The Integration of Metabolomic and Genomic Data

By

Burcu F. Darst

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Epidemiology)

at the

UNIVERSITY OF WISCONSIN-MADISON

2018

Date of final oral examination: 05/29/2018

The dissertation is approved by the following members of the Final Oral Committee:

Corinne D. Engelman, Associate Professor, Population Health Sciences
Qiongshi Lu, Assistant Professor, Biostatistics and Medical Informatics
Sterling C. Johnson, Professor, Geriatrics and Gerontology
C. David Page, Professor, Biostatistics and Medical Informatics
Ron E. Gangnon, Professor, Population Health Sciences
Paul E. Peppard, Associate Professor, Population Health Sciences, Biostatistics and Medical Informatics

Dedication

To Mom and to Dan, whose encouragement and love this dissertation is a reflection of.

If I could pinpoint one factor that enabled me to get to where I am today, it would not be natural intellect or luck—it would be my mom. My mom always made my education her priority, and through example, she provided me with every tool I could ever need to succeed. She taught me the value of hard work, how to be strong and independent, served as my moral compass, and has always made sure I know how precious life is. Because I will never be able to repay my mom for the life she has enabled me to lead, it is to this woman—the biggest advocate of my education—that I dedicate this dissertation.

I cannot imagine getting through such a trying time of life without the constant support of my significant other, whose encouragement, wisdom, and patience somehow managed to lift the weight of graduate school. Despite the many states between us, he's been there through every milestone and every roadblock, sometimes within hours of notice. It is to Dan, my source of inspiration and my reason to work hard, that I also dedicate this dissertation.

Acknowledgements

My graduate and dissertation research were generously supported for three years by an NLM training grant to the Computation and Informatics in Biology and Medicine Training Program [NLM 5T15LM007359], led by Mark Craven, Louise Pape, and Karen Nafzger, whose support I am incredibly grateful for. This research was also supported by the NIH [R01AG054047, R01AG27161, UL1TR000427, and P2C HD047873], Helen Bader Foundation, Northwestern Mutual Foundation, Extendicare Foundation, and State of Wisconsin. I would like to thank the University of Wisconsin Madison Biotechnology Center Gene Expression Center for providing Illumina Infinium genotyping services. I especially thank the WRAP participants for their dedication to curing Alzheimer's disease.

I have had so much support to help me get to this point. I realized my passion for genomics research while working at the Scripps Translational Science Institute with Cinnamon Bloss. She was instrumental in shaping my career, and her strong encouragement and persuasion led me to pursue graduate studies. Upon arriving at the Department of Population Health Sciences, I was warmly welcomed by its students, staff, and faculty. The friendships I've developed through the department, especially with the girls from my incoming class, have provided so many needed mental breaks to climb, bike, snowboard, etc.—I feel very fortunate to have gone through graduate school with friends who made it so much more enjoyable. Thank you, Quinn, for breaking up the long days with your sense of humor and attempts to remind me how to make jokes. I also owe a debt of gratitude to my dissertation committee for offering me

their expertise and support throughout my dissertation research: Corinne Engelman, Qiongshi Lu, Sterling Johnson, Ron Gangnon, Paul Peppard, and David Page.

The unconditional love and support from my mom and dad have always been my pillars, and I cannot thank them enough for making me the person I am today. I am so grateful for my loving parents and brothers, my closest although physically distant friends, my incredibly supportive and inspiring significant other, and my new Midwest family for their continuous love and support, generous offers to help me get through the long days, and the amazing times we've had together over these past five years that remind me of what's most important in life.

Finally, this work would not have been possible without the generous and unwavering support of my advisor, Corinne Engelman. My choice in graduate school was driven almost entirely by my desire to work with Corinne, and it is one of the best decisions I've ever made. Corinne began advocating on my behalf even before I started graduate school, and she has since provided me with countless invaluable opportunities to advance my career. Advisors play a very special role in the life of a trainee, and I consider myself incredibly privileged to have an advisor who feels so passionately about mentoring and truly has her students' best interests at heart. I have felt so encouraged by Corinne's ability to not only be an amazing mentor and researcher, but to also have a fulfilling family life outside of work. Thank you so much, Corinne, for always thinking of me when opportunities arose, for your careful and thoughtful feedback in all of the work I've done with you, for generously and freely sharing your time with me in the many long meetings we've had, for being an advocate for me anytime I needed one, for sharing

your life experiences and offering advice, and for being my advisor, role model, and friend for these past five years.

Abstract

A longitudinal multi-omics examination of Alzheimer's disease (AD) risk factors in the years prior to AD diagnosis is critical to better understand, prevent, diagnose, and treat the disease. Using longitudinal plasma samples from the Wisconsin Registry for Alzheimer's Prevention (WRAP), we investigated relationships between genomics, metabolomics, and risk factors for AD, including age, sex, cognitive function, and cerebral spinal fluid (CSF) markers of beta-amyloid deposition. The research conducted in this dissertation offers an important improvement regarding our knowledge of risk factors of AD in several key ways. First, it offers a deeper knowledge about the impact of aging and sex, two known risk factors for AD, on a panel of plasma metabolite levels, while also describing how heritable each of those metabolites are. Second, it has identified several metabolites and metabolic pathways, particularly those related to fatty acid metabolism and cysteine metabolism, to be associated with an aspect of cognition that declines early in the AD trajectory. Third, using a data-driven approach, it identifies many potential gene-metabolite relationships that could be associated with AD risk factors, and shows that most plasma and CSF metabolite levels are weakly correlated with each other. This information could be used to guide further investigations of specific biological mechanisms that could lead to the development of AD, and ultimately, could be used to improve predictive models of this devastating disease.

Table of contents

Introduction and literature review

A longitudinal multi-omics examination of Alzheimer's disease (AD) risk factors in the years prior to AD diagnosis is critical to better understand, prevent, diagnose, and treat the disease. First, studying individuals at risk for AD (due to a parental history of the disease, for example) in the decade or two prior to typical onset provides a window into the preclinical phase of the disease and the factors contributing to risk for or protection from AD. Second, the pathophysiology of AD is complex and integration of multiple types of data is key to understanding this complexity (Ritchie, Holzinger, Li, Pendergrass, & Kim, 2015; Shah & Newgard, 2015). Finally, several metabolomics studies of AD have been published (Enche Ady et al., 2017), but the examination of metabolomic profiles prior to AD diagnosis is important to identify and distinguish predictive versus diagnostic metabolomic profiles, since the disease process itself influences metabolites. This dissertation utilizes rich longitudinal neuropsychological, biometric, and biofluids data from the Wisconsin Registry for Alzheimer's Prevention (WRAP) cohort to investigate longitudinal metabolic changes and genomic factors that could influence risk for AD. This contribution will be significant because identification of novel genes, metabolic profiles, and interactions between the two is expected to lead to: identification of new pathways to target with therapeutic agents; novel risk prediction and diagnostic tools for early AD pathophysiology; and dietary, lifestyle, or environmental interventions to prevent or impede AD.

The proposed research has several innovative aspects with respect to genomics. First, it focuses on families with a strong history of AD, a more powerful approach for low frequency variants. While these variants are less common in the general population,

such variants will be found in about half the individuals from a family in which that variant is increasing risk for the disease. Therefore, while studies of unrelated cases and controls were successful for genome-wide association studies (GWAS) of common variants, studies of families with a history of the disease are ideal for the discovery of low frequency variants with a more substantial effect on disease risk. In previous research, we compared several statistical approaches designed to maximize the power of family-based genomic analyses of rare variants and found that all performed well (Darst & Engelman, 2016).

Further proof of this concept is provided by Cruchaga et al. who identified a low frequency variant in *PLD3* that segregated with AD status in 2 of the 14 families studied (Cruchaga et al., 2013). This variant was enriched in AD cases with a strong family history of dementia, with an OR of 6.7, and doubled the risk for AD in seven independent case-control studies with a total of over 11,000 individuals. In the WRAP sample, we found this variant to be twice as common in those with a parental history of AD than in those without a parental history. Moreover, in longitudinal analysis of WRAP participants with up to 12 years of follow-up data, we found that carriers of this variant had lower cognitive function in all 6 of the domains examined and the effect of carrying the risk allele was 2-6 times that of carrying the *APOE* ε4 allele (Engelman et al., 2017). Our proposed research builds on the success of this groundbreaking study, but is distinct and innovative in that it focuses on cognitive trajectories prior to AD diagnosis. This is an important time period since the pathophysiology of AD begins decades before the clinical diagnosis. And we have shown the heritability of certain aspects of cognitive

function to be as high as 0.52 (Darst et al., 2015), implying an important role of genetics in cognition.

With respect to metabolomics, most studies of AD to date have focused on the cross-sectional metabolic profile in one biofluid from cases versus controls(Enche Ady et al., 2017; Trushina & Mielke, 2014). The WRAP cohort is a longitudinal study of individuals at risk for AD due to parental history, but initially asymptomatic, with serial collection of both blood and cerebrospinal fluid (CSF) (Johnson et al., 2018; Sager, Hermann, & La Rue, 2005), allowing an innovative approach to investigate the metabolomics of AD. Examining the metabolome prior to AD diagnosis is important because the disease process can influence the metabolome. Moreover, it is unknown how the metabolome changes over time in those with and without preclinical changes associated with AD (beta-amyloid (Aβ) deposition and cognitive decline). Finally, it is critical to examine both the blood, which is much more practical to measure in a clinical setting, and CSF, which surrounds the brain and may be more relevant in AD pathophysiology.

Finally, metabolomics offers an unprecedented opportunity in epidemiologic research to measure environmental and exogenous exposures with far less measurement error than with standard epidemiologic questionnaires (2016; Tzoulaki, Ebbels, Valdes, Elliott, & Ioannidis, 2014). The integration of this rich longitudinal metabolomic data with genomic data allows for the discovery of novel pathways involved in AD pathophysiology.

Specific aims

A longitudinal multi-omics examination of AD risk factors in the years prior to Alzheimer's disease (AD) diagnosis is critical to better understand, prevent, diagnose, and treat the disease. Integrating genomic and metabolomic data will enable more thorough and comprehensive modeling of AD risk. Continuing to focus on single-data-type study designs, which do not accurately reflect the complexity of AD, will hinder progress in our understanding of this disease and of effective prevention, diagnosis, and treatment.

Our long-term goal is to elucidate complex interactions involved in AD pathology using a comprehensive set of omics data in order to inform precision medicine for AD. As a first step, this dissertation is focused on the role of genomics and metabolomics in risk factors for AD in the Wisconsin Registry for Alzheimer's Prevention (WRAP), which has collected over 16 years of serial cognitive data and plasma, DNA, and cerebrospinal fluid on a subset. Our central hypothesis is that genetic variants, metabolites, and interactions between these will influence longitudinal risk factors for AD. Integrating genomic and metabolomic data will enable modeling and identification of the complex interplay of genes and metabolites involved in AD pathology, which is necessary to achieve the goal of precision medicine for AD. We will test our central hypothesis by executing the following aims:

Aim 1: Characterize the metabolomics of aging and sex and the heritability of the metabolome.

H1a: Plasma metabolites will change with age and differ by sex. Further, plasma metabolite trajectories will differ by sex.

H1b: Plasma metabolites are influenced by a complex combination of genomic and environmental factors.

Aim 2: Determine the metabolomic profile of longitudinal cognitive decline.

H2: A set of plasma metabolites will be longitudinally associated with the earliest aspects of cognitive decline.

Aim 3: Detect complex relationships associated with AD risk factors by integrating genomic and metabolomic data.

H3a: One or more plasma or CSF metabolites, particularly those with low heritability (and a stronger environmental contribution), will interact with or mediate one or more genetic variants to steepen cognitive decline and increase Aβ deposition.

H3b: A set of plasma and CSF metabolites will be highly correlated between tissues, while others will have low correlations.

The findings from the proposed study are expected to have an important positive impact because integration of both genomic and metabolomic data lays the groundwork for precision medicine and provides a more comprehensive picture of the etiology of AD.

Methods

WRAP Study Design

WRAP is an ongoing longitudinal study of over 1,500 initially asymptomatic adults, age 40-85 at baseline, enriched for a parental history of dementia due to AD. This was determined by review of medical records of the parent, including autopsy records in some cases, or by administering a dementia questionnaire to the adult child regarding the parent with AD. Details of the study design and methods have been published (Johnson et al., 2018; Sager et al., 2005). Recruitment began in 2001; the study protocol targeted a window of four years between the baseline and Wave two visits, with ongoing follow up every two years. Wave six visits began in September 2016. The average follow up is nine years, with a maximum of 16 years thus far and an 81% retention rate. Assessments include a neuropsychological test battery, blood laboratory values, and extensive questionnaires ranging from medications and medical history to physical activity and diet. Blood specimens including plasma, serum, and whole blood for DNA extraction are collected and archived at each visit. A subset of the cohort is enrolled in one or more linked biomarker studies including lumbar puncture (LP) for collection of CSF (Johnson et al., 2014). The study design included participation of multiple siblings per family.

Neuropsychological assessment

The WRAP cognitive test battery consist of standardized, widely used clinical neuropsychological tests, which were selected to provide a comprehensive estimate of cognitive abilities, with an emphasis on abilities most likely to be affected in early-stage

AD. Cognitive composite scores were calculated in WRAP to reduce the number of outcome measures to a small number of reliable cognitive scores (Clark et al., 2016). These were calculated by standardizing (~N [0,1]) into z-scores, neuropsychological tests, using means and standard deviations obtained from the whole baseline sample, then averaging tests that were relevant to a particular aspect of cognition. WRAP participants have up to six time points of cognitive data.

Plasma collection

Fasting blood samples for this study were drawn the morning of each study visit, which was also the day cognitive testing was completed. Blood was collected in 10 mL ethylenediaminetetraacetic acid (EDTA) vacutainer tubes. They were immediately placed on ice, and then centrifuged at 3000 revolutions per minute for 15 minutes at room temperature. Plasma was pipetted off within one hour of collection. Plasma samples were aliquoted into 1.0 mL polypropylene cryovials and placed in -80°C freezers within 30 minutes of separation. Samples were never thawed before being shipped overnight on dry ice to Metabolon, Inc. (Durham, NC), where they were again stored in -80°C freezers and thawed once before testing.

Cerebrospinal fluid (CSF) collection and biomarker measurement

CSF has been collected in a subset of WRAP participants (~25%) via lumbar puncture (LP) in the morning after a 12-hour fast using a Sprotte 25- or 24-gauge spinal needle at the L3/4 or L4/5 interspace using gentle extraction into polypropylene syringes. CSF (22 mL) was then gently mixed, and centrifuged at 2000g for 10 minutes. Supernatants

were frozen in 0.5 mL aliquots in polypropylene tubes and stored at −80°C, as previously described (Darst et al., 2017). WRAP participants have up to four CSF samples.

The following have been measured or calculated in the WRAP samples: $A\beta_{42}$, total-tau (T-tau), phosphorylated-tau (P-tau), and ratios $A\beta_{42}/A\beta_{40}$, T-tau/$A\beta_{42}$, and P-tau/$A\beta_{42}$. Previous literature has demonstrated that $A\beta_{42}/A\beta_{40}$ is decreased in AD (Blennow, 2004; Skoog et al., 2003) and that the ratio normalizes the $A\beta_{42}$ concentration to a measure of overall amyloidogenic processing by amyloid precursor protein, making it possible to detect low $A\beta_{42}$ in high Aβ producers and vice versa (Lewczuk, Lelental, Spitzer, Maler, & Kornhuber, 2015). Similarly, studies have shown that combinations of tau and $A\beta_{42}$ improve sensitivity and specificity (Blennow, 2004; De Meyer et al., 2010; Duits et al., 2014), and predict subjective cognitive decline (Stomrud, Hansson, Blennow, Minthon, & Londos, 2007).

CSF $A\beta_{42}$, T-tau, and P-tau were quantified with sandwich ELISAs (INNOTEST β-amyloid1-42, hTAU-Ag, and Phospho-Tau[181P], respectively; Fujirebio Europe, Ghent, Belgium). For the $A\beta_{42}/A\beta_{40}$ ratio, CSF levels of $A\beta_{42}$ and $A\beta_{40}$ (a less amyloidogenic Aβ fragment as compared to $A\beta_{42}$) were quantified by electrochemiluminescence (ECL) using an Aβ triplex assay (MSD Human Aβ peptide Ultra-Sensitive Kit, Meso Scale Discovery, Gaithersburg, MD). Intra-assay coefficients of variation were below 10%.

Metabolomic profiling and quality control

An untargeted plasma metabolomics analysis was performed by Metabolon, Inc. using Ultrahigh Performance Liquid Chromatography-Tandom Mass Spectrometry (UPLC-MS/MS). Quantification was performed as previously described (Evans et al., 2014); details are described in Appendix A1. Metabolites within nine super pathways were identified: amino acids, carbohydrates, cofactors and vitamins, energy, lipids, nucleotides, partially characterized molecules, peptides, and xenobiotics.

Up to three longitudinal plasma samples were available for each participant. Metabolites with an interquartile range of zero (*i.e.*, those with very low or no variability) were excluded from analyses (n=178 metabolites). After removing these metabolites, samples were missing a median of 11.7% metabolites, while metabolites were missing a median of 1.2% of samples. Missing metabolite values were imputed to the lowest level of detection for each metabolite. Metabolite values were median-scaled and log-transformed to normalize metabolite distributions(van den Berg, Hoefsloot, Westerhuis, Smilde, & van der Werf, 2006). If a participant reported that they did not fast or withhold medications and caffeine for at least eight hours, the sample was excluded from analyses (n=159 samples). A total of 1,097 metabolites among 2,316 samples remained for analyses.

DNA collection and genomics quality control

DNA was extracted from whole blood samples using the PUREGENE® DNA Isolation Kit (Gentra Systems, Inc., Minneapolis, MN). DNA concentrations were quantified using the Invitrogen™ Quant-iT™ PicoGreen™ dsDNA Assay Kit (Thermo Fisher Scientific, Inc., Hampton, NH) analyzed on the Synergy 2 Multi-Detection

Microplate Reader (Biotek Instruments, Inc., Winooski, VT). Samples were diluted to 50 ng/ul following quantification.

A total of 1,340 samples were genotyped using the Illumina Multi-Ethnic Genotyping Array at the University of Wisconsin Biotechnology Center (Appendix A2). Thirty-six blinded duplicate samples were used to calculate a concordance rate of 99.99%, and discordant genotypes were set to missing. Sixteen samples missing >5% of variants were excluded, while 35,105 variants missing in >5% of individuals were excluded. No samples were removed due to outlying heterozygosity. Six samples were excluded due to inconsistencies between self-reported and genetic sex.

Due to sibling relationships in the WRAP cohort, genetic ancestry was assessed using Principal Components Analysis in Related Samples (PC-AiR), a method that makes robust inferences about population structure in the presence of relatedness (Conomos, Miller, & Thornton, 2015). This approach included several iterative steps and was based on 63,503 linkage disequilibrium (LD) pruned (r^2<0.10) and common (MAF>0.05) variants, using the 1000 Genomes data as reference populations (Genomes Project et al., 2015). First, kinship coefficients (KCs) were calculated between all pairs of individuals using genomic data with the Kinship-based Inference for Gwas (KING)-robust method (Manichaikul et al., 2010). PC-AiR was used to perform principal components analysis (PCA) on the reference populations along with a subset of unrelated individuals identified by the KCs. Resulting principal components (PCs) were used to project PC values onto the remaining related individuals. All PCs were then used to recalculate the KCs taking ancestry into account using the PC-Relate method, which estimates KCs robust to population structure (Conomos, Reiner, Weir, &

Thornton, 2016). PCA was performed again using the updated KCs, and KCs were also estimated again using updated PCs. The resulting PCs identified 1,198 WRAP participants whose genetic ancestry was primarily of European descent. This procedure was repeated within this subset of participants (excluding 1000 Genomes individuals) to obtain PC estimates used to adjust for population stratification in subsequent genomic analyses. Among European descendants, 160 variants were not in Hardy-Weinberg equilibrium (HWE) and 327,064 were monomorphic and thus, removed.

