187.

14. Liu LY, Schaub MA, Sirota M, Butte AJ. Sex differences in disease risk from reported genome-wide association study findings. *Human Genetics*. 2012;131(3):353-64.
15. Nikpay M, Goel A, Won H-H, Hall LM, Willenborg C, Kanoni S, et al. A comprehensive 1000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nature genetics*. 2015;47(10):1121-30.
16. The International Consortium for Blood Pressure Genome-Wide Association S. Genetic Variants in Novel Pathways Influence Blood Pressure and Cardiovascular Disease Risk. *Nature*. 2011;478(7367):103-9.
17. Traylor M, Farrall M, Holliday EG, Sudlow C, Hopewell JC, Cheng Y-C, et al. Genetic risk factors for ischaemic stroke and its subtypes (the METASTROKE Collaboration): a meta-analysis of genome-wide association studies. *The Lancet Neurology*. 2012;11(11):951-62.
18. Schunkert H, König IR, Kathiresan S, Reilly MP, Assimes TL, Holm H, et al. Large-scale association analyses identifies 13 new susceptibility loci for coronary artery disease. *Nature genetics*. 2011;43(4):333-8.
19. Kato N, Takeuchi F, Tabara Y, Kelly TN, Go MJ, Sim X, et al. Meta-analysis of genome-wide association studies identifies common variants associated with blood pressure variation in East Asians. *Nature genetics*. 2011;43(6):531-8.
20. Kim AM, Tingen CM, Woodruff TK. Sex bias in trials and treatment must end. *Nature*. 2010;465(7299):688-9.

21. Maher B. The case of the missing heritability. *Nature*. 2008;456(7218):18.
22. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. *Nature*. 2009;461(7265):747-53.
23. Ottman R. Gene–Environment Interaction: Definitions and Study Designs. *Preventive medicine*. 1996;25(6):764-70.
24. Mather K, Caligari PD. Genotype x environment interactions. IV. The effect of the background genotype. *Heredity*. 1976;36(1):41-8.
25. Thomas D. Gene–environment-wide association studies: emerging approaches. *Nature Reviews Genetics*. 2010;11(4):259-72.
26. Hunter DJ. Gene–environment interactions in human diseases. *Nature Reviews Genetics*. 2005;6(4):287-98.
27. Rao DC, Sung YJ, Winkler TW, Schwander K, Borecki I, Cupples LA, et al. Multiancestry Study of Gene–Lifestyle Interactions for Cardiovascular Traits in 610 475 Individuals From 124 Cohorts: Design and Rationale. *Circulation: Cardiovascular Genetics*. 2017;10(3):e001649.
28. Manning AK, Hivert M-F, Scott RA, Grimsby JL, Bouatia-Naji N, Chen H, et al. A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycaemic traits and insulin resistance. *Nature genetics*. 2012;44(6):659-69.

29. Winkler TW, Justice AE, Graff M, Barata L, Feitosa MF, Chu S, et al. The Influence of Age and Sex on Genetic Associations with Adult Body Size and Shape: A Large-Scale Genome-Wide Interaction Study. *PLOS Genetics*. 2015;11(10):e1005378.
30. Simino J, Shi G, Bis Joshua C, Chasman Daniel I, Ehret Georg B, Gu X, et al. Gene-Age Interactions in Blood Pressure Regulation: A Large-Scale Investigation with the CHARGE, Global BPgen, and ICBP Consortia. *The American Journal of Human Genetics*. 2014;95(1):24-38.
31. Taylor JY, Schwander K, Kardia SLR, Arnett D, Liang J, Hunt SC, et al. A Genome-wide study of blood pressure in African Americans accounting for gene-smoking interaction. *Scientific Reports*. 2016;6:18812.
32. Justice AE, Winkler TW, Feitosa MF, Graff M, Fisher VA, Young K, et al. Genome-wide meta-analysis of 241,258 adults accounting for smoking behaviour identifies novel loci for obesity traits. *Nature Communications*. 2017;8:14977.
33. Benjamin AM, Suchindran S, Pearce K, Rowell J, Lien LF, Guyton JR, et al. Gene by sex interaction for measures of obesity in the framingham heart study. *Journal of obesity*. 2011;2011:329038.
34. Tan A, Sun J, Xia N, Qin X, Hu Y, Zhang S, et al. A genome-wide association and gene-environment interaction study for serum triglycerides levels in a healthy Chinese male population. *Hum Mol Genet*. 2012;21(7):1658-64.