A total of 1,294,660 bi-allelic autosomal variants among 1,198 European descendants remained for imputation, which was performed with the Michigan Imputation Server v1.0.3 (Das et al., 2016), using the Haplotype Reference Consortium (HRC) v. r1.1 2016 (McCarthy et al., 2016) as the reference panel and Eagle2 v2.3 (Loh et al., 2016) for phasing. Prior to imputation, the HRC Imputation Checking Tool (Rayner, Robertson, Mahajan, & McCarthy, 2016) was used to identify variants that did not match those in HRC, were palindromic, differed in MAF>0.20, or that had non-matching alleles when compared to the same variant in HRC, leaving 898,220 for imputation. A total of 39,131,578 variants were imputed. Variants with a quality score R^2<0.80, MAF<0.001, or that were out of HWE were excluded, leaving 10,400,394 imputed variants. These were combined with the genotyped variants, leading to 10,499,994 imputed and genotyped variants for analyses. Data cleaning and file preparation were completed using PLINK v1.9 (Chang et al., 2015) and VCFtools v0.1.14 (Danecek et al., 2011). Coordinates are based on GRCh37 assembly hg19.

Chapter 1. Longitudinal metabolomics of aging and sex

Abstract

Understanding how metabolites are longitudinally influenced by age and sex could facilitate the identification of metabolomic profiles and trajectories that indicate disease risk. We investigated the metabolomics of age and sex using longitudinal plasma samples from the Wisconsin Registry for Alzheimer's Prevention (WRAP), a cohort of participants who were dementia free at enrollment. Metabolomic profiles were quantified for 2,316 fasting plasma samples among 1,187 participants, each with up to three study visits. Of 1,097 metabolites tested, 608 (55.4%) were associated with age and 680 (62.0%) with sex after correcting for multiple testing. Approximately twice as many metabolites were associated with age in stratified analyses of women versus men, and 63 metabolite trajectories significantly differed by sex, most notably including sphingolipids, which tended to increase in women and decrease in men with age. Using genome-wide genotyping, we also report the heritabilities of metabolites investigated, which ranged dramatically (0.2–99.2%); however, the median heritability of 36.2% suggests that many metabolites are highly influenced by a complex combination of genomic and environmental influences. These findings offer a more profound understanding of the aging process and may inform many new hypotheses regarding the role metabolites play in healthy aging.

Introduction

The metabolome represents the functional endpoints of a complex network of biological events, including genomic, epigenomic, transcriptomic, proteomic, and environmental factors (Deidda, Piras, Bassareo, Dessalvi, & Mercuro, 2015). Being the final downstream product, the metabolome is the closest to the phenotype among the biological systems (Horgan & Kenny, 2011), making it particularly relevant to investigate. Age is known to be the single largest risk factor of most prevalent diseases in developed countries (Niccoli & Partridge, 2012). A better understanding of how the metabolome changes with age could further reveal the mechanisms by which age influences disease risk and could facilitate the identification of high-risk metabolomic profiles that are suggestive of the early stages of particular diseases.

Previous studies have provided important evidence that age and sex influence the metabolome (Dunn et al., 2015; Krumsiek et al., 2015; Menni et al., 2013; Mittelstrass et al., 2011; Rist et al., 2017; Z. Yu et al., 2012). While informative, these studies are limited by their cross-sectional designs and the relatively small number of metabolites assessed by most. According to the Human Metabolite Database (HMDB) v4.0, there are an estimated 25,424 blood metabolites (Wishart et al., 2018). However, due to current technical limitations in identifying and quantifying metabolites, most recent studies have only been able to confidently capture ~100-600 of these. A larger panel of metabolites will provide a more comprehensive understanding of the metabolomics of age and sex. Further, in order to assess the metabolomics of aging, it is crucial to use a longitudinal study design that can capture age-related phenomena, particularly due to the high variability of metabolites (Makinen & Ala-Korpela, 2016).

Longitudinal assessments also facilitate the examination of metabolite trajectories, which can address important biological questions.

Using longitudinal plasma samples from the Wisconsin Registry for Alzheimer's Prevention (WRAP), we investigated how a large panel of metabolites is influenced by age and sex, and whether metabolite trajectories vary by sex. To facilitate the interpretation of our results and determine whether identified metabolites are more strongly influenced by genetic or environmental factors, we used genome-wide genotyping data to assess the heritability (h^2) of metabolites.

Methods

Participants

Study participants were from WRAP, a longitudinal study of initially dementia free middle-aged adults that allows for the enrollment of siblings and is enriched for a parental history of Alzheimer's disease (AD). Further details of the study design and methods used have been previously described (Johnson et al., 2018; Sager et al., 2005). For the current analyses, follow-up occurred every two years. This study was conducted with the approval of the University of Wisconsin Institutional Review Board and all subjects provided signed informed consent before participation.

Plasma collection and sample handling

Fasting blood samples for this study were drawn the morning of each study visit. Blood was collected in 10 mL ethylenediaminetetraacetic acid (EDTA) vacutainer tubes. They were immediately placed on ice, and then centrifuged at 3000 revolutions per minute for 15 minutes at room temperature. Plasma was pipetted off within one hour of

collection. Plasma samples were aliquoted into 1.0 mL polypropylene cryovials and placed in -80°C freezers within 30 minutes of separation. Samples were never thawed before being shipped overnight on dry ice to Metabolon, Inc. (Durham, NC), where they were again stored in -80°C freezers and thawed once before testing.

Metabolomic profiling and quality control

An untargeted plasma metabolomics analysis was performed by Metabolon, Inc. using Ultrahigh Performance Liquid Chromatography-Tandom Mass Spectrometry (UPLC-MS/MS). Quantification was performed as previously described (Evans et al., 2014); details are outlined in Appendix A1. Metabolites within nine super pathways were identified: amino acids, carbohydrates, cofactors and vitamins, energy, lipids, nucleotides, partially characterized molecules, peptides, and xenobiotics.

Up to three longitudinal plasma samples were available for each participant. Metabolites with an interquartile range of zero (*i.e.*, those with very low or no variability) were excluded from analyses (n=178 metabolites). After removing these metabolites, samples were missing a median of 11.7% metabolites, while metabolites were missing in a median of 1.2% of samples. Missing metabolite values were imputed to the lowest level of detection for each metabolite. Metabolite values were median-scaled and log-transformed to normalize metabolite distributions (van den Berg et al., 2006). If a participant reported that they did not fast or withhold medications and caffeine for at least eight hours, the sample was excluded from analyses (n=159 samples). A total of 1,097 metabolites among 2,344 samples remained for analyses.

DNA collection and genomics quality control

DNA was extracted from whole blood samples using the PUREGENE® DNA Isolation Kit (Gentra Systems, Inc., Minneapolis, MN). DNA concentrations were quantified using the Invitrogen™ Quant-iT™ PicoGreen™ dsDNA Assay Kit (Thermo Fisher Scientific, Inc., Hampton, NH) analyzed on the Synergy 2 Multi-Detection Microplate Reader (Biotek Instruments, Inc., Winooski, VT). Samples were diluted to 50 ng/ul following quantification.

A total of 1,340 samples were genotyped using the Illumina Multi-Ethnic Genotyping Array at the University of Wisconsin Biotechnology Center (Appendix A2). Thirty-six blinded duplicate samples were used to calculate a concordance rate of 99.99%, and discordant genotypes were set to missing. Sixteen samples missing >5% of variants were excluded, while 35,105 variants missing in >5% of individuals were excluded. No samples were removed due to outlying heterozygosity. Six samples were excluded due to inconsistencies between self-reported and genetic sex.

Due to sibling relationships in the WRAP cohort, genetic ancestry was assessed using Principal Components Analysis in Related Samples (PC-AiR), a method that makes robust inferences about population structure in the presence of relatedness (Conomos et al., 2015). This approach included several iterative steps and was based on 63,503 linkage disequilibrium (LD) pruned (r^2<0.10) and common (MAF>0.05) variants, using the 1000 Genomes data as reference populations (Genomes Project et al., 2015). First, kinship coefficients (KCs) were calculated between all pairs of individuals using genomic data with the Kinship-based Inference for Gwas (KING)-robust method (Manichaikul et al., 2010). PC-AiR was used to perform principal components analysis (PCA) on the reference populations along with a subset of

unrelated individuals identified by the KCs. Resulting principal components (PCs) were used to project PC values onto the remaining related individuals. All PCs were then used to recalculate the KCs taking ancestry into account using the PC-Relate method, which estimates KCs robust to population structure (Conomos et al., 2016). PCA was performed again using the updated KCs, and KCs were also estimated again using updated PCs. The resulting PCs identified 1,198 WRAP participants whose genetic ancestry was primarily of European descent. This procedure was repeated within this subset of participants (excluding 1000 Genomes individuals) to obtain PC estimates used to adjust for population stratification in subsequent genomic analyses. Among European descendants, 160 variants were not in Hardy-Weinberg equilibrium (HWE) and 327,064 were monomorphic and thus, removed.

A total of 1,294,660 bi-allelic autosomal variants among 1,198 European descendants remained for imputation, which was performed with the Michigan Imputation Server v1.0.3 (Das et al., 2016), using the Haplotype Reference Consortium (HRC) v. r1.1 2016 (McCarthy et al., 2016) as the reference panel and Eagle2 v2.3 (Loh et al., 2016) for phasing. Prior to imputation, the HRC Imputation Checking Tool (Rayner et al., 2016) was used to identify variants that did not match those in HRC, were palindromic, differed in MAF>0.20, or that had non-matching alleles when compared to the same variant in HRC, leaving 898,220 for imputation. A total of 39,131,578 variants were imputed. Variants with a quality score R^2<0.80, MAF<0.001, or that were out of HWE were excluded, leaving 10,400,394 imputed variants. These were combined with the genotyped variants, leading to 10,499,994 imputed and genotyped variants for analyses. Data cleaning and file preparation were completed

using PLINK v1.9 (Chang et al., 2015) and VCFtools v0.1.14 (Danecek et al., 2011). Coordinates are based on GRCh37 assembly hg19.

Metabolite correlations

Correlations between metabolites were assessed using Pearson r and the first available sample for each individual (*i.e.*, using a cross-sectional approach).

Metabolome-wide association study

Associations were tested using linear mixed effects regression models implemented in the SAS MIXED procedure (Singer, 1998). Primary predictors included age and sex, which were assessed within the same models. Reported effects are from the main effects (β_1) of age and sex, which are calculated by combining the inverse variance weights of the between individual (β_B) and within individual (β_C) effects:

$$\beta_1 = \frac{weight_B \beta_B + \sum_{i=1}^{n} weight_{C_i} \beta_{C_i}}{weight_B + \sum_{i=1}^{n} weight_{C_i}}$$

To examine effect modification of the metabolomics trajectories by sex, analyses were repeated stratifying the sample by sex. To assess the statistical significance of the effect modification, separate models were run that included an interaction term for age-by-sex using the full sample (men and women combined). All models included random intercepts for within-subject correlations (due to repeated measures) and within-family correlations (due to siblings). Models included fixed effects for age, sex, self-reported race, and cholesterol lowering medication use, which was the most commonly used class of medications in our sample. Sensitivity analyses were performed with an additional fixed effect for plasma sample storage time. Each set of analyses was corrected for multiple testing using the Benjamini-Hochberg (Benjamini & Hochberg, 1995) adjustment with an alpha of 0.05.

Metabolite heritability estimates

The h^2 of each metabolite was estimated using a variance components method that jointly models narrow-sense h^2 and the h^2 explained by genotyped variants (Zaitlen et al., 2013), which allows for the inclusion of both closely and distantly related individuals, as implemented in GCTA (Yang, Lee, Goddard, & Visscher, 2011). A genetic relationship matrix was created from 272,839 weakly linked (R^2<0.50) and common (MAF>0.05) directly genotyped variants. Analyses of h^2 were cross-sectional, using the first available metabolomics sample for 1,111 Caucasians that had both metabolomic and genomic data, and adjusted for sex and age. To assess whether metabolite h^2 could influence the effect of age or sex on metabolite levels, Pearson r was used to calculate correlations between h^2 estimates and the strength of associations (*i.e.*, *P*-values) for age and sex.

Results

Participants

A total of 1,212 WRAP participants with 2,344 longitudinal plasma samples were available for analyses. At the baseline visit for the current study, participants were 61 years old on average, 69% were female, and 94% were Caucasian (Table 1). Most individuals were unrelated (n=825), but 147 families had >1 individual (family sizes ranged from 1-9 members, with an average of 1.2 individuals per family). Analyses stratified by sex included 838 women and 374 men, who had similar characteristics with the exception of more men taking cholesterol lowering medications than women. Participants each had 1,097 plasma metabolites available for analyses, 347 (31.6%) of

which were of unknown chemical structure. Metabolites were largely uncorrelated with each other (Appendix A3). Properties of each metabolite, such as biochemical name, super pathway, and sub pathway, are described in Appendix A4.

Metabolome-wide association study

Aging Metabolomics

All metabolome-wide association results are summarized in Table 2 and detailed in Appendix A4. After adjusting for multiple testing, the levels of 637 metabolites (58.1% of metabolites assessed) significantly changed with age, 516 of which increased with age (Figures 1A and 2). Of the total 34 steroid lipids tested, 28 significantly decreased with age (including 20/22 androgenic, 5/5 progestin, 4/4 pregnenolone, and 1/3 corticosteroids), while two, 11-ketoetiocholanolone glucuronide, an androgenic steroid, and cortisol, significantly increased with age.

Higher levels of most fatty acid lipids were associated with increased age (including 13/14 long chain fatty acids, 28/34 acylcarnitines, and 41/78 other fatty acids), with the exception of eicosanodioate (C20-DC), a dicarboxylate fatty acid that decreased with age. Higher levels of sphingolipids tended to be associated with increased age (19/21 associated sphingolipids).

The majority of amino acids associated with age increased with age (82.6% or 90/109 associated amino acids), including glutamine, one of the 20 common amino acids that are encoded directly by the genetic code. Seven other common amino acids decreased with age (histidine, threonine, tryptophan, leucine, methionine, aspartate, and asparagine), while the 12 others were not associated with age.

Sex Metabolomics

Six hundred and ninety-eight metabolites (63.6% of metabolites assessed) significantly differed by sex, with the slight majority (388 metabolites or 55.6%) found in lower levels in women (Figures 1B and 3). Of the metabolites associated with sex, 415 were also associated with age. Twenty-nine steroid lipids were associated with sex, all of which were found in significantly lower levels in women, with the exception of two corticosteroids (cortisol and corticosterone), which were found in higher levels in women. Androgenic steroids constituted the three metabolites most strongly associated with sex (5alpha-androstan-3alpha, 17beta-diol monosulfate, P<4.0e-308, 5alpha-androstan-3alpha, 17beta-diol 17-glucuronide, P=4.3e-226, and 5alpha-androstan-3alpha, 17beta-diol disulfate, P=2.6e-185).

Ninety fatty acids were associated with sex, 60 of which were found in higher levels in women. Acylcarnitine fatty acids were an exception, as 17/26 significantly associated acylcarnitines were found in lower levels in women. Among all tested phospholipids, 73.8% (48/65) were higher in women, as were 85% (34/40) of all tested sphingolipids.

The majority of amino acids associated with sex were found in lower levels in women (76.8% or 86/112), including 13 of the 20 common amino acids (alanine, tyrosine, methionine, arginine, proline, aspartate, asparagine, tryptophan, glutamate, phenylalanine, and the three branched-chain amino acids (BCAAs): leucine, isoleucine, and valine), while two were found in higher levels in women (glycine and serine). The remaining five did not significantly differ by sex.

Effect Modification of Metabolomics Trajectories by Sex

Analyses stratified by sex identified 588 metabolites (53.6% of metabolites assessed) that were significantly associated with age among women (Figures 4A and Appendix A5) and 297 metabolites (27.1% of metabolites assessed) among men (Figures 4B and Appendix A6), with 206 being common to both groups.

The trajectories of 80 metabolites (7.3%) significantly differed over time by sex (Figures 4C and 5). Of the four most significant metabolites, three were sphingolipids, which were also the largest group of metabolites whose trajectories differed by sex (20.0% or 16/80). Fifteen of these sphingolipids increased with age among women and decreased with age among men. Several other groups of metabolites had trajectories that also differed by sex, including seven fatty acids, six of which showed larger increases with age among women than men; nine steroid lipids, eight of which showed larger decreases with age among women than men; eight phospholipids, five of which increased in women and decreased in men with age; and cholesterol, which increased in women and decreased in men with age.

Metabolite heritability estimates

Metabolite h^2 estimates ranged widely (0.2–99.1%) and had a median h^2 of 36.3%, with a first quartile (Q_1) of 25.5% and a third quartile (Q_3) of 49.7% (Figure 6 and Appendix A4). The metabolites with the lowest h^2 were three lipids: adipoylcarnitine (C6-DC), an acylcarnitine (h^2 = 0.2%), 15-methylpalmitate (i17:0), a branched fatty acid (h^2=0.2%), and glycosyl-N-stearoyl-sphingosine (d18:1/18:0), a ceramide lipid (h^2=0.6%). The metabolites with the highest h^2 estimates were two unknown metabolites (X-12093 and X-24328, h^2=99.1% and 91.1%, respectively) and a nucleotide involved in purine metabolism (N2,N2-dimethylguanosine, h^2=90.0%).

Metabolon recently identified X-12093 as N2-acetyl, N6 methyllysine, an amino acid in the lysine catabolic sub pathway.

Super pathway median h^2 estimates ranged from 23.2–46.3%, with peptides having the highest median, followed by amino acids (40.4%), and partially characterized molecules having the lowest median, although the latter pathway only contained five metabolites. Among the metabolite subgroups that were recurrent themes in our association results (*i.e.*, sub pathways highlighted in Table 2), the 20 common amino acids had a median h^2 of 49.3% (Q_1–Q_3: 36.9–65.1%); fatty acids overall had a median h^2 of 30.3% (Q_1–Q_3: 16.9–42.4%), while acylcarnitines had a slightly higher median h^2 of 41.3% (Q_1–Q_3: 26.6–56.0%); phospholipids overall had a median h^2 of 35.9 (Q_1–Q_3: 24.7–53.3%), while phosphatidylcholines had a slightly lower median h^2 of 30.6% (Q_1–Q_3: 22.6–39.6%); sphingolipids had a median h^2 of 42.0% (Q_1–Q_3: 31.8–51.7%); and steroid lipids overall had a median h^2 of 39.6% (Q_1–Q_3: 35.0–50.7%), while androgenic steroids had a median h^2 of 42.5% (Q_1–Q_3: 37.6–50.7%).

Metabolites associated with age and sex had h^2 estimates that were representative of overall metabolite h^2 estimates. Among the 608 metabolites associated with age, the median h^2=36.1% and Q_1–Q_3: 25.9–50.0%. Similarly, among the 680 metabolites associated with sex, the median h^2=37.1% and Q_1-Q_3: 26.0–50.6%. Overall, metabolite h^2 estimates were not correlated with the strength of associations for age or sex (Pearson r=-0.03 and -0.02, respectively).

Discussion

To our knowledge, this is the first longitudinal metabolomics assessment of aging and sex and uses one of the largest panels of metabolites reported to date. Our results provide strong evidence that most plasma metabolite levels are highly influenced by aging and that aging has a broader effect on metabolites in women than men. Metabolites are also highly influenced by sex, with men and women having substantially different metabolomic profiles. We report h^2 estimates on more metabolites than previously reported and find that the variation of only a few metabolite levels can be attributed almost entirely to either genetic or environmental influences. Rather, most are influenced by a complex combination of genetic and environmental factors, consistent with previous studies (Long et al., 2017; Shin et al., 2014). How heritable a metabolite was did not appear to influence the effect of age or sex on metabolite levels.

Differences in levels of plasma lipid steroids, including androgens, progestins, and pregnenolones, were among the most significant findings for both age and sex. The steroid sex differences serve as a proof of concept, as it is well established that androgens are present in lower levels in women than men (Goodman-Gruen & Barrett-Connor, 2000). Androgens are also known to decrease with age among men in both plasma (Ferrini & Barrett-Connor, 1998) and serum (Harman et al., 2001), and also decline with age in serum among women (Spencer, Klein, Kumar, & Azziz, 2007), perhaps most steeply during early reproductive years (Davison, Bell, Donath, Montalto, & Davis, 2005).

The plasma metabolites we identified to be associated with sex and age are consistent with findings from previous cross-sectional studies. The UK Adult Twin

Registry (TwinsUK) study reported 165 out of 280 (58.9%) tested serum and plasma metabolites to be associated with age in cross-sectional analyses (Menni et al., 2013). Our data had 114 of these 165 metabolites, of which 71 were significantly associated with age, and 65 had effects that were in the same direction as those reported in the TwinsUK study (Appendix A7). The metabolites that had the opposite direction of effect between studies were four amino acids (dimethylarginine, leucine, asparagine, and tryptophan), one nucleotide (uridine), and one xenobiotic (theophylline), all of which we reported decreased with age, with the exception of dimethylarginine, which increased with age, contradictory to findings from the TwinsUK study. However, other studies have reported that serum tryptophan levels decrease with age (Dunn et al., 2015; Z. Yu et al., 2012). Among the 65 metabolites with the same effect, 29 were lipids, all of which increased with age (the majority were fatty acids, including 10 long chain fatty acids, six polyunsaturated fatty acids, and six other fatty acids), and 13 were amino acids (including glutamine, which increased with age, and histidine and aspartate, which both decreased with age).

The Cooperative Health Research in the Region of Augsburg (KORA F4) study, which was also cross-sectional, reported 180 out of 507 (35.5%) tested serum metabolites to be associated with sex (Krumsiek et al., 2015). Our data had 98 of these 180 metabolites, of which 84 were significantly associated with sex, and all had effects that were in the same direction as those reported in KORA F4 (Appendix A8). Among these were 33 amino acids (including 11 common amino acids, all of which were lower in women except glycine and serine, which is also consistent with Mittelstrass et al. (Mittelstrass et al., 2011)); 18 lipids (including five long chain fatty acids and three

medium chain acids, all of which were higher in women, and three androgenic steroids, all of which were lower in women); and 18 unknown metabolites (all but one were lower in women). The single most significant finding in the KORA F4 study was the third most significant in our study (5alpha-androsta-3beta, 17beta-diol disulfate, an androgenic steroid; the two other androgenic steroids that were our first and second most significant sex findings were not tested in the KORA F4 study). Also consistent with our findings, other studies have reported serum and plasma phosphatidylcholines and sphingolipids levels to be higher in women than men (Gonzalez-Covarrubias et al., 2013; Mittelstrass et al., 2011; Rist et al., 2017), and serum acylcarnitines to be lower in women (Mittelstrass et al., 2011).

Consistent with results from our sex-stratified analyses, a previous KORA F4 publication also reported serum sphingolipids to increase in concentrations with age among women and acylcarnitines to increase with age among both women and men (Z. Yu et al., 2012). The KORA F4 study, which had a sample of 1,038 women and 1,124 men, also similarly found twice as many metabolites associated with age among women than men. This suggests that our similar observation may not be driven solely by the differences in sample sizes between women and men in our study and that it may have biological implications; *i.e.*, aging may influence a wider breadth of metabolites in women than men. A probable cause for such a difference may be that during menopause, women experience very abrupt and dramatic hormone changes and loss of ovarian function, whereas during “andropause”, men experience a gradual loss of hormones and decline in fertility (Vermeulen, 2000). These hormonal changes could be associated with other metabolic changes as well. Post-menopausal women have higher

levels of sphingomyelins, fatty acids, acylcarnitines, lysophosphatidylcholines, and several amino acids than pre-menopausal women (Auro et al., 2014; Ke et al., 2015), and a recent study found that plasma and urine metabolomics can be used to predict menopause status with 90% accuracy. Moreover, androgenic steroids have been linked to lipid levels in postmenopausal women (Noyan, Yucel, & Sagsoz, 2004). Given that the baseline average ages of women and men in our sample are each ~61 years old, it is likely that our results are indicative of hormonal changes that occur in later ages and that most of our female participants have undergone menopause. It will be crucial to replicate these findings with a metabolomics panel that captures a larger proportion of the ~25,000 known blood metabolites in order to determine the validity of this hypothesis.

Among the 80 metabolites with different trajectories between women and men were sphingolipids, phosphatidylcholines, and cholesterol. Metabolites from the latter two subgroups have been previously reported to have similar trajectories as what we identified, *i.e.*, increasing with age in women and decreasing in men (Auro et al., 2014). To our knowledge, a decrease of sphingolipid levels in men as our results suggest has not been previously reported. However, greater sphingomyelin increases with age in women than men have been previously described (Mielke et al., 2015).

We compared our metabolite h^2 estimates to those recently estimated from a twin study of 1,930 individuals in the TwinsUK cohort (Long et al., 2017). Among the 466 metabolites overlapping with our study, h^2 estimates were only moderately correlated (Pearson r=0.36) and our estimates were 9.6 percentage points lower on average. However, our metabolite h^2 estimates were 8.9 percentage points higher on average

(and had a lower correlation of r=0.25) when comparing 191 overlapping metabolite h^2 estimates from an earlier twin study based on 7,824 individuals from both the KORA F4 and TwinsUK cohorts (Shin et al., 2014). Interestingly, despite having some overlapping participants, h^2 estimates between these two previous studies were only moderately correlated: among 163 overlapping metabolite h^2 estimates, Pearson r was 0.38, with estimates based on the TwinsUK cohort being 18.8 percentage points higher on average than the combined KORA and TwinsUK study. Differences in h^2 estimates may be driven by differences in population composition and size, phenotypic variation, and analytic approaches.

This study was not without limitations. Our findings are likely driven by our panel of metabolites, and it is possible that a different panel of metabolites could produce different results. Many of our findings are in accordance with previous publications, thereby strengthening confidence in our results that have not been previously investigated with regards to age and sex. Accordingly, it will be crucial to replicate novel findings with an external cohort. However, we also identified several inconsistencies between our study and others regarding h^2 estimates and a few of our association results, which could have been due to differences in study designs and sample populations. This challenge is common (Enche Ady et al., 2017), as the field of metabolomics is rapidly developing and widely accepted standards for quality control techniques are forthcoming. Differences in platforms, quantification techniques, statistical analysis methods, laboratory techniques for sample handling (*i.e.*, anti-coagulation method, preservation, storage duration), and fasting status at the time of the sample draw may result in large variations from one study to another (Gonzalez-

Dominguez, Sayago, & Fernandez-Recamales, 2017). The metabolomics quality control process we have outlined here as well as that described in Voyle et al. (Voyle et al., 2016) could serve as guidelines for future studies. Many of our findings included metabolites that had unknown chemical structures, which is a current limitation of the field of metabolomics, as it can be difficult and costly to accurately identify metabolites. Further, we only investigated linear effects of age, but non-linear age effects may exist and should be investigated in future investigations.

Using a large panel of longitudinal metabolomics data, we conducted a comprehensive investigation of the influence of aging and sex on metabolomics. Our findings suggest that levels of most metabolites are highly influenced by sex and age, and that sex differentially influences levels and trajectories of many metabolites. These findings underscore the importance of incorporating age and sex in the design and analysis of metabolomics investigations. We also report that many metabolite levels are influenced by a complex combination of both genomic and environmental influences. These findings offer a deeper understanding of the aging process and could inform many novel hypotheses regarding the role of metabolites in healthy and accelerated aging.

Tables

Table 1. WRAP Participant Characteristics at Baseline for the Current Study.

Mean (SD) or N (%).

Characteristic	Overall (N=1,212, obs=2,344)	Male (n=374, obs=731)	Female (n=838, obs=1,613)
Age (years)	60.8 (6.7)	61.2 (6.9)	60.7 (6.6)
Caucasian	1,135 (93.7)	351 (93.9)	784 (93.6)
Cholesterol lowering medication	387 (31.9)	146 (39.0)*	241 (28.8)*
Sample storage (days)	1,510.5 (415.7)	1,511.2 (424.3)	1,510.2 (412.0)

obs=number of longitudinal observations

*Significantly differs between men and women (*P*=3.9e-4)

Table 2. Metabolome-wide association results summary.

Number of metabolites associated with each trait by pathways and recurrent sub pathways after correcting for multiple comparisons.

	Age			Sex			Age in Women			Age in Men			Age*Sex (female/male)				
Metabolite Super/Sub Pathways (#Metabolites)	Tot	+β	–β	Tot	+β	–β	Tot	+β	–β	Tot	+β	–β	Tot	+/+	–/–	+/–	–/+
Amino Acids (175)	109	90	19	112	26	86	106	93	13	51	39	12	8	3	0	4	1
Common Amino Acids (20)	8	1	7	15	2	13	7	2	5	4	0	4	0	0	0	0	0
Carbohydrates (23)	17	16	1	13	6	7	17	17	0	7	6	1	1	1	0	0	0
Cofactors and Vitamins (28)	19	16	3	16	8	8	20	17	3	14	9	5	1	0	0	1	0
Energy (8)	6	6	0	6	4	2	6	6	0	2	2	0	1	1	0	0	0
Lipids (353)	198	152	46	252	176	76	191	158	33	86	44	42	50	9	9	28	4
Fatty Acids (126)	83	82	1	90	60	30	83	82	1	23	22	1	7	4	0	2	1
Acylcarnitines (34)	28	28	0	26	9	17	32	32	0	7	7	0	5	4	0	1	0
Phospholipids (65)	22	9	13	52	48	4	12	10	2	16	5	11	8	0	3	5	0
Lysophospholipids (24)	3	1	2	17	15	2	1	1	0	2	0	2	3	0	0	3	0
Phosphatidylcholines (19)	8	4	4	17	16	1	5	4	1	8	3	5	4	0	2	2	0
Phosphatidylethanolamine (9)	2	1	1	7	7	0	2	2	0	2	0	2	1	0	1	0	0
Sphingolipids (40)	21	19	2	34	34	0	26	25	1	8	0	8	16	1	0	15	0
Steroids (34)	30	2	28	29	2	27	29	1	28	19	1	18	9	0	6	0	3

Androgenic (22)	20	1	19	20	0	20	20	1	19	15	0	15	4	0	2	0	2
Progestin (5)	5	0	5	2	0	2	5	0	5	0	0	0	5	0	4	0	1
Pregnenolone (4)	4	0	4	4	0	4	4	0	4	3	0	3	0	0	0	0	0
Corticosteroids (3)	1	1	0	3	2	1	0	0	0	1	1	0	0	0	0	0	0
Nucleotides (35)	20	19	1	24	1	23	18	18	0	11	10	1	1	1	0	0	0
Partially Characterized Molecules (5)	3	3	0	2	0	2	4	4	0	0	0	0	0	0	0	0	0
Peptides (22)	12	11	1	16	3	13	14	13	1	4	3	1	1	1	0	0	0
Xenobiotics (101)	44	36	8	52	19	33	39	32	7	18	13	5	3	1	1	1	0
Unknown (347)	209	167	42	205	67	138	173	146	27	104	78	26	14	7	1	3	3
Total (1,097)	637	516	121	698	310	388	588	504	84	297	204	93	80	24	11	37	8

Shaded rows represent super pathways, which sum to the “Total” row. Sub pathways are indented. In the Sex columns, + means the metabolite was higher in women, whereas – means the metabolite was higher in men. For all other columns, + means the metabolite increased with age, whereas – means it decreased with age. In the Age*Sex columns, +/+ means the metabolite increased with age in both women and men, –/– means it decreased with age in both women and men, +/– means it increased with age in women and decreased with age in men, and –/+ means it decreased with age in women and increased with age in men. Results from the Age and Sex columns were assessed within the same model; results from the Age in Women and Age in Men columns were assessed within separate models stratifying the sample by sex; and results from the Age*Sex column were assessed within a separate model including an age-by-sex interaction term.

Figures

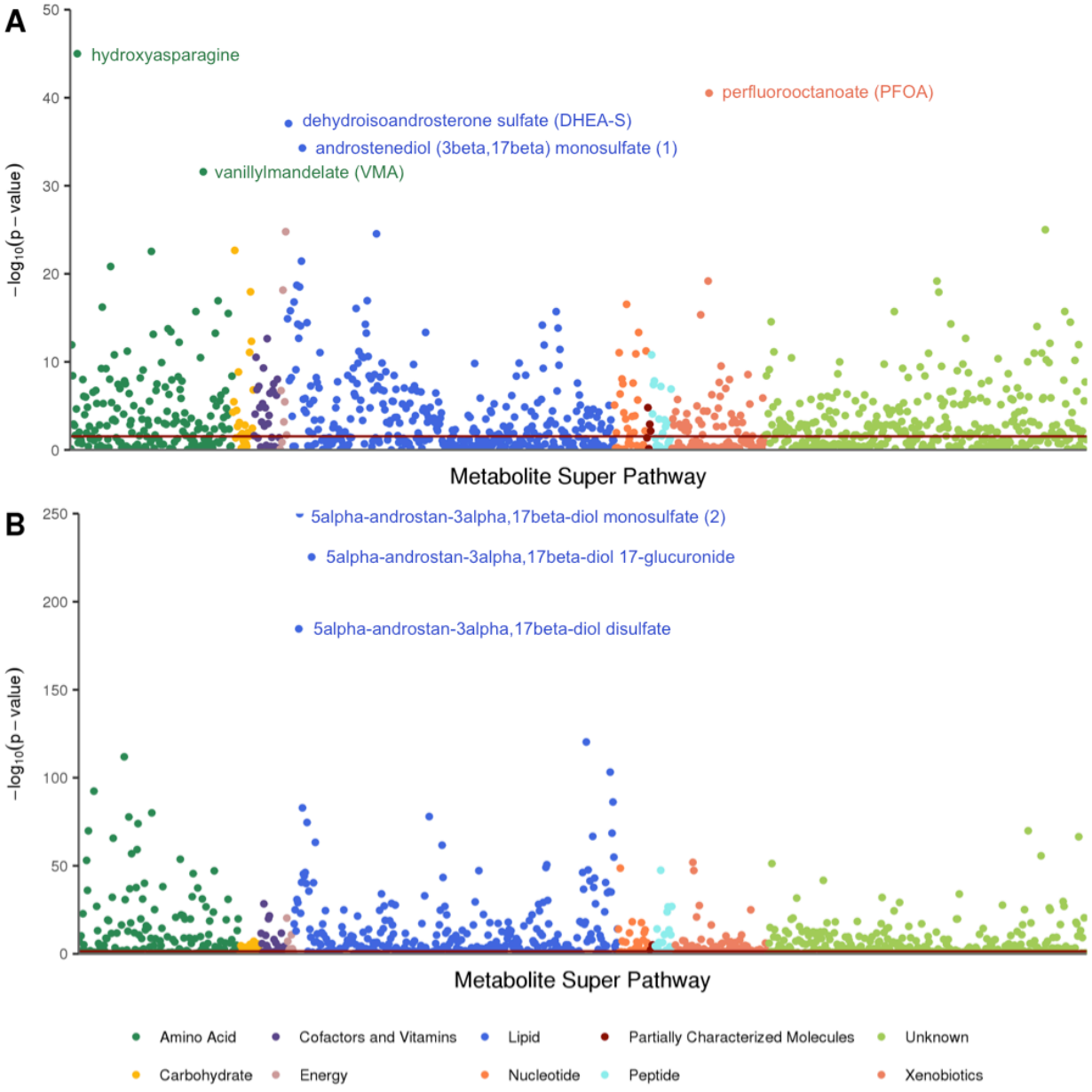


Figure 1. Manhattan plots of metabolome-wide associations results.

A. Age significantly influenced 637 metabolites. B. Sex significantly influenced 698 metabolites. Both sets of results use a Benjamini-Hochberg adjusted p-value threshold (red horizontal line).