35. Li C, He J, Chen J, Zhao J, Gu D, Hixson JE, et al. Genome-Wide Gene-Potassium Interaction Analyses on Blood Pressure: The GenSalt Study (Genetic Epidemiology Network of Salt Sensitivity). *Circ Cardiovasc Genet*. 2017;10(6).
36. Singh A, Babyak MA, Nolan DK, Brummett BH, Jiang R, Siegler IC, et al. Gene by stress genome-wide interaction analysis and path analysis identify EBF1 as a cardiovascular and metabolic risk gene. *European journal of human genetics : EJHG*. 2015;23(6):854-62.
37. Feitosa MF, Kraja AT, Chasman DI, Sung YJ, Winkler TW, Ntalla I, et al. Novel genetic associations for blood pressure identified via gene-alcohol interaction in up to 570K individuals across multiple ancestries. *PLoS One*. 2018;13(6):e0198166.
38. Sung YJ, Winkler TW, Manning AK, Aschard H, Gudnason V, Harris TB, et al. An Empirical Comparison of Joint and Stratified Frameworks for Studying $G \times E$ Interactions: Systolic Blood Pressure and Smoking in the CHARGE Gene-Lifestyle Interactions Working Group. *Genetic Epidemiology*. 2016;40(5):404-15.
39. Das S, Forer L, Schonherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. *Nat Genet*. 2016;48(10):1284-7.
40. Kraft P, Yen Y-C, Stram DO, Morrison J, Gauderman WJ. Exploiting gene-environment interaction to detect genetic associations. *Human heredity*. 2007;63(2):111-9.

41. Manning AK, LaValley M, Liu C-T, Rice K, An P, Liu Y, et al. Meta-analysis of gene-environment interaction: joint estimation of SNP and SNP \times environment regression coefficients. *Genetic Epidemiology*. 2011;35(1):11-8.
42. Lin X, Lee S, Wu MC, Wang C, Chen H, Li Z, et al. Test for Rare Variants by Environment Interactions in Sequencing Association Studies. *Biometrics*. 2016;72(1):156-64.
43. Tzeng J-Y, Zhang D, Pongpanich M, Smith C, McCarthy MI, Sale MM, et al. Studying gene and gene-environment effects of uncommon and common variants on continuous traits: a marker-set approach using gene-trait similarity regression. *The American Journal of Human Genetics*. 2011;89(2):277-88.
44. Zhao G, Marceau R, Zhang D, Tzeng J-Y. Assessing gene-environment interactions for common and rare variants with binary traits using gene-trait similarity regression. *Genetics*. 2015;199(3):695-710.
45. Jiao S, Hsu L, Bézieau S, Brenner H, Chan AT, Chang-Claude J, et al. SBERIA: Set-Based Gene-Environment Interaction Test for Rare and Common Variants in Complex Diseases. *Genetic epidemiology*. 2013;37(5):452-64.
46. Jiao S, Peters U, Berndt S, Bézieau S, Brenner H, Campbell PT, et al. Powerful set-based gene-environment interaction testing framework for complex diseases. *Genetic epidemiology*. 2015;39(8):609-18.

47. Marceau R, Lu W, Holloway S, Sale MM, Worrall BB, Williams SR, et al. A Fast Multiple-Kernel Method With Applications to Detect Gene-Environment Interaction. *Genetic epidemiology*. 2015;39(6):456-68.
48. Su Y-R, Di C-Z, Hsu L. A unified powerful set-based test for sequencing data analysis of GxE interactions. *Biostatistics*. 2017;18(1):119-31.
49. Chen H, Meigs JB, Dupuis J. Incorporating Gene-Environment Interaction in Testing for Association with Rare Genetic Variants. *Human Heredity*. 2014;78(2):81-90.
50. Villas-Bôas SG, Mas S, Åkesson M, Smedsgaard J, Nielsen J. Mass spectrometry in metabolome analysis. *Mass spectrometry reviews*. 2005;24(5):613-46.
51. German JB, Hammock BD, Watkins SM. Metabolomics: building on a century of biochemistry to guide human health. *Metabolomics*. 2005;1(1):3-9.
52. Dettmer K, Aronov PA, Hammock BD. Mass spectrometry-based metabolomics. *Mass spectrometry reviews*. 2007;26(1):51-78.
53. Watson AD. Thematic review series: systems biology approaches to metabolic and cardiovascular disorders. Lipidomics: a global approach to lipid analysis in biological systems. *Journal of lipid research*. 2006;47(10):2101-11.
54. Koal T, Deigner H-P. Challenges in mass spectrometry based targeted metabolomics. *Current molecular medicine*. 2010;10(2):216-26.
55. Bain JR, Stevens RD, Wenner BR, Ilkayeva O, Muoio DM, Newgard CB. Metabolomics Applied to Diabetes Research. Moving From Information to Knowledge. 2009;58(11):2429-43.