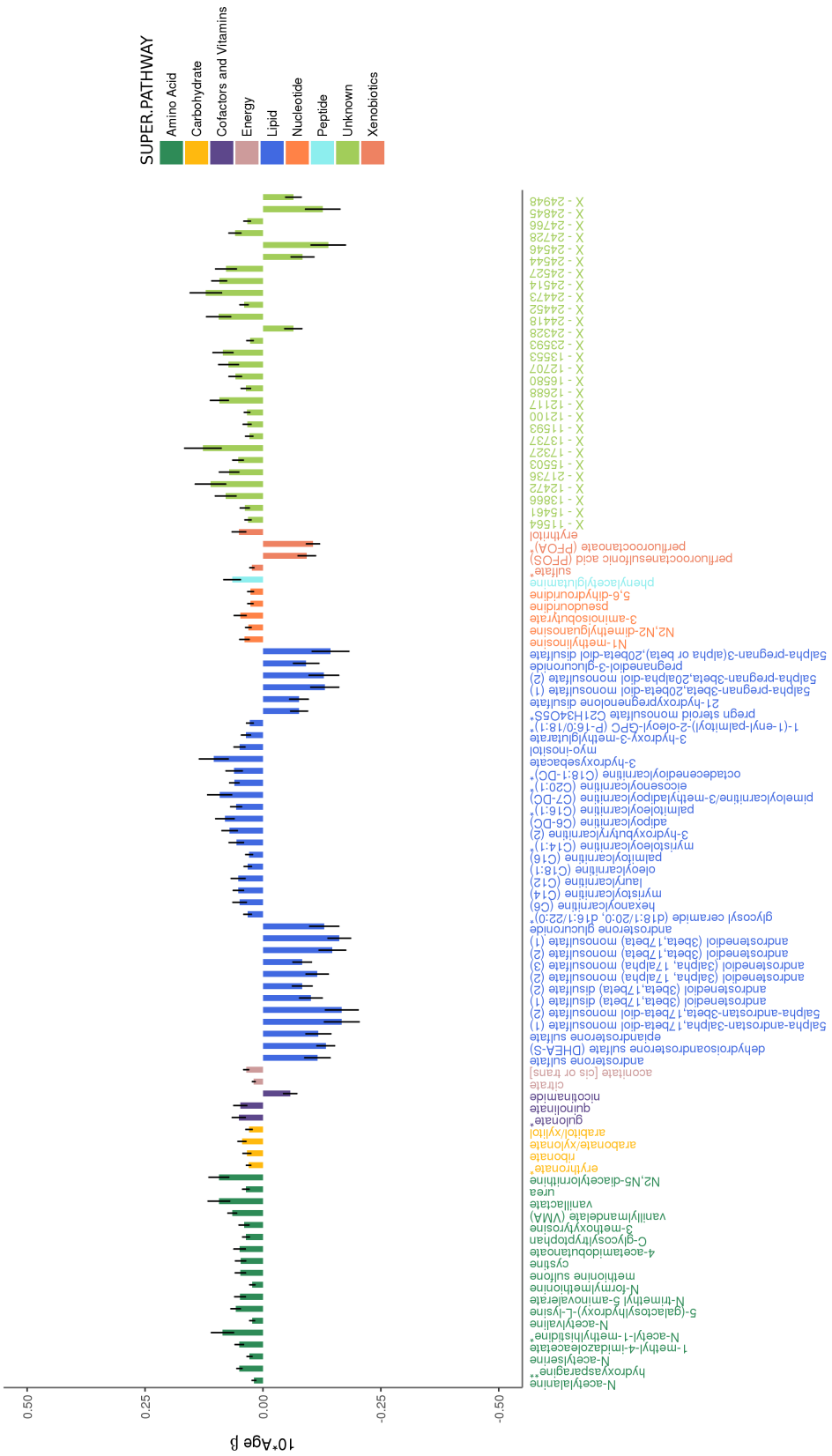

SUPER.PATHWAY
Amino Acid
Carbohydrate
Cofactors and Vitamins
Energy
Lipid
Nucleotide
Peptide
Unknown
Xenobiotics
10*Age β
0.50
0.25
0.00
-0.25
-0.50
N-acetylalanine
hydroxyasparagine**
N-acetylserine
1-methyl-4-imidazoleacetate
N-acetyl-1-methylhistidine*
N-acetylvaline
5-(galactosylhydroxy)-L-lysine
N-trimethyl 5-aminovalerate
N-formylmethionine
methionine sulfone
cystine
4-acetamidobutanoate
C-glycosyltryptophan
3-methoxytyrosine
vanillylmandelate (VMA)
vanillactate
urea
N2,N5-diacetylornithine
erythronate*
ribonate
arabonate/xylonate
arabitol/xylitol
gulonate*
quinolinate
nicotinamide
citrate
aconitate [cis or trans]
androsterone sulfate
epiandrosterone sulfate
dehydroisoandrosterone sulfate (DHEA-S)
5alpha-androstan-3alpha,17beta-diol monosulfate (1)
5alpha-androstan-3beta,17beta-diol monosulfate (2)
androstenediol (3beta,17beta) disulfate (1)
androstenediol (3beta,17beta) disulfate (2)
androstenediol (3alpha, 17alpha) monosulfate (2)
androstenediol (3alpha, 17alpha) monosulfate (3)
androstenediol (3beta,17beta) monosulfate (2)
androstenediol (3beta,17beta) monosulfate (1)
androsterone glucuronide
glycosyl ceramide (d18:1/20:0, d16:1/22:0)*
hexanoylcarnitine (C6)
myristoylcarnitine (C14)
laurylcarnitine (C12)
oleoylcarnitine (C18:1)
palmitoylcarnitine (C16)
myristoleoylcarnitine (C14:1)*
adipoylcarnitine (C6-DC)
3-hydroxybutyrylcarnitine (2)
palmitoleoylcarnitine (C16:1)*
pimeloylcarnitine/3-methyladipoylcarnitine (C7-DC)
eicosenoylcarnitine (C20:1)*
octadecenedioylcarnitine (C18:1-DC)*
3-hydroxysebacate
myo-inositol
3-hydroxy-3-methylglutarate
1-(1-enyl-palmitoyl)-2-oleoyl-GPC (P-16:0/18:1)*
pregn steroid monosulfate C21H34O5S*
21-hydroxypregnenolone disulfate
5alpha-pregnan-3beta,20beta-diol monosulfate (1)
5alpha-pregnan-3beta,20alpha-diol monosulfate (2)
pregnanediol-3-glucuronide
5alpha-pregnan-3(alpha or beta),20beta-diol disulfate
N1-methylinosine
N2,N2-dimethylguanosine
3-aminoisobutyrate
pseudouridine
5,6-dihydrouridine
phenylacetylglutamine
sulfate*
perfluorooctanesulfonic acid (PFOS)
perfluorooctanoate (PFOA)*
erythritol
X - 11564
X - 15461
X - 13866
X - 12472
X - 21736
X - 15503
X - 17327
X - 13737
X - 11593
X - 12100
X - 12117
X - 12688
X - 16580
X - 12707
X - 13553
X - 23593
X - 24328
X - 24418
X - 24452
X - 24473
X - 24514
X - 24527
X - 24544
X - 24546
X - 24728
X - 24766
X - 24845
X - 24948

Figure 2. Adjusted effects of a 10-year increase in age on the top 100 metabolites most strongly influenced by age.

Positive values indicate the amount a metabolite increased over 10 years, whereas negative values indicate the amount a metabolite decreased over 10 years. Black vertical lines indicate 10*standard errors.

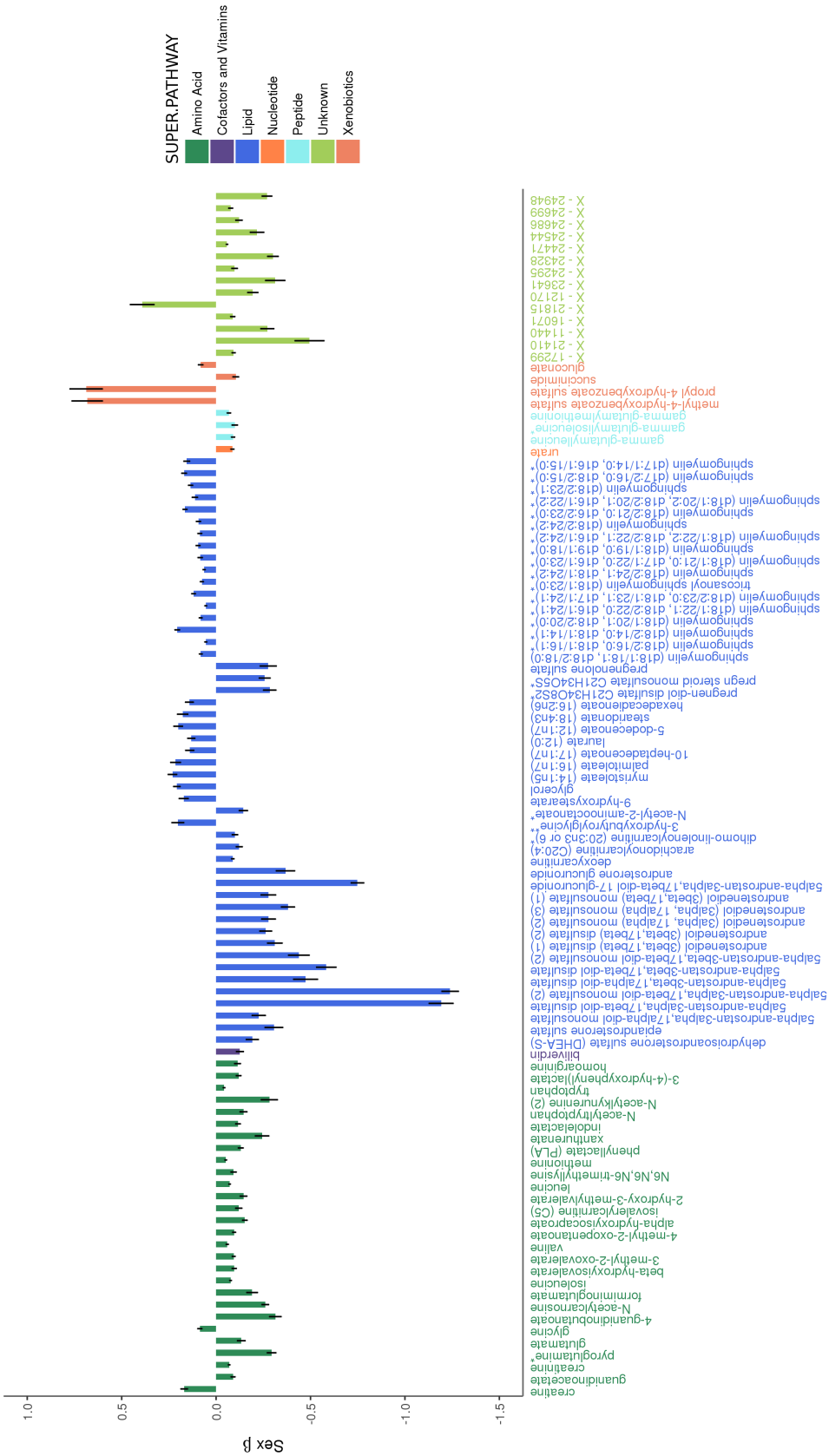

SUPER.PATHWAY
Amino Acid
Cofactors and Vitamins
Lipid
Nucleotide
Peptide
Unknown
Xenobiotics
Sex β
1.0
0.5
0.0
-0.5
-1.0
-1.5
creatine
guanidinoacetate
creatinine
pyroglutamine*
glutamate
glycine
4-guanidinobutanoate
N-acetylcarnosine
formiminoglutamate
isoleucine
beta-hydroxyisovalerate
3-methyl-2-oxovalerate
valine
4-methyl-2-oxopentanoate
alpha-hydroxyisocaproate
isovalerylcarnitine (C5)
2-hydroxy-3-methylvalerate
leucine
N6,N6,N6-trimethyllysine
methionine
phenyllactate (PLA)
xanthurenate
indolelactate
N-acetyltryptophan
N-acetyllysurenine (2)
tryptophan
3-(4-hydroxyphenyl)lactate
homoarginine
bilirubin
dehydroisoandrosterone sulfate (DHEA-S)
epiandrosterone sulfate
5alpha-androstan-3alpha,17alpha-diol monosulfate
5alpha-androstan-3alpha,17beta-diol disulfate
5alpha-androstan-3beta,17beta-diol monosulfate (2)
5alpha-androstan-3beta,17alpha-diol disulfate
5alpha-androstan-3beta,17beta-diol disulfate
5alpha-androstan-3beta,17beta-diol monosulfate (2)
androsterone sulfate (1)
androsterone sulfate (2)
androsterone monosulfate (1)
androsterone monosulfate (2)
androsterone monosulfate (3)
androsterone monosulfate (1)
5alpha-androstan-3alpha,17beta-diol 17-glucuronide
androsterone glucuronide
deoxycarnitine
arachidonoylcarnitine (C20:4)
dihomo-linolenoylcarnitine (20:3n3 or 6)*
3-hydroxybutyroylglycine**
N-acetyl-2-aminooctanoate*
9-hydroxystearate
glycerol
myristoleate (14:1n5)
palmitoleate (16:1n7)
10-heptadecenoate (17:1n7)
laurate (12:0)
5-dodecenoate (12:1n7)
stearidonate (18:4n3)
hexadecadienoate (16:2n6)
pregnen-diol disulfate C21H34O8S2*
pregn steroid monosulfate C21H34O5S*
pregnenolone sulfate
sphingomyelin (d18:1/18:1, d18:2/18:0)
sphingomyelin (d18:2/16:0, d18:1/16:1)*
sphingomyelin (d18:2/14:0, d18:1/14:1)*
sphingomyelin (d18:1/20:1, d18:2/20:0)*
sphingomyelin (d18:1/22:1, d18:2/22:0, d16:1/24:1)*
sphingomyelin (d18:2/23:0, d18:1/23:1, d17:1/24:1)*
tricosanoyl sphingomyelin (d18:1/23:0)*
sphingomyelin (d18:2/24:1, d18:1/24:2)*
sphingomyelin (d18:1/21:0, d17:1/22:0, d16:1/23:0)*
sphingomyelin (d18:1/19:0, d19:1/18:0)*
sphingomyelin (d18:1/22:2, d18:2/22:1, d16:1/24:2)*
sphingomyelin (d18:2/24:2)*
sphingomyelin (d18:2/21:0, d16:2/23:0)*
sphingomyelin (d18:1/20:2, d18:2/20:1, d16:1/22:2)*
sphingomyelin (d18:2/23:1)*
sphingomyelin (d17:2/16:0, d18:2/15:0)*
sphingomyelin (d17:1/14:0, d16:1/15:0)*
urate
gamma-glutamylleucine
gamma-glutamylmethionine
gamma-glutamylisoleucine*
methyl-4-hydroxybenzoate sulfate
propyl 4-hydroxybenzoate sulfate
succinimide
gluconate
X - 17299
X - 21410
X - 11440
X - 16071
X - 21815
X - 12170
X - 23641
X - 24295
X - 24328
X - 24471
X - 24544
X - 24686
X - 24699
X - 24948

Figure 3. Adjusted effects of the top 100 metabolites most strongly influenced by sex. Positive values indicate that the metabolite was higher in women, whereas negative values indicate that the metabolite was higher in men. Black vertical lines indicate 10*standard errors.

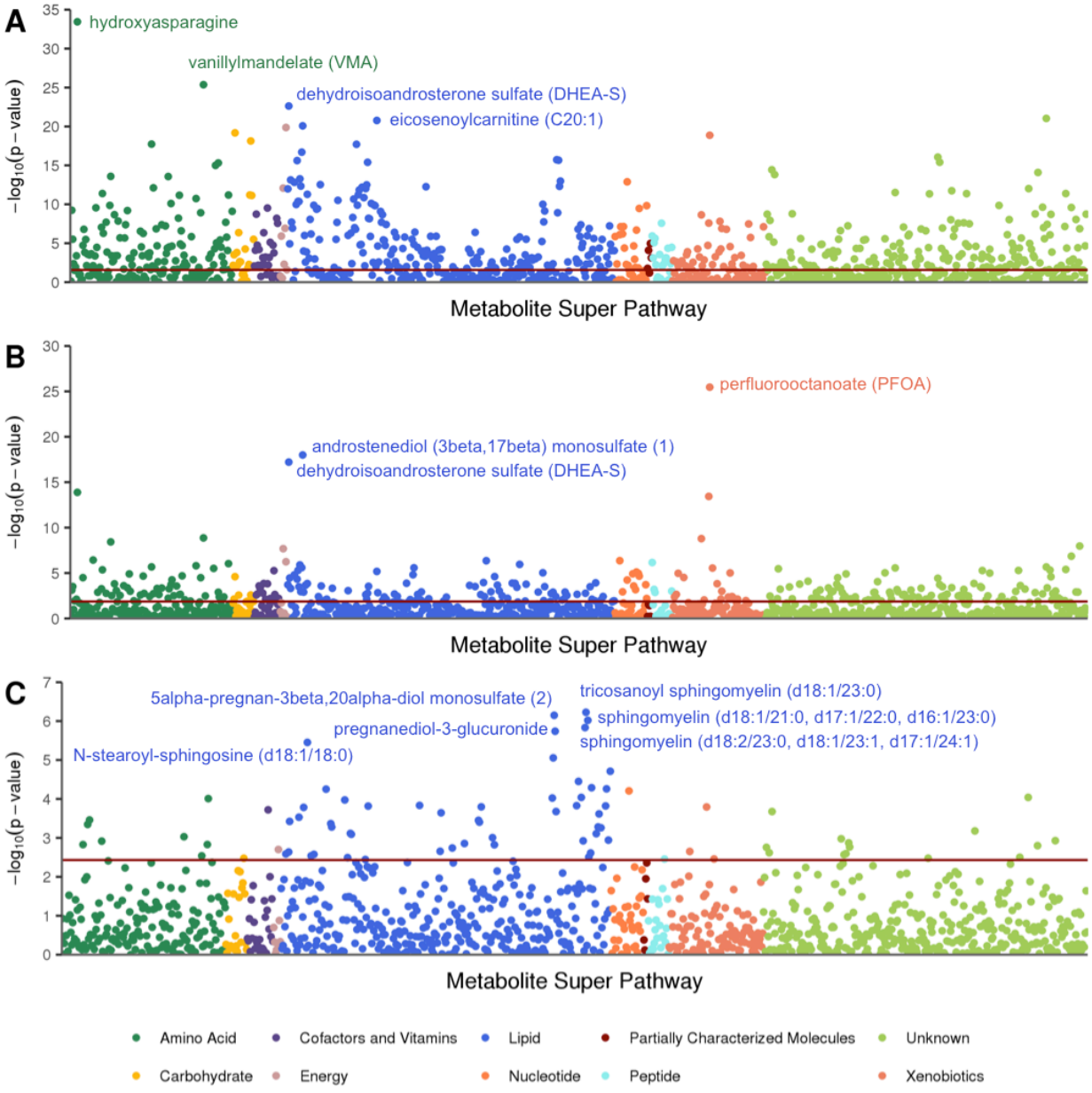


Figure 4. Manhattan plots of metabolome-wide associations results.

A. Age significantly influenced 588 metabolites in women. B. Age significantly influenced 297 metabolites in men. C. Trajectories of 80 metabolites significantly differ by sex. Each set of results uses a Benjamini-Hochberg adjusted p-value threshold (red horizontal line).