56. Rhee EP, Gerszten RE. Metabolomics and Cardiovascular Biomarker Discovery. *Clinical chemistry*. 2012;58(1):139-47.
57. Floegel A, Stefan N, Yu Z, Mühlenbruch K, Drogan D, Joost H-G, et al. Identification of Serum Metabolites Associated With Risk of Type 2 Diabetes Using a Targeted Metabolomic Approach. *Diabetes*. 2013;62(2):639-48.
58. Zheng Y, Yu B, Alexander D, Mosley TH, Heiss G, Nettleton JA, et al. Metabolomics and Incident Hypertension Among Blacks. The Atherosclerosis Risk in Communities Study. 2013.
59. Menni C, Graham D, Kastenmüller G, Alharbi NHJ, Alsanosi SM, McBride M, et al. Metabolomic Identification of a Novel Pathway of Blood Pressure Regulation Involving Hexadecanedioate. *Hypertension*. 2015;66(2):422-9.
60. Wang TJ, Larson MG, Vasan RS, Cheng S, Rhee EP, McCabe E, et al. Metabolite profiles and the risk of developing diabetes. *Nature medicine*. 2011;17(4):448-53.
61. Shah SH, Sun J-L, Stevens RD, Bain JR, Muehlbauer MJ, Pieper KS, et al. Baseline metabolomic profiles predict cardiovascular events in patients at risk for coronary artery disease. *American Heart Journal*. 2012;163(5):844-50.e1.
62. Wang Z, Klipfell E, Bennett BJ, Koeth R, Levison BS, DuGar B, et al. Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. *Nature*. 2011;472(7341):57-63.

63. Zheng Y, Yu B, Alexander D, Manolio TA, Aguilar D, Coresh J, et al. Associations between metabolomic compounds and incident heart failure among African Americans: the ARIC Study. *American journal of epidemiology*. 2013:kwt004.
64. Tai ES, Tan MLS, Stevens RD, Low YL, Muehlbauer MJ, Goh DLM, et al. Insulin resistance is associated with a metabolic profile of altered protein metabolism in Chinese and Asian-Indian men. *Diabetologia*. 2010;53(4):757-67.
65. Rizza S, Copetti M, Rossi C, Cianfarani MA, Zucchelli M, Luzi A, et al. Metabolomics signature improves the prediction of cardiovascular events in elderly subjects. *Atherosclerosis*. 2014;232(2):260-4.
66. Ganna A, Salihovic S, Sundström J, Broeckling CD, Hedman ÅK, Magnusson PKE, et al. Large-scale Metabolomic Profiling Identifies Novel Biomarkers for Incident Coronary Heart Disease. *PLOS Genetics*. 2014;10(12):e1004801.
67. Vaarhorst AA, Verhoeven A, Weller CM, Bohringer S, Goralier S, Meissner A, et al. A metabolomic profile is associated with the risk of incident coronary heart disease. *Am Heart J*. 2014;168(1):45-52.e7.
68. Wurtz P, Havulinna AS, Soininen P, Tynkkynen T, Prieto-Merino D, Tillin T, et al. Metabolite profiling and cardiovascular event risk: a prospective study of 3 population-based cohorts. *Circulation*. 2015;131(9):774-85.
69. Paynter NP, Balasubramanian R, Giulianini F, Wang DD, Tinker LF, Gopal S, et al. Metabolic Predictors of Incident Coronary Heart Disease in Women. *Circulation*. 2018;137(8):841-53.

70. Harris WS, Miller M, Tighe AP, Davidson MH, Schaefer EJ. Omega-3 fatty acids and coronary heart disease risk: clinical and mechanistic perspectives. *Atherosclerosis*. 2008;197(1):12-24.
71. Harris WS, Bulchandani D. Why do omega-3 fatty acids lower serum triglycerides? *Curr Opin Lipidol*. 2006;17(4):387-93.
72. Davidson MH. Mechanisms for the hypotriglyceridemic effect of marine omega-3 fatty acids. *Am J Cardiol*. 2006;98(4a):27i-33i.
73. Gieger C, Geistlinger L, Altmaier E, De Angelis MH, Kronenberg F, Meitinger T, et al. Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. *PLoS Genet*. 2008;4(11):e1000282.
74. Shin S-Y, Fauman EB, Petersen A-K, Krumsiek J, Santos R, Huang J, et al. An atlas of genetic influences on human blood metabolites. *Nat Genet*. 2014;46(6):543-50.
75. Suhre K, Shin S-Y, Petersen A-K, Mohnen RP, Meredith D, Wägele B, et al. Human metabolic individuality in biomedical and pharmaceutical research. *Nature*. 2011;477(7362):10.1038/nature10354.
76. Krumsiek J, Suhre K, Evans AM, Mitchell MW, Mohnen RP, Milburn MV, et al. Mining the Unknown: A Systems Approach to Metabolite Identification Combining Genetic and Metabolic Information. *PLoS Genetics*. 2012;8(10):e1003005.
77. Draisma HHM, Pool R, Kobl M, Jansen R, Petersen A-K, Vaarhorst AAM, et al. Genome-wide association study identifies novel genetic variants contributing to variation in blood metabolite levels. *Nature communications*. 2015;6:7208-.

78. Ried JS, Shin S-Y, Krumsiek J, Illig T, Theis FJ, Spector TD, et al. Novel genetic associations with serum level metabolites identified by phenotype set enrichment analyses. *Human Molecular Genetics*. 2014;23(21):5847-57.
79. Rhee EP, Ho JE, Chen M-H, Shen D, Cheng S, Larson MG, et al. A Genome-Wide Association Study of the Human Metabolome in a Community-Based Cohort. *Cell metabolism*. 2013;18(1):130-43.
80. Rhee EP, Yang Q, Yu B, Liu X, Cheng S, Deik A, et al. An exome array study of the plasma metabolome. *Nature Communications*. 2016;7:12360.
81. Yu B, Li AH, Metcalf GA, Muzny DM, Morrison AC, White S, et al. Loss-of-function variants influence the human serum metabolome. *Science Advances*. 2016;2(8):e1600800.
82. Long T, Hicks M, Yu H-C, Biggs WH, Kirkness EF, Menni C, et al. Whole-genome sequencing identifies common-to-rare variants associated with human blood metabolites. *Nat Genet*. 2017;49(4):568-78.
83. Yu B, de Vries PS, Metcalf GA, Wang Z, Feofanova EV, Liu X, et al. Whole genome sequence analysis of serum amino acid levels. *Genome Biology*. 2016;17(1):237.
84. Yu B, Zheng Y, Alexander D, Morrison AC, Coresh J, Boerwinkle E. Genetic determinants influencing human serum metabolome among African Americans. *PLoS Genet*. 2014;10.

85. Feofanova EV, Yu B, Metcalf GA, Liu X, Muzny D, Below JE, et al. Sequence-Based Analysis of Lipid-Related Metabolites in a Multiethnic Study. *Genetics*. 2018;209(2):607-16.
86. Mittelstrass K, Ried JS, Yu Z, Krumsiek J, Gieger C, Prehn C, et al. Discovery of Sexual Dimorphisms in Metabolic and Genetic Biomarkers. *PLOS Genetics*. 2011;7(8):e1002215.
87. Krumsiek J, Mittelstrass K, Do KT, Stücker F, Ried J, Adamski J, et al. Gender-specific pathway differences in the human serum metabolome. *Metabolomics*. 2015;11(6):1815-33.
88. Hartiala JA, Wilson Tang WH, Wang Z, Crow AL, Stewart AFR, Roberts R, et al. Genome-wide association study and targeted metabolomics identifies sex-specific association of CPS1 with coronary artery disease. *Nature Communications*. 2016;7:10558.
89. Klein MS, Connors KE, Shearer J, Vogel HJ, Hittel DS. Metabolomics Reveals the Sex-Specific Effects of the SORT1 Low-Density Lipoprotein Cholesterol Locus in Healthy Young Adults. *Journal of Proteome Research*. 2014;13(11):5063-70.
90. Maric-Bilkan C, Arnold AP, Taylor DA, Dwinell M, Howlett SE, Wenger N, et al. Report of the National Heart, Lung, and Blood Institute Working Group on Sex Differences Research in Cardiovascular Disease. *Scientific Questions and Challenges*. 2016;67(5):802-7.