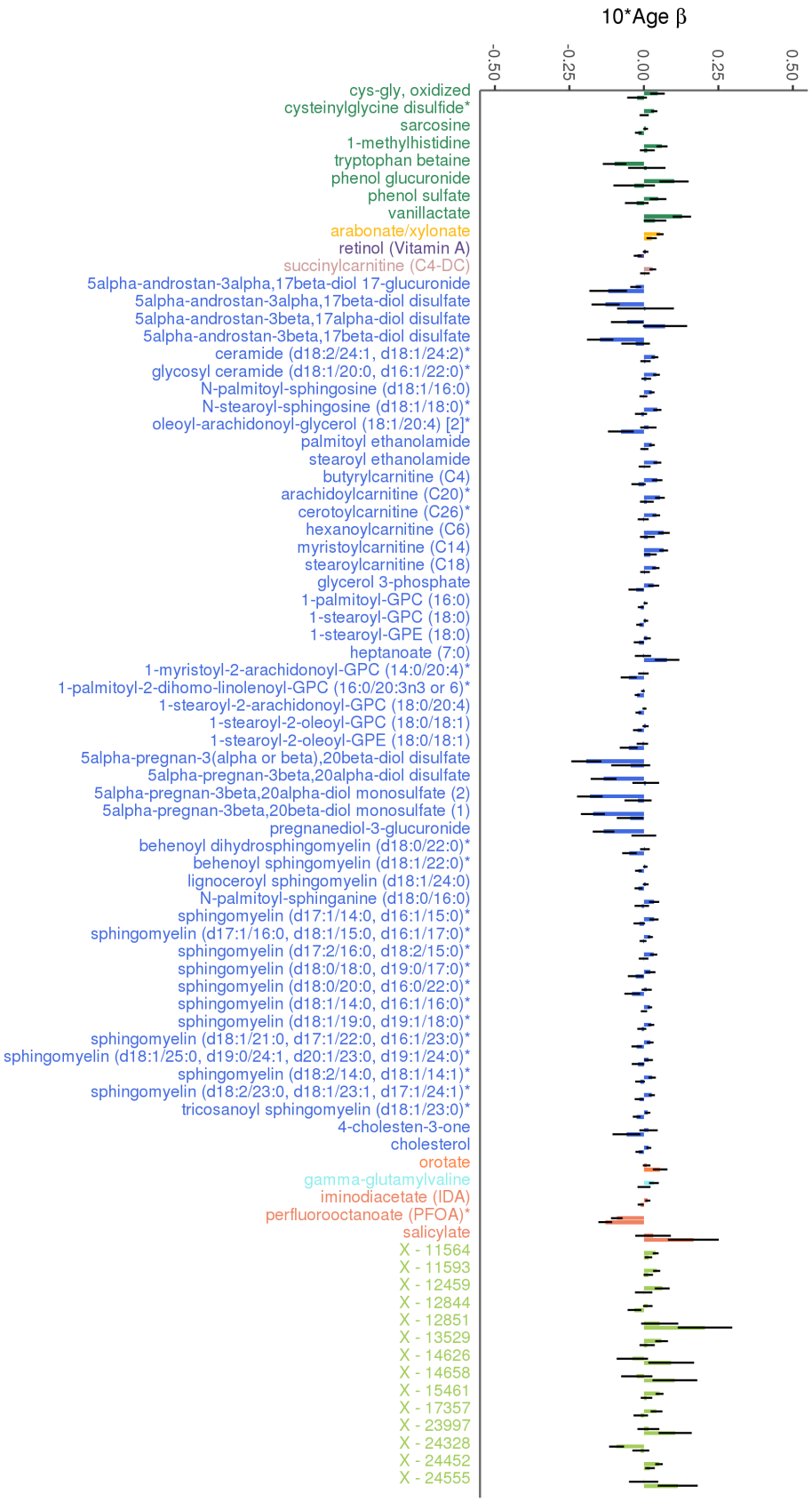

10*Age β
-0.50
-0.25
0.00
0.25
0.50
cys-gly, oxidized
cysteinylglycine disulfide*
sarcosine
1-methylhistidine
tryptophan betaine
phenol glucuronide
phenol sulfate
vanillactate
arabonate/xylonate
retinol (Vitamin A)
succinylcarnitine (C4-DC)
5alpha-androstan-3alpha,17beta-diol 17-glucuronide
5alpha-androstan-3alpha,17beta-diol disulfate
5alpha-androstan-3beta,17alpha-diol disulfate
5alpha-androstan-3beta,17beta-diol disulfate
ceramide (d18:2/24:1, d18:1/24:2)*
glycosyl ceramide (d18:1/20:0, d16:1/22:0)*
N-palmitoyl-sphingosine (d18:1/16:0)
N-stearoyl-sphingosine (d18:1/18:0)*
oleoyl-arachidonoyl-glycerol (18:1/20:4) [2]*
palmitoyl ethanolamide
stearoyl ethanolamide
butyrylcarnitine (C4)
arachidoylcarnitine (C20)*
cerotoylcarnitine (C26)*
hexanoylcarnitine (C6)
myristoylcarnitine (C14)
stearoylcarnitine (C18)
glycerol 3-phosphate
1-palmitoyl-GPC (16:0)
1-stearoyl-GPC (18:0)
1-stearoyl-GPE (18:0)
heptanoate (7:0)
1-myristoyl-2-arachidonoyl-GPC (14:0/20:4)*
1-palmitoyl-2-dihomo-linolenoyl-GPC (16:0/20:3n3 or 6)*
1-stearoyl-2-arachidonoyl-GPC (18:0/20:4)
1-stearoyl-2-oleoyl-GPC (18:0/18:1)
1-stearoyl-2-oleoyl-GPE (18:0/18:1)
5alpha-pregnan-3(alpha or beta),20beta-diol disulfate
5alpha-pregnan-3beta,20alpha-diol disulfate
5alpha-pregnan-3beta,20alpha-diol monosulfate (2)
5alpha-pregnan-3beta,20beta-diol monosulfate (1)
pregnanediol-3-glucuronide
behenoyl dihydrosphingomyelin (d18:0/22:0)*
behenoyl sphingomyelin (d18:1/22:0)*
lignoceroyl sphingomyelin (d18:1/24:0)
N-palmitoyl-sphinganine (d18:0/16:0)
sphingomyelin (d17:1/14:0, d16:1/15:0)*
sphingomyelin (d17:1/16:0, d18:1/15:0, d16:1/17:0)*
sphingomyelin (d17:2/16:0, d18:2/15:0)*
sphingomyelin (d18:0/18:0, d19:0/17:0)*
sphingomyelin (d18:0/20:0, d16:0/22:0)*
sphingomyelin (d18:1/14:0, d16:1/16:0)*
sphingomyelin (d18:1/19:0, d19:1/18:0)*
sphingomyelin (d18:1/21:0, d17:1/22:0, d16:1/23:0)*
sphingomyelin (d18:1/25:0, d19:0/24:1, d20:1/23:0, d19:1/24:0)*
sphingomyelin (d18:2/14:0, d18:1/14:1)*
sphingomyelin (d18:2/23:0, d18:1/23:1, d17:1/24:1)*
tricosanoyl sphingomyelin (d18:1/23:0)*
4-cholesten-3-one
cholesterol
orotate
gamma-glutamylvaline
iminodiacetate (IDA)
perfluorooctanoate (PFOA)*
salicylate
X - 11564
X - 11593
X - 12459
X - 12844
X - 12851
X - 13529
X - 14626
X - 14658
X - 15461
X - 17357
X - 23997
X - 24328
X - 24452
X - 24555

Figure 5. Adjusted effects of a 10-year increase in age on the 80 metabolites with trajectories that significantly differ by sex.

For each metabolite, the bar on the left represents the change in metabolite level over 10 years in women, whereas the bar on the right represents the change in metabolite level over 10 years in men. Black vertical lines indicate 10*standard errors.

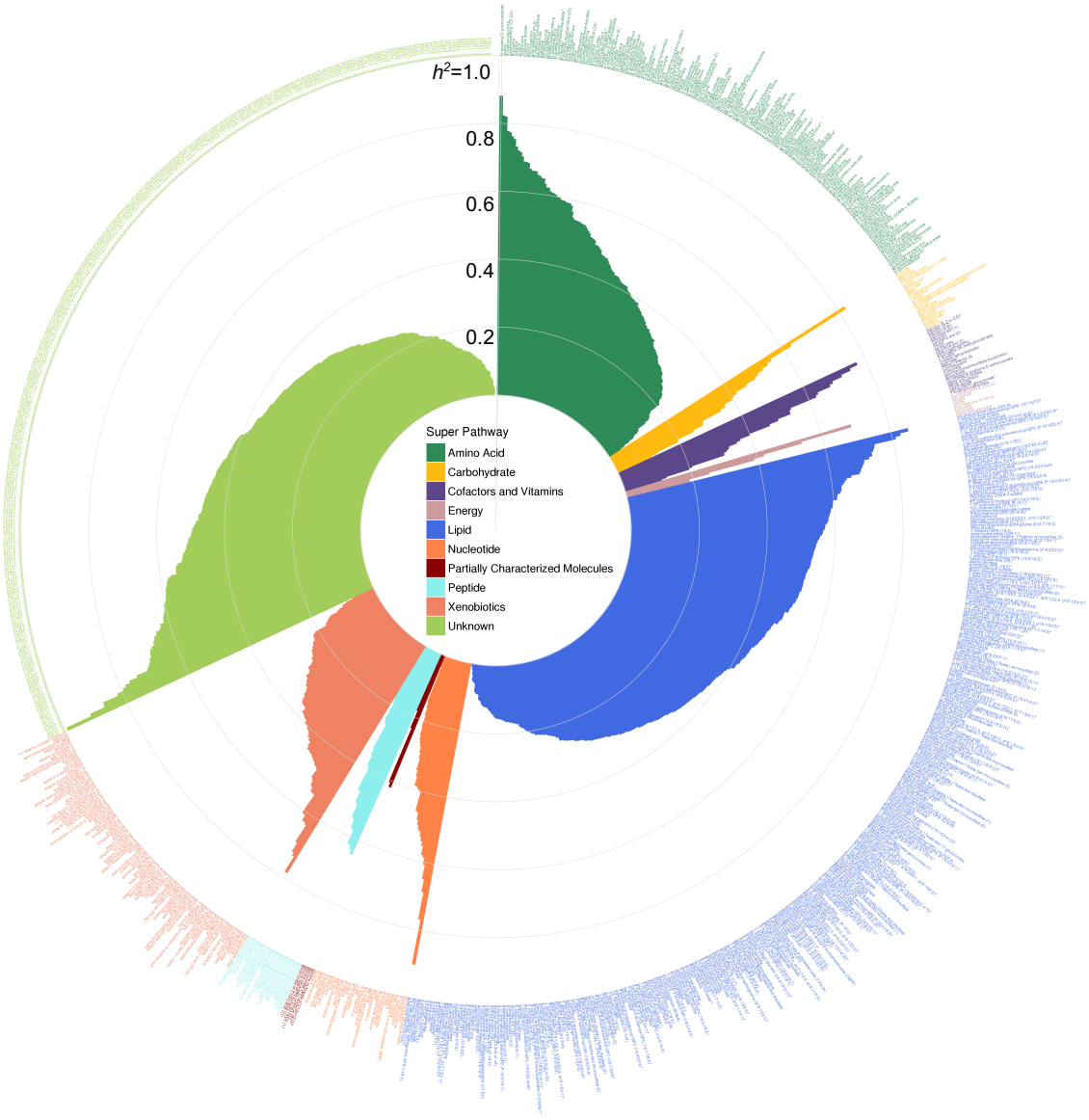


Figure 6. Pinwheel plot of metabolite heritabilities.

Each bar indicates the heritability of the corresponding metabolite. Metabolite names are indicated in the outer circle.

Chapter 2. Metabolites associated with early age-related cognitive change in a cohort at risk for Alzheimer's

Abstract

Objective: We investigated the metabolomics of early cognitive changes related to Alzheimer's disease (AD) in order to better understand mechanisms that could contribute to early stages and progression of this disease. Methods: This investigation used longitudinal plasma samples from the Wisconsin Registry for Alzheimer's Prevention (WRAP), a cohort of participants who were dementia free at enrollment and enriched with a parental history of AD. Metabolomic profiles were quantified for 2,338 fasting plasma samples among 1,206 participants, each with up to three study visits. Results: Of 1,097 metabolites tested, levels of seven were associated with executive function trajectories, including an amino acid and three fatty acids, but none were associated with delayed recall trajectories. Conclusions: Our time-varying metabolomic results suggest potential mechanisms that could contribute to the earliest signs of cognitive decline. In particular, fatty acids may be associated with cognition in a manner that is more complex than previously suspected.

Background

Recent technological advances have made metabolomic studies increasingly favorable among Alzheimer's disease (AD) researchers (Enche Ady et al., 2017; Gonzalez-Dominguez et al., 2017; Trushina & Mielke, 2014); however, these studies have been limited to cross-sectional approaches comparing patients with either AD or mild cognitive impairment (MCI) to controls. In these early stages of AD metabolomics research, few metabolites have been found to be associated with AD in more than one study (Enche Ady et al., 2017). Because neuropathological changes that lead to the development of AD occur decades before its clinical presentation (Jack et al., 2010; Serrano-Pozo, Frosch, Masliah, & Hyman, 2011), longitudinal investigations preceding its diagnosis could add to our current knowledge. In particular, understanding how biomarkers correlate with subtle changes in cognition prior to AD diagnosis could help identify causal mechanisms contributing to its onset.

Executive function and memory deficits occur in the very early stages of AD, prior to deficits of language and visuospatial functions (Albert, 1996; Baudic et al., 2006; Lafleche & Albert, 1995), and are associated with AD pathology and subsequent global cognitive decline (Clark et al., 2016; Clark et al., 2012). Metabolite levels associated with these early changes in cognition could be indicative of underlying biological mechanisms and pathways contributing to the pathology of AD and could ultimately inform stronger predictive models and preventative trials for this disease.

Using longitudinal plasma samples from the Wisconsin Registry for Alzheimer's Prevention (WRAP), we investigated whether time-varying metabolite levels predicted age-related cognitive changes (*i.e.*, trajectories) for executive function and memory,

specifically, delayed recall. Results from each of these association analyses were further explored using Mendelian randomization (MR) and in a metabolite pathway analysis.

Methods

Participants

Study participants were from WRAP, a longitudinal study of initially dementia free middle-aged adults that allows for the enrollment of siblings and is enriched for a parental history of AD. Further details of the study design and methods used have been previously described (Johnson et al., 2018; Sager et al., 2005). Analyses did not include the baseline WRAP visit due to subsequent protocol changes regarding sample collection procedures and tests included in the neuropsychological battery. Study visits included in the current analyses occurred every two years. This study was conducted with the approval of the University of Wisconsin Institutional Review Board, and all participants provided signed informed consent before participation.

Biological samples

Plasma collection and sample handling

Fasting blood samples for this study were drawn the morning of each study visit, which was also the day cognitive testing was completed. Blood was collected in 10 mL ethylenediaminetetraacetic acid (EDTA) vacutainer tubes. They were immediately placed on ice, and then centrifuged at 3000 revolutions per minute for 15 minutes at room temperature. Plasma was pipetted off within one hour of collection. Plasma samples were aliquoted into 1.0 mL polypropylene cryovials and placed in -80°C

freezers within 30 minutes of separation. Samples were never thawed before being shipped overnight on dry ice to Metabolon, Inc. (Durham, NC), where they were again stored in -80°C freezers and thawed once before testing.

Metabolomic profiling and quality control

An untargeted plasma metabolomics analysis was performed by Metabolon, Inc. using Ultrahigh Performance Liquid Chromatography-Tandom Mass Spectrometry (UPLC-MS/MS). Quantification was performed as previously described (Evans et al., 2014); details are outlined in Appendix A1. Metabolites within nine super pathways were identified: amino acids, carbohydrates, cofactors and vitamins, energy, lipids, nucleotides, partially characterized molecules, peptides, and xenobiotics.

Up to three longitudinal plasma samples were available for each participant. Metabolites with an interquartile range of zero (*i.e.*, those with very low or no variability) were excluded from analyses (n=178 metabolites). After removing these metabolites, samples were missing a median of 11.7% metabolites, while metabolites were missing in a median of 1.2% of samples. Missing metabolite values were imputed to the lowest level of detection for each metabolite. Metabolite values were median-scaled and log-transformed to normalize metabolite distributions (van den Berg et al., 2006). If a participant reported that they did not fast or withhold medications and caffeine for at least eight hours, the sample was excluded from analyses (n=159 samples). A total of 1,097 metabolites among 2,338 samples remained for analyses.

DNA collection and genomics quality control

DNA was extracted from whole blood samples using the PUREGENE® DNA Isolation Kit (Gentra Systems, Inc., Minneapolis, MN). DNA concentrations were

quantified using the Invitrogen™ Quant-iT™ PicoGreen™ dsDNA Assay Kit (Thermo Fisher Scientific, Inc., Hampton, NH) analyzed on the Synergy 2 Multi-Detection Microplate Reader (Biotek Instruments, Inc., Winooski, VT). Samples were diluted to 50 ng/ul following quantification.

A total of 1,340 samples were genotyped using the Illumina Multi-Ethnic Genotyping Array at the University of Wisconsin Biotechnology Center (Appendix A2). Thirty-six blinded duplicate samples were used to calculate a concordance rate of 99.99%, and discordant genotypes were set to missing. Sixteen samples missing >5% of variants were excluded, while 35,105 variants missing in >5% of individuals were excluded. No samples were removed due to outlying heterozygosity. Six samples were excluded due to inconsistencies between self-reported and genetic sex.

Due to sibling relationships in the WRAP cohort, genetic ancestry was assessed using Principal Components Analysis in Related Samples (PC-AiR), a method that makes robust inferences about population structure in the presence of relatedness (Conomos et al., 2015). This approach included several iterative steps and was based on 63,503 linkage disequilibrium (LD) pruned (r^2<0.10) and common (MAF>0.05) variants, using the 1000 Genomes data as reference populations (Genomes Project et al., 2015). First, kinship coefficients (KCs) were calculated between all pairs of individuals using genomic data with the Kinship-based Inference for Gwas (KING)-robust method (Manichaikul et al., 2010). PC-AiR was used to perform principal components analysis (PCA) on the reference populations along with a subset of unrelated individuals identified by the KCs. Resulting principal components (PCs) were used to project PC values onto the remaining related individuals. All PCs were then

used to recalculate the KCs taking ancestry into account using the PC-Relate method, which estimates KCs robust to population structure (Conomos et al., 2016). PCA was performed again using the updated KCs, and KCs were also estimated again using updated PCs. The resulting PCs identified 1,198 WRAP participants whose genetic ancestry was primarily of European descent. This procedure was repeated within this subset of participants (excluding 1000 Genomes individuals) to obtain PC estimates used to adjust for population stratification in subsequent genomic analyses. Among European descendants, 160 variants were not in Hardy-Weinberg equilibrium (HWE) and 327,064 were monomorphic and thus, removed.

A total of 1,294,660 bi-allelic autosomal variants among 1,198 European descendants remained for imputation, which was performed with the Michigan Imputation Server v1.0.3 (Das et al., 2016), using the Haplotype Reference Consortium (HRC) v. r1.1 2016 (McCarthy et al., 2016) as the reference panel and Eagle2 v2.3 (Loh et al., 2016) for phasing. Prior to imputation, the HRC Imputation Checking Tool (Rayner et al., 2016) was used to identify variants that did not match those in HRC, were palindromic, differed in MAF>0.20, or that had non-matching alleles when compared to the same variant in HRC, leaving 898,220 for imputation. A total of 39,131,578 variants were imputed. Variants with a quality score R^2<0.80, MAF<0.001, or that were out of HWE were excluded, leaving 10,400,394 imputed variants. These were combined with the genotyped variants, leading to 10,499,994 imputed and genotyped variants for analyses. Data cleaning and file preparation were completed using PLINK v1.9 (Chang et al., 2015) and VCFtools v0.1.14 (Danecek et al., 2011). Coordinates are based on GRCh37 assembly hg19.

Cognitive phenotypes

Composite scores were calculated for executive function and delayed recall based on a previous analysis (Clark et al., 2016). Each composite score was calculated from three neuropsychological tests, which were each converted to z-scores using baseline means and standard deviations. The executive function composite score included the Trails Making Test Part B (TMTB) (Reitan & Wolfson, 1985) total time to completion, Stroop Neuropsychological Screen Test (Trenerry, Crosson, DeBoe, & Leber, 1989) color-word interference total items completed in 120 second, and Wechsler Abbreviated Intelligence Scale-Revised (WAIS-R) digit symbol coding subtest total items completed in 90 seconds (Wechsler, 1981). The delayed recall composite score included the Rey Auditory Verbal Learning Test (RAVLT) (Schmidt, 1996) long-delay free recall, Wechsler Memory Scale-Revised Logical Memory (WMS-R LM) (Wechsler, 1987) delayed recall, and Brief Visuospatial Memory Test (BVMT-R) (Benedict, 1997) delayed recall. The TMTB was multiplied by negative one prior to being converted to z-scores, so that higher z-scores indicated better performance. These z-scores were then averaged to derive executive function and delayed recall composite scores at each visit for each individual.

Statistical analyses

Metabolome-wide association studies

All associations were tested using linear mixed effects regression models implemented in the SAS MIXED procedure. To assess whether metabolite levels were associated with age-related cognitive trajectories, an interaction term between metabolite level and age was used to predict cognitive composite scores (*i.e.*, executive

function and delayed recall). Models included fixed effects for centered age, sex, self-reported race, cholesterol-lowering medication use (the most commonly used class of medications in our sample), sample storage time, education, a genetic risk score for *APOE* (Darst et al., 2017), and practice effects (using visit number). Random intercepts were included for within-subject correlations due to repeated measures and within-family correlations due to the enrollment of siblings. The two sets of *P*-values resulting from testing executive function and delayed recall trajectories were separately corrected for multiple testing using the Benjamini-Hochberg (Benjamini & Hochberg, 1995) adjustment with an alpha of 0.05.

Mendelian randomization

MR (Smith & Ebrahim, 2003) was used to assess whether levels of any individual metabolite identified in our association analyses (*i.e.*, metabolites associated with either executive function or delayed recall trajectories) could causally influence cognition. Metabolic quantitative trait loci (mQTL) were identified as genomic variants influencing metabolite levels with a $P<0.001$ using genome-wide association study (GWAS) summary statistics provided by the authors of a recent publication by Long et al., 2017 (Long et al., 2017). A polygenic score (PS) was created for each metabolite identified in our association analyses that also had GWAS summary statistics available. PSs were defined as the sum of an individual's metabolite-increasing alleles weighted by the effect sizes from GWAS summary statistics. PSs were created using the additive allelic scoring function in PLINK 1.9 (Chang et al., 2015) after LD pruning variants within each PS ($R^2>0.50$). To be consistent with our discovery models, interactions between each PS and age were tested for association with cognition using linear mixed effects

regression models. Models included fixed effects for age, sex, education, practice effects, and the first four PCs to account for population stratification, and included random intercepts for repeated measures and sibling relationships.

Metabolite pathway analysis

Results from association analyses were further investigated using a metabolite pathway analysis. Metabolites included in this analysis were those associated with either executive function or delayed recall trajectories with an unadjusted $P<0.05$ and that had a Kyoto Encyclopedia of Genes and Genomics (KEGG) compound ID (Kanehisa & Goto, 2000). Metabolites on our panel with KEGG compound IDs were used as the reference panel for this analysis. The pathway analysis was conducted using MetaboAnalyst 4.0 and included both an overrepresentation analysis, which was assessed using a hypergeometric test, and a pathway topology analysis, which was assessed using relative-betweenness centrality (Xia & Wishart, 2016). The overrepresentation analysis tests whether a user-defined list of metabolites represents a particular pathway of metabolites more than expected by chance. The pathway topology analysis considers the structure of a pathway by assessing how connected metabolites are within a pathway. If a pathway contains metabolites that connect dense clusters of other metabolites, the pathway would have a high impact score, as changes to its metabolites would likely have a strong impact on other metabolites within the pathway.