91. Greef Jvd, Stroobant P, Heijden Rvd. The role of analytical sciences in medical systems biology. *Current Opinion in Chemical Biology*. 2004;8(5):559-65.
92. Ordovas JM. Nutrigenetics, Plasma Lipids, and Cardiovascular Risk. *Journal of the American Dietetic Association*. 2006;106(7):1074-81.
93. Gorlov IP, Gorlova OY, Frazier ML, Spitz MR, Amos CI. Evolutionary evidence of the effect of rare variants on disease etiology. *Clinical genetics*. 2011;79(3):199-206.
94. Grove ML, Yu B, Cochran BJ, Haritunians T, Bis JC, Taylor KD, et al. Best practices and joint calling of the HumanExome BeadChip: the CHARGE Consortium. *PLoS One*. 2013;8(7):e68095.
95. Tishkoff SA, Williams SM. Genetic analysis of African populations: human evolution and complex disease. *Nature Reviews Genetics*. 2002;3:611.
96. Du X-J. Gender modulates cardiac phenotype development in genetically modified mice. *Cardiovascular Research*. 2004;63(3):510-9.
97. Djouadi F, Weinheimer CJ, Saffitz JE, Pitchford C, Bastin J, Gonzalez FJ, et al. A gender-related defect in lipid metabolism and glucose homeostasis in peroxisome proliferator-activated receptor alpha-deficient mice. *The Journal of clinical investigation*. 1998;102(6):1083-91.
98. Czubryt MP, McAnally J, Fishman GI, Olson EN. Regulation of peroxisome proliferator-activated receptor γ coactivator 1 α (PGC-1 α) and mitochondrial function by MEF2 and HDAC5. *Proceedings of the National Academy of Sciences*. 2003;100(4):1711-6.

99. Orlowska-Baranowska E, Gora J, Baranowski R, Stoklosa P, vel Betka LG, Pedzich-Placha E, et al. Association of the common genetic polymorphisms and haplotypes of the chymase gene with left ventricular mass in male patients with symptomatic aortic stenosis. *PloS one*. 2014;9(5):e96306.
100. Wu H, Roks AJ, Leijten FP, Garrelds IM, Musterd-Bhaggoe UM, van den Bogaerdt AJ, et al. Genetic variation and gender determine bradykinin type 1 receptor responses in human tissue: implications for the ACE-inhibitor-induced effects in patients with coronary artery disease. *Clinical Science*. 2014;126(6):441-9.
101. Morrison AC, Bare LA, Chambless LE, Ellis SG, Malloy M, Kane JP, et al. Prediction of coronary heart disease risk using a genetic risk score: the Atherosclerosis Risk in Communities Study. *American journal of epidemiology*. 2007;166(1):28-35.
102. Khera AV, Emdin CA, Drake I, Natarajan P, Bick AG, Cook NR, et al. Genetic risk, adherence to a healthy lifestyle, and coronary disease. *New England Journal of Medicine*. 2016;375(24):2349-58.
103. Inouye M, Abraham G, Nelson CP, Wood AM, Sweeting MJ, Dudbridge F, et al. Genomic risk prediction of coronary artery disease in 480,000 adults: implications for primary prevention. *Journal of the American College of Cardiology*. 2018;72(16):1883-93.

104. Gaziano JM, Concato J, Brophy M, Fiore L, Pyarajan S, Breeling J, et al. Million Veteran Program: a mega-biobank to study genetic influences on health and disease. *Journal of clinical epidemiology*. 2016;70:214-23.
105. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*. 2015;12(3):e1001779.
106. Sugiyama MG, Agellon LB. Sex differences in lipid metabolism and metabolic disease risk. *Biochemistry and cell biology = Biochimie et biologie cellulaire*. 2012;90(2):124-41.
107. Hunter DJ. Gene–environment interactions in human diseases. *Nature Reviews Genetics*. 2005;6(4):287.
108. VanderWeele TJ. Sample Size and Power Calculations for Additive Interactions. *Epidemiologic methods*. 2012;1(1):159-88