Results

Participants

A total of 1,206 WRAP participants with 2,338 longitudinal plasma samples were available for analyses. At baseline for the current study, 69.2% of participants were female, 93.7% were Caucasian, and participants were 60.9 years old with a bachelor's degree, on average (Table 1). Participants each had 1,097 plasma metabolites available for analyses, 347 (31.6%) of which were of unknown chemical structure. Properties of each metabolite, such as biochemical name, super pathway, and sub pathway are described in Appendix B1.

Metabolome-wide association studies

Executive Function

All metabolome-wide association results are detailed in Appendix B1. Seven metabolite-by-age interactions were associated with executive function (Figures 1A and 2). Levels of cysteine S-sulfate, an amino acid, had the strongest association (unadjusted *P*=5.2e-05), with lower levels associated with better executive function in midlife and poorer executive function later in life. The six other significant metabolites showed the opposite relationship with age and executive function, which included erucate (22:1n9) (a monosaturated omega-9 fatty acid), four partially characterized molecules (glycine conjugate of $C_{10}H_{12}O_2$, fatty acid 8:1 acyl glutamine conjugate, fatty acid 6:1 acyl glutamine conjugate, and $C_{12}H_{18}O_5$), and one unknown metabolite (X-18887).

Delayed Recall

No metabolite-by-age interactions were associated with delayed recall after adjusting for multiple comparisons. The three strongest interactions included heneicosapentaenoate (21:5n3) (a polysaturated fatty acid, unadjusted *P*=0.00009), X –

02269 (an unknown metabolite, *P*=0.0004), and erucate (22:1n9) (unadjusted *P*=0.0005) (Figure 1B). Four of the seven metabolites associated with executive function showed a similar relationship with delayed recall, although none were statistically significant (erucate (22:1n9), X – 13866, X – 12104, and cysteine S-sulfate, all unadjusted *P*-values <0.20) (Appendix B2).

Mendelian randomization

GWAS summary statistics were available for three of the seven metabolites associated with executive function (cysteine S-sulfate, erucate (22:1n9), and X-13866, an unknown metabolite) and used to create a PS for each metabolite. The three PSs were fairly weak instruments, with correlations with corresponding metabolites ranging from r=-0.04 to 0.004 and the largest *F*-statistic=1.71, well below the commonly used *F*-statistic threshold of 10 (Stock, Wright, & Yogo, 2002) (Appendix B3). Not surprisingly, associations between executive function and the PSs-by-age were not significant (each *P*≥0.54). Thus, MR analyses were insufficient to draw conclusions about the nature of the relationship between the metabolites and executive function.

Metabolite pathway analysis

Of the 1,097 metabolites tested, only 291 had KEGG compound IDs that were recognized by MetaboAnalyst and could be used as the reference panel for the pathway analysis. A total of 254 metabolites met the inclusion threshold of an unadjusted *P*<0.05 for the cognitive metabolite pathway analysis; however, only 82 of these were identified metabolites with KEGG compound IDs. These metabolites most strongly represented pathways involved in inositol phosphate, ether lipid, and amino sugar and nucleotide

metabolism, although none of the pathways identified were statistically significant (Appendix B4 and Appendix B5).

Discussion

We analyzed the metabolomics of cognitive trajectories using time-varying plasma metabolomic samples and a large panel of metabolites. Our findings suggest that specific metabolite levels, particularly cysteine s-sulfate and fatty acid lipids, correspond with executive function trajectories in late middle-aged adults at increased risk for AD. However, metabolite levels were not statistically associated with delayed recall trajectories.

The associations we observed between metabolite levels and executive function trajectories could provide insight to mechanisms contributing to cognitive decline. In particular, lower levels of the amino acid cysteine S-sulfate were associated with better executive function in midlife, but poorer executive function later in life. The involvement of cysteine metabolism in AD has been implicated in a pathway analysis of previous AD metabolomics studies (Enche Ady et al., 2017). Our results further suggest that such a relationship could depend on age. Cysteine S-sulfate is a glutamate receptor agonist that can lead to calcium influx in nerve cells and neurotoxicity when present in high levels (Olney, Misra, & de Gubareff, 1975; Snowden et al., 2017). It has been shown to drive excitotoxic neurodegeneration in individuals with molybdenum cofactor deficiency, an autosomal recessive inborn error of metabolism characterized by early childhood death (Kumar et al., 2017). This supports our finding that high levels of cysteine S-sulfate may be detrimental to cognitive function in midlife. However, further

investigations using longitudinal cohorts, and perhaps experiments using model organisms, will be crucial to validate our findings and determine whether high levels could have protective effects later in life.

The opposite pattern was seen for the six other metabolites associated with executive function, which included three fatty acids, where higher metabolite levels in younger years, but lower levels in older years, were associated with better executive function. One of these fatty acids was erucate (22:1n9), an omega-9 fatty acid that readily crosses the blood brain barrier (Golovko & Murphy, 2006) and has been shown to enhance memory performance in mice (Kim et al., 2016). Fatty acids have long been suspected to influence cognitive performance, but studies have had mixed findings regarding their role, particularly of omega-3 fatty acids (Cederholm, Salem, & Palmblad, 2013; Mazereeuw, Lanctot, Chau, Swardfager, & Herrmann, 2012). Our results suggest that this role may be difficult to define because the implications of these metabolite levels change as individuals age. This is further supported by the consistent relationships seen by two other partially characterized fatty acids (fatty acid 8:1 acyl glutamine conjugate and fatty acid 6:1 acyl glutamine conjugate). More information about these two particular metabolites could prove useful in understanding the relationship between fatty acids and cognitive function. Beyond cognitive performance, omega fatty acids have also been shown to be dysregulated in certain brain regions of patients with AD pathology (Snowden et al., 2017), further strengthening the potential relevance of fatty acids.

This study had several limitations. The pathway analysis we performed was highly limited due to the large number of metabolites in our panel that did not have

KEGG compound IDs. This greatly underscores the importance of continued efforts to identify and characterize metabolites. The PSs we developed for our MR assessment were weak instrumental variables and did not allow us to determine whether levels of the metabolites we identified are causally related to executive function. Our cohort may not have experienced sufficient cognitive decline at this point to identify metabolite levels associated with delayed recall trajectories. Because our analyses are based on an average of two and up to three longitudinal samples per participant, we are somewhat limited in our ability to assess cognitive and metabolite trajectories. Although our metabolite panel is large relative to previous investigations, it is possible that a different panel of metabolites could produce different results; however, quantifying and identifying metabolites is a challenging task that is highly dependent on technological advances. It will be critical to replicate findings presented here with an independent cohort. To assist with replication of metabolomics findings, as the field of metabolomics rapidly develops, it will be crucial to develop standard methods of measuring and analyzing metabolites.

Using a large panel of longitudinal metabolomics data, we found that levels of certain plasma metabolites, including cysteine S-sulfate and erucate, were associated with age-related change in executive function, one of the earliest aspects of cognitive function to change in the course of AD development. Replication in cohorts with longitudinal metabolomics data will be necessary to confirm whether these metabolites contribute to the development of AD. If these metabolites are shown to have causal influences on cognition through future longitudinal and experimental research studies,

subsequent investigations of their nutritional influences could further elucidate the mechanisms influencing early stages of AD and perhaps inform preventative measures.

Tables

Table 1. WRAP Participant Characteristics at Baseline.

Mean (SD) or N (%).

Characteristic	Overall (N=1,206, obs=2,338)
Age (years)	60.9 (6.7)
Female	834 (69.2)
Years of education	16.3 (2.9)
Caucasian	1,130 (93.7)
APOE ε4 carrier	462 (38.3)
Cholesterol lowering medication	385 (31.9)
Sample storage (days)	1,517.3 (405.1)
# Visits	1.9 (0.6)

obs=observations

Table 2. Top ten metabolite*age interactions on executive function.

Metabolite	Super Pathway	Sub Pathway	*P*-value
Cysteine S-sulfate	Amino acid	Methionine, Cysteine, SAM and Taurine Metabolism	**5.2e-05**
X – 13866 ($C_{12}H_{18}O_5$)	Partially Characterized Molecules	Partially Characterized Molecules	**7.3e-05**
X – 12839 (fatty acid 8:1 acyl glutamine conjugate)	Partially Characterized Molecules	Partially Characterized Molecules	**7.9e-05**
Erucate (22:1n9)	Lipid	Long Chain Fatty Acid	**1.5e-04**
Glycine conjugate of $C_{10}H_{12}O_2$	Partially Characterized Molecules	Partially Characterized Molecules	**1.8e-04**
X – 18887	Unknown	Unknown	**2.2e-04**
X – 12104 (fatty acid 6:1 acyl glutamine conjugate)	Partially Characterized Molecules	Partially Characterized Molecules	**5.4e-04**
Dihomo-linolenoyl-choline	Lipid	Fatty Acid Metabolism (Acyl Choline)	5.4e-04
N6-acetyllysine	Amino acid	Lysine Metabolism	5.5e-04
Heptenedioate (C7:1-DC)	Lipid	Fatty Acid, Dicarboxylate	5.9e-04

P-values are unadjusted.

Bold values are statistically significant using a Benjamini-Hochberg adjustment for multiple comparisons.

Figures

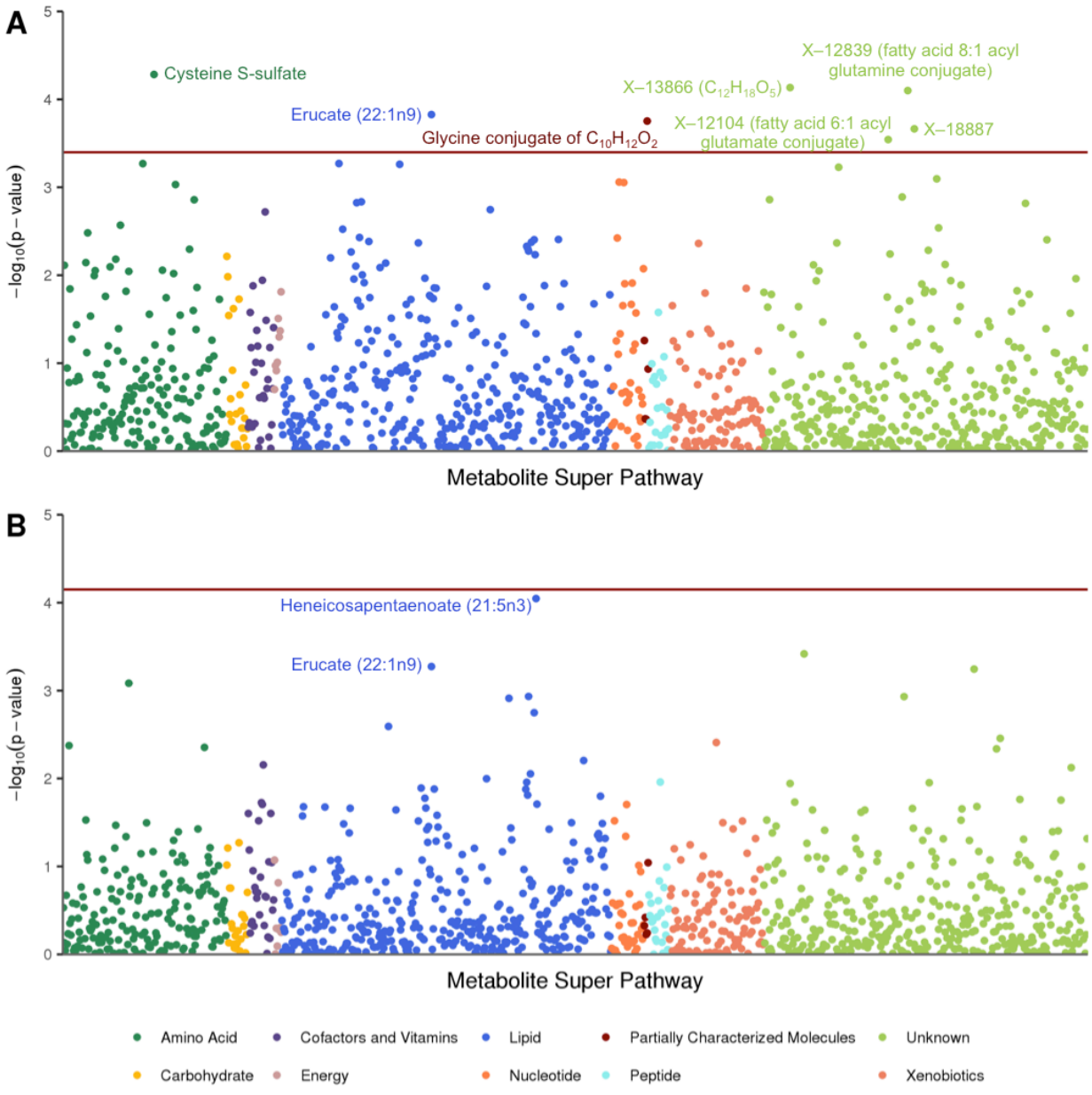


Figure 1. Manhattan plot of metabolome-wide association results for cognitive composite scores.

A. Seven metabolite*age interactions were significantly associated with executive function. B. No metabolite*age interactions were significantly associated with delayed

recall. Both sets of results used a Benjamini-Hochberg adjusted *P*-value threshold (red horizontal line).

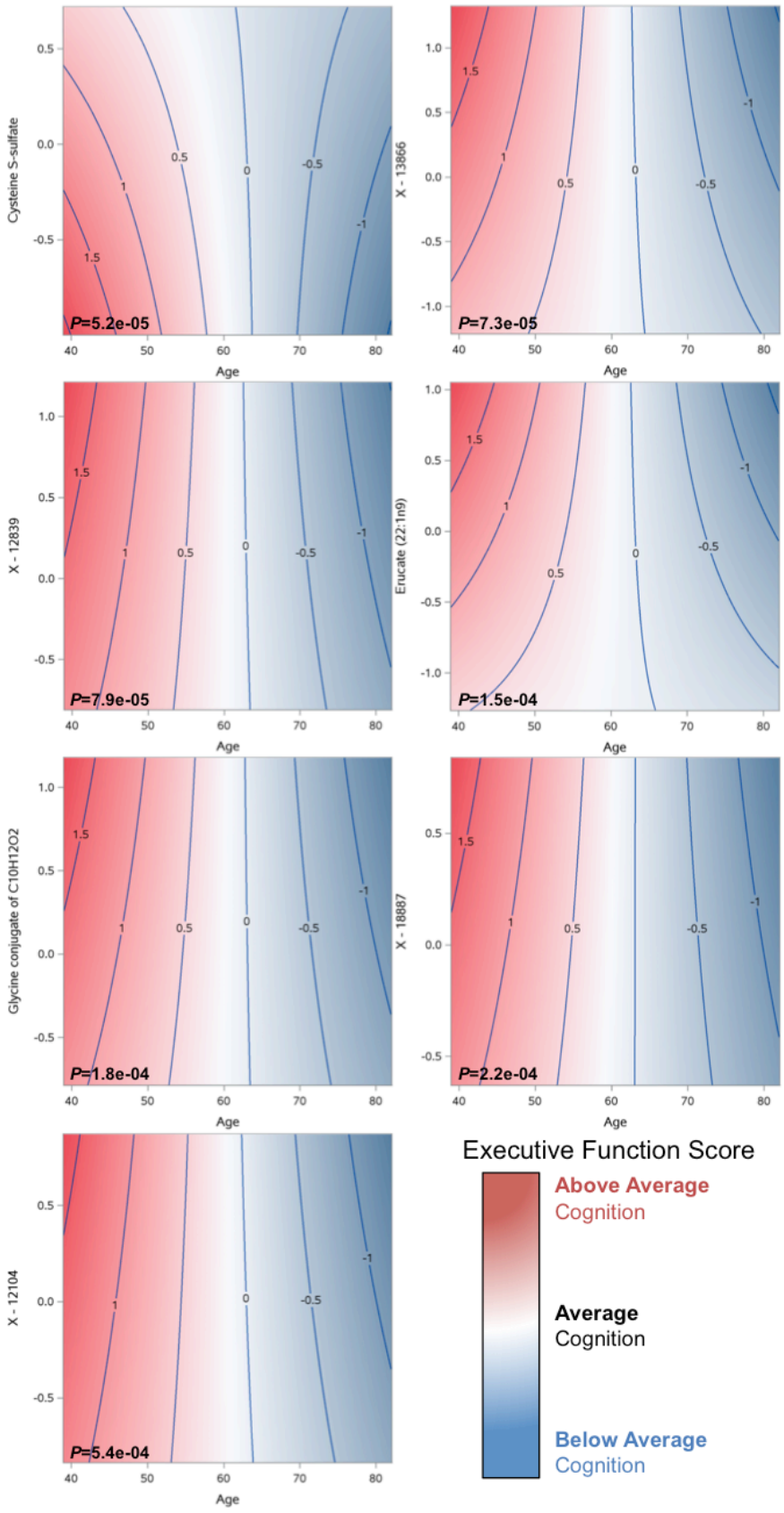

Cysteine S-sulfate
X - 13866
X - 12839
Erucate (22:1n9)
Glycine conjugate of C10H12O2
X - 18887
X - 12104
Age
P=5.2e-05
P=7.3e-05
P=7.9e-05
P=1.5e-04
P=1.8e-04
P=2.2e-04
P=5.4e-04
Executive Function Score
Above Average Cognition
Average Cognition
Below Average Cognition

Figure 2. Contour plots showing executive function trajectories by seven time-varying metabolite levels.

The y-axis represents standardized metabolite levels, whereas the z-axis represents the executive function composite score. In younger ages, higher levels of most metabolite are associated with above average cognition, whereas in older ages, higher levels are associated with below average cognition, with the exception of cysteine s-sulfate, where there opposite is true. Unadjusted *P*-values are indicated for each test.

Chapter 3: Integrated network analysis of genomics, metabolomics, and Alzheimer's risk factors

Abstract

Although Alzheimer's disease (AD) is highly heritable, genetic variants known to be associated with AD only explain a small proportion of its heritability. Genetic factors may only convey disease risk in individuals with certain environmental exposures, suggesting that a multi-omics approach could reveal the underlying mechanisms contributing to complex traits, such as AD. We developed an integrated network to investigate relationships between metabolomics, genomics, and AD risk factors using Wisconsin Registry for Alzheimer's Prevention participants. Analyses included 1,111 non-Hispanic Caucasian participants with whole blood expression for 11,376 genes (imputed from dense genome-wide genotyping), 1,097 fasting plasma metabolites, and 17 AD risk factors. A subset of 155 individuals also had 364 fasting cerebral spinal fluid (CSF) metabolites. After adjusting each of these 12,854 variables for potential confounders, we developed an undirected graphical network, representing all significant pairwise correlations upon adjusting for multiple testing. There were many instances of genes being indirectly linked to AD risk factors through metabolites, suggesting that genes may influence AD risk through particular metabolites. Follow-up analyses suggested that glycine mediates the relationship between *CPS1* and measures of cardiovascular and diabetes risk, including body mass index, waist-hip ratio, inflammation, and insulin resistance. We also found that 38 CSF metabolites explained more than 60% of the variance of CSF levels of tau, a detrimental protein that accumulates in the brain of AD patients and is necessary for its diagnosis. These results

further our understanding of underlying mechanisms contributing to AD risk while demonstrating the utility of generating and integrating multiple omics data types.

Background

Genome-wide association studies (GWAS) have identified thousands of single nucleotide polymorphism (SNP)-trait associations (MacArthur et al., 2017). However, these variants tend to have very small effect sizes and typically explain a small portion of trait heritability. Alzheimer's disease (AD) is an example of such a trait: 53% of its phenotypic variance can be explained by genomic variants, collectively (*i.e.*, SNP heritability); yet, the 21 GWAS variants identified in a meta-analysis to be associated with AD only account for 31% of its genetic variance, leaving 69% unaccounted for (Ridge et al., 2016). In order to more comprehensively understand the disease risk conveyed by genetic factors, it is crucial to consider genomics in combination with other omics data types and to use integrative multi-omics approaches that can capture intricate relationships.

Although there has been great interest recently in the integration of multi-omics datasets, progress in this field is still fairly limited and it faces many challenges (Bersanelli et al., 2016; Buescher & Driggers, 2016; Gligorijevic & Przulj, 2015; Huang, Chaudhary, & Garmire, 2017; Lopez de Maturana, Pineda, Brand, Van Steen, & Malats, 2016; Ritchie et al., 2015). However, studies have been able to show that the use of multiple omics data types is more predictive than single data types (Buescher & Driggers, 2016; Mankoo, Shen, Schultz, Levine, & Sander, 2011). A recent study with dense longitudinal omics data displayed the utility of integrating such data with regards

to personalized medicine (Price et al., 2017). Although limited by its sample size of 108 participants, this investigation identified meaningful systems biology relationships that were able to improve the health of its participants. As it is becoming more feasible and common to acquire multiple omics data types, it is essential that we move towards systems biology approaches of understanding complex diseases, rather than focusing on single data types that are unable to capture the intricacies imposed by biology.

Recent technological advances have made metabolomics studies increasingly favorable among investigations of AD (Enche Ady et al., 2017), obesity (Zhang, Sun, & Wang, 2017), and cardiovascular disease (Ussher, Elmariah, Gerszten, & Dyck, 2016), to name a few. An appeal of the metabolome is that of the biological systems, metabolomics could offer an effective way to accurately capture individual-level environmental exposures; it is the most proximal to the development of the phenotype (Horgan & Kenny, 2011) and many metabolites have a low heritability (Long et al., 2017; Shin et al., 2014), implying that such metabolites are more strongly influenced by the environment than genomics. Metabolomic variations that precede disease onset could prove to be highly informative for predictive models as well as preventative and therapeutic medicine. Pathological changes that cause AD are known to begin decades before the diagnosis of AD (Jack et al., 2009). As such, an integrated approach of studying the genomics and metabolomics of risk factors that precede an AD diagnosis could provide a better understanding of the underlying biological and environmental mechanisms that lead to the onset of AD.

We developed an integrative network to investigate relationships between plasma metabolomics, cerebral spinal fluid (CSF) metabolomics, genomics, and AD risk

factors using 1,111 participants with deep longitudinal phenotypes from the Wisconsin Registry for Alzheimer's Prevention (WRAP). AD risk factors included neuropsychological measures of cognitive function, CSF levels of the two proteins required for an AD diagnosis that are known to accumulate in the brains of AD patients, amyloid-beta (Aβ) and tau, and measures of cardiovascular disease and diabetes risk, two diseases that are known to increase AD risk. Further, in order to understand whether plasma metabolite levels are representative of metabolites in CSF, which may be a more relevant tissue for neurological diseases, we also assessed the correlation of plasma and CSF metabolite levels.

Methods

Participants

Study participants were from WRAP, a longitudinal study of initially dementia free middle-aged adults that allows for the enrollment of siblings and is enriched for a parental history of Alzheimer's disease. Further details of the study design and methods used have been previously described (Johnson et al., 2018; Sager et al., 2005). Participants included in this analysis had genetic ancestry that was primarily of European descent, had both genomic and metabolomic data available, and up to nineteen AD risk factors (Table 1; of note, cholesterol is not included in this table because it was measured on the metabolite panel). This study was conducted with the approval of the University of Wisconsin Institutional Review Board, and all subjects provided signed informed consent before participation.

Plasma and CSF collection and sample handling

Fasting blood samples for this study were drawn the morning of each study visit. Plasma samples were stored in ethylenediaminetetraacetic acid (EDTA) tubes at -80°C. Blood was collected in 10 mL ethylenediaminetetraacetic acid (EDTA) vacutainer tubes. They were immediately placed on ice, and then centrifuged at 3000 revolutions per minute for 15 minutes at room temperature. Plasma was pipetted off within one hour of collection. Plasma samples were aliquoted into 1.0 mL polypropylene cryvolials and placed in -80°C freezers within 30 minutes of separation.

As previously described (Darst et al., 2017), CSF was collected via lumbar puncture (LP) in the morning after a 12-hour fast, not necessarily on the same day as a study visit (LPs were drawn within a median of 120 days of the study visit, ranging from 0-661 days). LPs were performed using a Sprotte 25- or 24-gauge spinal needle at the L3/4 or L4/5 interspace using gentle extraction into polypropylene syringes. CSF (22 mL) was then gently mixed and centrifuged at 2000g for 10 minutes. Supernatants were frozen in 0.5 mL aliquots in polypropylene tubes and stored at −80°C.

Plasma and CSF samples were never thawed before being shipped overnight on dry ice to Metabolon (Durham, NC), where they were again stored in -80°C freezers and thawed once before testing.

CSF biomarker quantification

CSF $A\beta_{42}$, total tau (T-tau), and phosphorylated tau (P-tau) were quantified with sandwich ELISAs (INNOTEST β-amyloid1-42, hTAU-Ag, and Phospho-Tau[181P], respectively; Fujirebio Europe, Ghent, Belgium). CSF levels of $A\beta_{42}$ and $A\beta_{40}$ (a less amyloidogenic Aβ fragment as compared to $A\beta_{42}$) used to calculate the ratio of $A\beta_{42}/A\beta_{40}$ were quantified by electrochemiluminescence (ECL) using an Aβ triplex

assay (MSD Human Aβ peptide Ultra-Sensitive Kit, Meso Scale Discovery, Gaithersburg, MD). A total of 223 samples with CSF biomarkers among 141 individuals were available for this analysis.

Plasma and CSF metabolomic profiling and quality control

Untargeted plasma and CSF metabolomic analyses and quantification were performed by Metabolon (Durham, NC) using Ultrahigh Performance Liquid Chromatography-Tandom Mass Spectrometry (UPLC-MS/MS) (Evans et al., 2014); details are outlined in Appendix A1. Metabolites within eight super pathways were identified: amino acids, carbohydrates, cofactors and vitamins, energy, lipids, nucleotides, peptides, and xenobiotics.

Up to three longitudinal plasma samples were available for each participant. Plasma metabolites with an interquartile range of zero (*i.e.*, those with very low or no variability) were excluded from analyses (178 metabolites). After removing these metabolites, samples were missing a median of 11.7% plasma metabolites, while plasma metabolites were missing in a median of 1.2% of samples.

Up to four longitudinal CSF samples were available for each participant. Similarly, CSF metabolites with an interquartile range of zero were excluded from analyses (48 CSF metabolites). After removing these metabolites, samples were missing a median of 6.9% CSF metabolites, while CSF metabolites were missing in a median of 0.3% of samples.

Missing plasma and CSF metabolite values were imputed to the lowest level of detection for each metabolite. Metabolite values were median-scaled and log-transformed to normalize metabolite distributions (van den Berg et al., 2006). If a

participant reported that they did not fast or withhold medications and caffeine for at least eight hours prior to the blood draw, the plasma sample was excluded from analyses (159 plasma samples), leaving 1,097 plasma metabolites among 2,189 plasma samples (1,111 individuals) for analyses. Similarly, if a participant reported that they did not fast for at least eight hours prior to the LP, the CSF sample was excluded from analyses (4 CSF samples), leaving 364 CSF metabolites among 346 CSF samples (155 individuals) for analyses.

CSF and plasma metabolite correlations

A total of 326 metabolites were captured in both CSF and plasma. The correlation of these metabolites between tissue types was calculated using the Pearson correlation coefficient. In order to reduce within individual variability due to metabolite level changes over time, correlations were based on 141 pairs of plasma and CSF samples that were collected within a timespan of four months of each other. After removing these samples, plasma and CSF samples were collected a median of 27 days apart.

DNA collection and genomics quality control

DNA was extracted from whole blood samples using the PUREGENE® DNA Isolation Kit (Gentra Systems, Inc., Minneapolis, MN). DNA concentrations were quantified using the Invitrogen™ Quant-iT™ PicoGreen™ dsDNA Assay Kit (Thermo Fisher Scientific, Hampton, NH) analyzed on the Synergy 2 Multi-Detection Microplate Reader (Biotek Instruments, Winooski, VT). Samples were normalized to 50 ng/ul following quantification.

A total of 1,340 samples were genotyped using the Illumina Multi-Ethnic Genotyping Array at the University of Wisconsin Biotechnology Center (Appendix A2). Thirty-six blinded duplicate samples were used to calculate a concordance rate of 99.99%, and discordant genotypes were set to missing. Sixteen samples missing >5% of variants were excluded, while 35,105 variants missing in >5% of individuals were excluded. No samples were removed due to outlying heterozygosity. Six samples were excluded due to inconsistencies between self-reported and genetic sex.

Due to the sibling relationships present in the WRAP cohort, genetic ancestry was assessed using Principal Components Analysis in Related Samples (PC-AiR), a method that makes robust inferences about population structure in the presence of relatedness(Conomos et al., 2015). This approach included several iterative steps and was based on 63,503 linkage disequilibrium (LD) pruned (r^2<0.10) and common (MAF>0.05) variants, using the 1000 Genomes data as reference populations (Genomes Project et al., 2015). First, kinship coefficients (KCs) were calculated between all pairs of individuals using genomic data with the Kinship-based Inference for Gwas (KING)-robust method (Manichaikul et al., 2010). PC-AiR was used to perform principal components analysis (PCA) on the reference populations along with a subset of unrelated individuals identified by the KCs. Resulting principal components (PCs) were used to project PC values onto the remaining related individuals. All PCs were then used to recalculate the KCs taking ancestry into account using the PC-Relate method, which estimates KCs robust to population structure (Conomos et al., 2016). PCA was performed again using the updated KCs, and KCs were also estimated again using updated PCs. The resulting PCs identified 1,198 WRAP participants whose

genetic ancestry was primarily of European descent. This procedure was repeated within this subset of participants (excluding 1000 Genomes individuals) to obtain PC estimates used to adjust for population stratification in subsequent genomic analyses. Among European descendants, 160 variants were not in Hardy-Weinberg equilibrium (HWE) and 327,064 were monomorphic and thus, removed.

A total of 1,294,660 bi-allelic autosomal variants among 1,198 European descendants remained for imputation, which was performed with the Michigan Imputation Server v1.0.3 (Das et al., 2016), using the Haplotype Reference Consortium (HRC) v. r1.1 2016 (McCarthy et al., 2016) as the reference panel and Eagle2 v2.3 (Loh et al., 2016) for phasing. Prior to imputation, the HRC Imputation Checking Tool(Rayner et al., 2016) was used to identify variants that did not match those in HRC, were palindromic, differed in MAF>0.20, or that had non-matching alleles when compared to the same variant in HRC, leaving 898,220 for imputation. A total of 39,131,578 variants were imputed. Variants with a quality score R^2<0.80, MAF<0.001, or that were out of HWE were excluded, leaving 10,400,394 imputed variants. These were combined with the genotyped variants, leading to 10,499,994 imputed and genotyped variants for analyses. Data cleaning and file preparation were completed using PLINK v1.9 (Chang et al., 2015) and VCFtools v0.1.14 (Danecek et al., 2011). Coordinates are based on GRCh37 assembly hg19.

Whole blood gene expression imputation

The resulting 10,499,994 imputed and genotyped variants were used to impute whole blood gene expression using PrediXcan (Gamazon et al., 2015) with the Depression Genes and Networks reference dataset (Battle et al., 2014), PrediXcan's

largest reference sample consisting of 922 individuals with RNA sequencing on whole blood and GWAS data. PrediXcan filters results to only include genes that are imputed with reasonable accuracy, using a false discovery rate of 0.05. After removing genes with zero variability between individuals (162 genes), whole blood gene expression data for 11,376 genes were available for analyses.

Integrative network analysis

The approach we used for our network analysis was similar to that of Price et al., 2017 (Price et al., 2017). A total of 12,854 variables, including 11,376 expressed genes, 1,097 plasma metabolites, 364 CSF metabolites, and 17 AD risk factors, were available for the network analysis. Linear mixed models, as implemented by the lme4 package in R (Bates, Machler, Bolker, & Walker, 2015), were used to adjust each variable for age and sex and included a random intercept for family. Further adjustments were made specific to the variable being assessed: imputed gene expression was also adjusted for the first four principal components to account for ancestry; CSF and plasma metabolites were adjusted for cholesterol lowering medication use and sample storage time; the executive function and delayed recall composite scores were adjusted for practice effects; and systolic and diastolic blood pressure were adjusted for ace inhibitor and beta blocker medication use. For longitudinal traits (such as metabolites), random intercepts were used as the new outcomes for each individual, whereas for constant traits (such as imputed gene expression values), residuals were used as the new outcomes for each individual. These adjusted outcomes were used to assess all 82,606,231 pairwise correlations between traits using Spearman rank, and significance was determined using a Bonferroni-adjusted *P*-value (0.05/82,606,231=6.1e-10). To

identify relationships between omics data, significant inter-omic associations and significant associations with an AD risk factor were used to develop an integrative network, which was created using the igraph R package (Csardi & Nepusz, 2006). Dense subgraphs were identified using a community detection algorithm that maximizes the modularity of the network, such that there is high connectivity within communities (or groups of distinct variables), but low connectivity between communities (Clauset, Newman, & Moore, 2004).

Targeted mediation and interaction analyses

Results from the integrated network analysis were used to identify potential mediation and interactions between imputed gene expression and metabolite levels that could impact AD risk factors, as a proof of concept. Although our network analysis suggested many potentially meaningful mediation or interaction relationships, we only investigated the gene-metabolite correlation with the most consistent biological support from the GWAS catalog (www.ebi.ac.uk/gwas, date accessed: May 9, 2018, MacArthur et al., 2017) to illustrate the utility of the network analysis results. Such relationships were investigated using the longitudinal data (2,198 observations among 1,111 individuals) with linear mixed models, again as implemented by the lme4 package in R (Bates et al., 2015), including random intercepts for within-individual repeated measures and within-family relationships. To assess whether a metabolite mediated the relationship between imputed gene expression and am AD risk factor, models were run to assess whether: 1) the gene predicted the AD risk factor, 2) the gene predicted metabolite levels, 3) the metabolite predicted the AD risk factor, and 4) the gene predicted the AD risk factor while adjusting for the metabolite. The causal mediation

effect, or the indirect effect of a gene on an AD risk factor through a metabolite, was calculated as the difference between the effect of the gene in model 1 and model 4, as implemented in the R mediation package (Imai, Keele, & Tingley, 2010). To determine whether this difference was significant, standard errors and *P*-values were estimated using the quasi-Bayesian Monte Carlo method with 1000 simulations. Because the mediation package can only handle mixed models with one random effect, the mediation analysis was run with models 1 and 4 excluding the random effect for family. As a sensitivity analysis, the mediation analysis was rerun limiting models 1 and 4 to unrelated individuals (n=898 with 1,774 observations). A fifth linear mixed model was used to assess interactions by adding a gene*metabolite interaction term to model 4. Model 5 did not use the mediation package and was thus able to include random intercepts for both within-individual repeated measures and within-family relationships. All models including a gene had covariates for age, sex, and the first four PCs, while models including a metabolite had covariates for age, sex, cholesterol lowering medication use, and sample storage time.

Results

Participants

A total of 1,111 WRAP participants had both genomic and plasma metabolomic data. At baseline, 68.9% of participants were female and participants were 61.0 years old with a bachelor's degree, on average (Table 2). Participants each had 1,097 plasma metabolites available for analyses, 347 (31.6%) of which were of unknown chemical structure, whole blood gene expression for 11,376 genes, and up to 17 AD risk factors.

A subset of 155 individuals also had 364 CSF metabolites available for analyses, 56 (15.4%) of which were of unknown chemical structure. Participants with CSF metabolomic data had similar characteristics as the full sample (Table 2). Properties of each plasma and CSF metabolite, such as biochemical name, super pathway, and sub pathway are described in Appendix C1, and numbers of metabolites within each super pathway are summarized in Appendix C2.

Correlation between plasma and CSF metabolomics

The median correlation between the 326 metabolites common to both plasma and CSF was r=0.26, with some variability existing between different metabolite pathways (Figure 1). Xenobiotics had the highest median correlation (r=0.53), while lipids had the lowest (r=0.11). Overall, metabolite correlations ranged from |r|=0.0002 (inosine, a nucleotide) to |r|=0.88 (quinate, a xenobiotic). Interestingly, one of the highest correlations was caffeine (r=0.81). Correlations between each of the 326 CSF and plasma metabolites are described in Appendix C3.

Integrated network

After applying a Bonferroni correction for multiple testing, a total of 90,308 significant correlations (edges) among 10,869 variables (nodes) were used to develop an overall ‘hairball’ network (Appendix C4). Notably, although there were far fewer metabolites than genes in the network (1,387 metabolites versus 9,481 genes), there were more edges between metabolites than genes (49,499 versus 37,473 edges, respectively).

The inter-omic network is shown in Figure 2 (a labeled version is shown in Appendix C5), and its corresponding community partitions are shown in Appendix C6.

This network had 1,224 edges and 635 nodes, including 171 metabolite-gene and 833 metabolite-AD risk factor edges. Of these, there were only four CSF metabolite-gene edges and 73 CSF metabolite-AD risk factor edges, likely due to the much smaller number of CSF metabolomic samples. No genes were directly linked to AD risk factors; however, many genes were indirectly linked to AD risk factors through metabolites, as described below. Each of the 1,224 correlations is described in Appendix C7.

The largest community contained 680 edges among 289 nodes, which included 264 plasma metabolites, ten CSF metabolites, eight genes, and seven AD risk factors related to cardiovascular disease and diabetes: body mass index (BMI), waist-hip ratio (WHR), homeostatic model assessment of insulin resistance (HOMA-IR), interleukin 6 (IL-6), metabolic equivalents (METs), diastolic blood pressure (DBP), and systolic blood pressure (SBP) (Appendix C8). Expression levels of these eight genes were all indirectly linked to AD risk factors within this community through plasma metabolites. *CPS1* expression levels were negatively correlated with plasma gamma-glutamylglycine, proprionylglycine, and glycine levels, all of which were negatively correlated with BMI, WHR, IL-6, and/or HOMA-IR (Figure 3). *TMEM229B* and *PLEKHH1* were both negatively correlated with two glycerophosphatidylcholines (1-(1-enyl-palmitoyl)-2-palmitoleoyl-GPC (P-16:0/16:1) and 1-(1-enyl-palmitoyl)-2-palmitoyl-GPC (P-16:0/16:0)), which were also negatively correlated with BMI, WHR, and/or HOMA-IR. *NAALAD2* was negatively correlated with an amino acid beta-citrylglutamate, which was positively correlated with BMI, WHR, IL-6, and HOMA-IR. *ZNF655* and *ZKSCAN1* were both positively correlated with X-12063, which was also positively correlated with BMI, WHR, and HOMA-IR. *CHRNA5* was positively correlated with 5-

hydroxylysine, which was positively correlated with BMI, WHR, IL-6, and HOMA-IR, and negatively correlated with METs. *ARVCF* was negatively correlated with X-11593, which was positively correlated with BMI, IL-6, and HOMA-IR.

Several genes outside of the cardiovascular and diabetes community were indirectly linked to AD risk factors within this community. Gene expression of *FOSL2*, *KRTCAP3*, and *ZNF513* were positively correlated, while *IFT172*, *NRBP1*, *PPM1G*, and *ZNF512* were negatively correlated, with levels of plasma mannose, a carbohydrate that was positively correlated with BMI, WHR, IL-6, and HOMA-IR (Appendix C9A). *CABP1*, *SPPL3*, and *UNC119B* expression levels were negatively correlated with plasma butyrylcarnitine (C4), which was positively correlated with BMI, WHR, IL-6, and HOMA-IR (Appendix C9B). *SLC27A4*, *PHYHD1*, *ENDOG*, and *SH3GLB2* expression levels were negatively correlated with plasma 2'-O-methyluridine and 2'-O-methylcytidine levels, both nucleotides involved in pyrimidine metabolism, and the latter nucleotide is also negatively correlated with BMI and WHR (Appendix C9C). *PHYHD1* was also negatively correlated with CSF levels of 2'-O-methylcytidine.

The only correlations identified among the CSF biomarkers (*i.e.*, amyloid and tau) are shown in Figure 4. Higher CSF T-tau and P-tau levels were correlated with higher levels of 38 CSF metabolites, collectively. These metabolites included 13 lipids (six phosphatidylcholines, two lysophosphatidylcholines, five sphingolipids, and cholesterol), seven amino acids, five carbohydrates, one nucleotide, one energy metabolite, one cofactor and vitamin metabolite, one xenobiotic, and nine unknown metabolites. However, none of the CSF amyloid biomarkers were correlated with CSF metabolites. We investigated how much of the variance of T-tau and P-tau could be explained by

these metabolites with linear mixed models that included random intercepts for within-subject repeated measures and within-family relationships, using the R^2 statistic for mixed models as defined by Edwards et al., 2008 (Edwards, Muller, Wolfinger, Qaqish, & Schabenberger, 2008) and implemented in the r2glmm R package. After removing the variation explained by age and sex, the 37 metabolites correlated with T-tau explained 60.7% of the variation of T-tau, while the 35 metabolites correlated with P-tau explained 64.0% of the variation of P-tau.

Targeted mediation and interaction analyses

Targeted mediation and interaction analyses were focused on a particular pathway identified within the large cardiovascular and diabetes community involving *CPS1*, glycine plasma metabolites (glycine, proprionylglycine, and gamma-glutamylglycine), BMI, WHR, IL-6, and HOMA-IR. Associations between *CPS1* variants and glycine have been reported in at least nine studies (Demirkan et al., 2015; Draisma et al., 2015; Kettunen et al., 2016; Long et al., 2017; Raffler et al., 2015; Shin et al., 2014; Suhre et al., 2011; Xie et al., 2013; B. Yu et al., 2014), more than any of the other gene-metabolite associations identified in our network analysis, and these studies were based not only on Caucasian populations, but also on Japanese and African American populations. Many previous studies have also reported associations between glycine and cardiovascular risk factors, including BMI, waist circumference, inflammation, and HOMA-IR (Cheng et al., 2012; Demirkan et al., 2015; Ding et al., 2015; El Hafidi, Perez, & Banos, 2006; Felig, Marliss, & Cahill, 1969; Geidenstam et al., 2017; Kraus et al., 2016; Takashina et al., 2016; Thalacker-Mercer et al., 2014). This evidence made this pathway a strong candidate for mediation and interaction analyses.

Figure 5 shows results from the mediation analyses using glycine as the mediator, including the total effect (*i.e.*, the effect of *CPS1* in the model unadjusted for glycine), the direct effect (*i.e.*, the effect of *CPS1* in the model adjusted for glycine), and the indirect effect (*i.e.*, the effect of *CPS1* due to the effect of *CPS1* on glycine) for BMI (Figures 5A and 5B), WHR (Figures 5C and 5D), IL-6 (Figure 5E and Figure 5F), and HOMA-IR (Figures 5G and 5H). The total effect of *CPS1* was null for each of these four outcomes, likely due to the negative association between *CPS1* and glycine coupled with the negative association between glycine and the AD risk factor, resulting in direct and indirect effects that had opposing directions (Richiardi, Bellocco, & Zugna, 2013). Our results show that lower levels of *CPS1* expression lead to increased glycine levels, and higher glycine levels lead to decreased BMI, WHR, IL-6, and HOMA-IR. Thus, with glycine as a mediator, lower levels of *CPS1* lead to decreased BMI, WHR, IL-6, and HOMA-IR. Mediation analyses using propionylglycine and gamma-glutamylglycine as the mediator showed similar results and can be found in Appendix C10 and Appendix C11. We did not identify any interactions between *CPS1* and glycine that were associated with BMI, WHR, IL-6, or HOMA-IR (all *P*-values>0.07).

Discussion

We developed an integrative network to investigate relationships between genomics, plasma metabolomics, CSF metabolomics, and AD risk factors. Although no gene expression levels were directly correlated with AD risk factors, there were many instances of genes being indirectly correlated with AD risk factors though metabolites. Building on one such instance, we found that glycine mediated the pathway between

CPS1 expression and cardiovascular and diabetes risk factors. This suggests that our results may have generated many valid hypotheses that warrant further investigation. We also found that correlations between plasma and CSF metabolites ranged widely but had a low average correlation. This could suggest that most plasma metabolites are not representative of certain metabolic changes occurring in the brain, although we cannot rule out the possibility that the low average correlation is, at least partially, due to the time difference between the plasma and CSF sample collection.

The low correlation we observed between plasma and CSF metabolite levels could be related to ~98% of small molecules not being able to pass the blood-brain barrier (BBB) (Pardridge, 2005). Cholesterol is an example of a lipid metabolite that typically cannot pass the BBB (Bjorkhem & Meaney, 2004), and was not correlated between tissues (r=-0.07). On the other hand, caffeine (a xenobiotic) readily crosses the BBB (McCall, Millington, & Wurtman, 1982) and it was highly correlated between tissues (r=0.81), as was 5-acetylamino-6-amino-3-methyluracil (r=0.82), which is a caffeine metabolite, and theophylline (r=0.82), which is structurally and pharmacologically similar to caffeine. This could contribute to lipids having the weakest average correlation and xenobiotics having the strongest average correlation between plasma and CSF tissues. However, it is important to note that metabolites within a given pathway can vary widely from each other and should be considered on an individual basis, accordingly, as the averages presented here may not reflect a particular metabolite's unique properties. The hypothesis about plasma and CSF differing due to the BBB is also supported by the only correlations in the network analysis involving CSF biomarkers (*i.e.*, tau) being with CSF metabolites, although we cannot rule out the possibility that this correlation is

related to CSF biomarkers and CSF metabolomics being analyzed from the same sample and thus, not having time-related variation.

Our network analysis revealed that 38 CSF metabolites were highly predictive of CSF T-tau and P-tau, collectively explaining 60.7% and 64.0% of the variance of T-tau and P-tau, respectively. Further investigations of these CSF metabolites could lead to a better understanding of mechanisms and pathways that influence the development of tau tangles. In contrast, no CSF metabolites were correlated with CSF amyloid biomarkers, which could have implications about the biological function of amyloid versus tau. It is possible that we did not capture small molecules that amyloid may be associated with, or that amyloid is generally not associated with small molecules. Although our CSF findings were limited by their small sample size, they offer potentially novel information regarding the interface between CSF biomarkers and CSF metabolites, as we have not identified previous studies investigating these relationships.

One advantage of using imputed gene expression data is that it only represents the genetically regulated component of gene expression, reducing the risk of confounding due to environmental factors and reverse causality in mediation analyses. We found that glycine mediated the relationship between *CPS1* and BMI, WHR, IL-6, and HOMA-IR, such that lower *CPS1* expression was associated with higher levels of glycine, which were associated with lower BMI, WHR, IL-6, and HOMA-IR. Relationships between *CPS1*, glycine, and cardiovascular risk factors have been hypothesized recently, but not clearly defined (Matone et al., 2016; Xie et al., 2013). The *CPS1* (Carbamoyl-Phosphate Synthase 1) gene encodes for a mitochondrial enzyme that catalyzes the first step of the hepatic urea cycle by synthesizing carbamoyl

phosphate from ammonia, bicarbonate, and two molecules of ATP, and is important for removal of urea from cells (Haberle et al., 2011). Notably, all genes encoding enzymes involved in the urea cycle are expressed in the brain, including *CPS1* (Hansmannel et al., 2010), and levels of enzymes and metabolic intermediates involved in the urea cycle are altered in AD patients (Griffin & Bradshaw, 2017). *CPS1* variants have been linked to CPS1 deficiency (Haberle et al., 2011), neonatal pulmonary hypertension (Pearson et al., 2001), vascular function (Summar et al., 2004), traits related to blood clotting, such as fibrinogen levels and platelet count (Astle et al., 2016; Danik et al., 2009; de Vries et al., 2016; Sabater-Lleal et al., 2013), homocysteine levels (Lange et al., 2010; Pare et al., 2009; van Meurs et al., 2013; Williams et al., 2014), HDL cholesterol (Willer et al., 2013), kidney function and disease (Gorski et al., 2017; Kottgen et al., 2010; Mahajan et al., 2016; Pattaro et al., 2016), AD (Jun et al., 2016), and BMI (Locke et al., 2015; Melen et al., 2013). Higher adipose tissue expression of *CPS1* has been associated with detrimental traits, including weight gain (Matone et al., 2016). At least nine studies have reported associations between *CPS1* variants and glycine (Demirkan et al., 2015; Draisma et al., 2015; Kettunen et al., 2016; Long et al., 2017; Raffler et al., 2015; Shin et al., 2014; Suhre et al., 2011; Xie et al., 2013; B. Yu et al., 2014) and others have reported associations with betaine, a derivative of glycine (Hartiala et al., 2016; Long et al., 2017; Shin et al., 2014). Glycine is a common amino acid involved in the production of DNA, phospholipids, and collagen, and in the release of energy. Previous studies have identified negative correlations between glycine and cardiovascular and diabetes risk factors such as BMI, waist circumference, HOMA-IR, obesity and visceral obesity, subcutaneous and visceral fat area, hypertension, and acute myocardial infarction

(Cheng et al., 2012; Demirkan et al., 2015; Ding et al., 2015; El Hafidi et al., 2006; Felig et al., 1969; Geidenstam et al., 2017; Kraus et al., 2016; Takashina et al., 2016; Thalacker-Mercer et al., 2014). These previous findings are in the same direction as our findings and are highly supportive of the biological relevance of our results, which lead us to hypothesize that the *CPS1*-cardiovascular risk pathway is linked through the mediation of glycine.

One particular *CPS1* variant, rs715, has been linked to urine and blood glycine levels (Demirkan et al., 2015; Long et al., 2017; Raffler et al., 2015; Shin et al., 2014; Xie et al., 2013), blood levels of betaine (Hartiala et al., 2016; Long et al., 2017; Shin et al., 2014), blood levels of fibrinogen (de Vries et al., 2016; Sabater-Lleal et al., 2013), and BMI(Locke et al., 2015). This is a common variant, with a MAF=0.27 based on 62,784 whole genome sequences from Trans-Omics for Precision Medicine (TOPMed) (TNT, 2017). The minor C allele of rs715 decreases *CPS1* expression (Hartiala et al., 2016). To further test our findings, we conducted additional mediation analyses using this variant and found highly consistent results, suggesting that having one or two minor alleles of rs715 (which decreases *CPS1* expression) increases levels of the three glycine plasma metabolites, which decreases BMI, WHR, IL-6, and HOMA-IR (Appendix C12-C14). Thus, the minor C allele of rs715 may have a protective role in cardiovascular risk.

One of the primary strengths of this analysis is that it shows the feasibility of performing integrated omics analyses and the potential utility of such approaches. It is becoming more common for cohorts to collect such datasets; for example, the National Institutes of Health is sponsoring the new TOPMed nation-wide consortium that aims to

deeply phenotype its participants utilizing omics technologies (www.nhlbiwgs.org). It is anticipated that initiatives such as TOPMed will greatly advance our knowledge of many complex diseases and traits. However, to fully utilize these rich data, it will be crucial to identify effective means of integrating them and maximize their potential to provide a more holistic understanding of the disease process. While there is still a great need for such methods, our inter-omic network analysis and subsequent targeted follow-up analyses outlines one approach to effectively integrate omics data.

This study was not without limitations. Due to computational burdens, our network analysis did not fully utilize the longitudinal aspect of our data. Further, our sample sizes for CSF biomarkers and metabolites were limited, which is likely why we had few CSF findings in our network analysis. Plasma and CSF samples typically were not collected on the same day, which could influence our correlation results. However, this may not have influenced our network analysis to a large extent because we averaged the residuals of longitudinal traits. We were unable to include smoking behavior in our network analysis due to the prohibitive number of smokers in our cohort (n=48). Despite these limitations, we were encouraged to find that many of our results had been previously reported, thereby strengthening confidence in our novel findings.

Opponents of the “big data” era have criticized omics approaches because they are not hypothesis driven and do not follow the standard scientific method (Duncan, 2007; Stieb, Boot, & Turner, 2017). However, we know biology to be complex far beyond our current understanding. To believe that we currently have the ability to generate valid biological hypotheses to understand complex conditions without data would be a fallacy. This was a lesson learned in the years preceding the completion of

the human genome sequence in 2001 when research efforts were heavily invested into targeted genetic loci and genome-wide linkage screens of ~500 loci (Fallin, Duggal, & Beaty, 2016). While this approach was successful for genes that follow Mendelian patterns, such as highly penetrant variants in the *BRCA1* and *BRCA2* genes that are responsible for inherited forms of breast cancer and the *APP*, *PSEN1*, and *PSEN2* genes that cause the inherited early onset form of AD, it had limited success for traits that follow complex inheritance patterns (Fallin et al., 2016). The utility of omics data, and particularly integrated omics approaches, is the ability to generate data driven hypotheses. Our knowledge of biology has been evolving for centuries; however, with the data we are able to generate due to recent biotechnological advances, we now have the opportunity to advance our knowledge of biology at an unprecedented rate. Such data could lead to dramatic improvements in the state of preventative and therapeutic medicine, particularly for complex diseases such as AD, for which few such preventative or therapeutic methods exist and little is known about the underlying biological mechanisms.

By integrating genomics, metabolomics, and clinical risk factors for AD, we were able to identify complex relationships that could lend to a better understanding of the onset of AD and risk factors associated with its onset. As the generation of omics data accelerates across investigations of a variety of research fields, continued efforts to navigate statistical and computational issues will be critical. The work presented here represents early efforts to integrate omics data, but much more research is needed to identify the most effective means of doing so and thereby maximize the utility of such rich sources of data.

Tables

Table 1. Nineteen AD Risk Factors Included in Network Analysis.

Table 1. Seventeen AD Risk Factors Included in Network Analysis		
Category of Risk Factor	Risk Factor	N
Cognitive	Executive function Composite Score	1,096
	Delayed Recall Composite Score	1,107
	Education	1,111
	Mom's age at memory loss	608
	Dad's age at memory loss	340
Cerebral Spinal Fluid	$A\beta_{42}$	141
	T-tau	141
	P-tau	141
	$A\beta_{42}/A\beta_{40}$	141
Cardiovascular/Diabetic	BMI	1,111
	WHR	1,111
	METs	1,108
	Alcohol use	1,104
	IL-6	1,088
Cardiovascular	SBP	1,111
	DBP	1,111
Diabetic	HOMA-IR	1,107

AD: Alzheimer's disease, $A\beta_{42}$: β-$Amyloid_{42}$, T-tau: Total-tau, P-tau: Phosphorylated-tau, $A\beta_{40}$: β-$Amyloid_{40}$, BMI: Body-mass index, WHR: Waist-hip ratio, METs: Metabolic equivalents, IL-6: Interleukin 6, SBP: Systolic blood pressure, DBP: Diastolic blood pressure, HOMA-IR: Homeostatic model assessment of insulin resistance

Alcohol use=(#drinks/day)*(#days/week)

Table 2. WRAP Participant Characteristics at Baseline Sample. Mean (SD) or N (%).

Characteristic	Overall (N=1,111, obs=2,191)	CSF Metabolomics (N=155, obs=346)
Age (years)	61.0 (6.7)	61.2 (6.6)
Female	766 (68.9)	103 (66.5)
Years of education	16.4 (2.8)	16.7 (2.9)
Parental history of AD	803 (72.3)	112 (72.3)
Use of cholesterol-lowering medication	354 (31.9)	45 (29.0)
# Visits	2.0 (0.6)	2.2 (1.0)

obs=observations

Figures

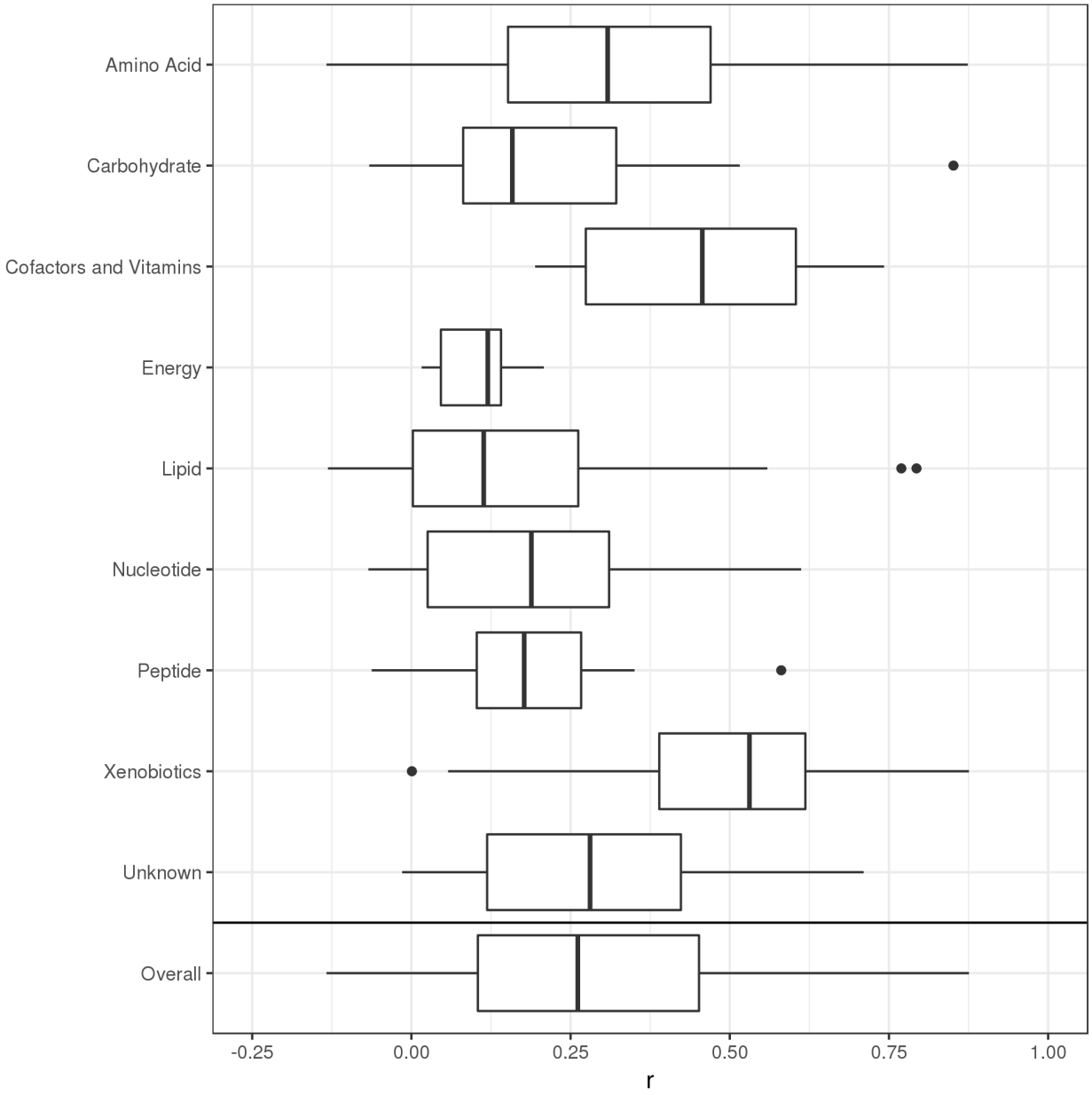


Figure 1. Correlations between plasma and CSF metabolites by super pathway. Vertical bars represent median correlations; box width represents the first and third quartiles; horizontal bars (whiskers) represent the range of correlations that are within 1.5 times the interquartile range; and dots represent outlier correlations that exceed 1.5 times the interquartile range.

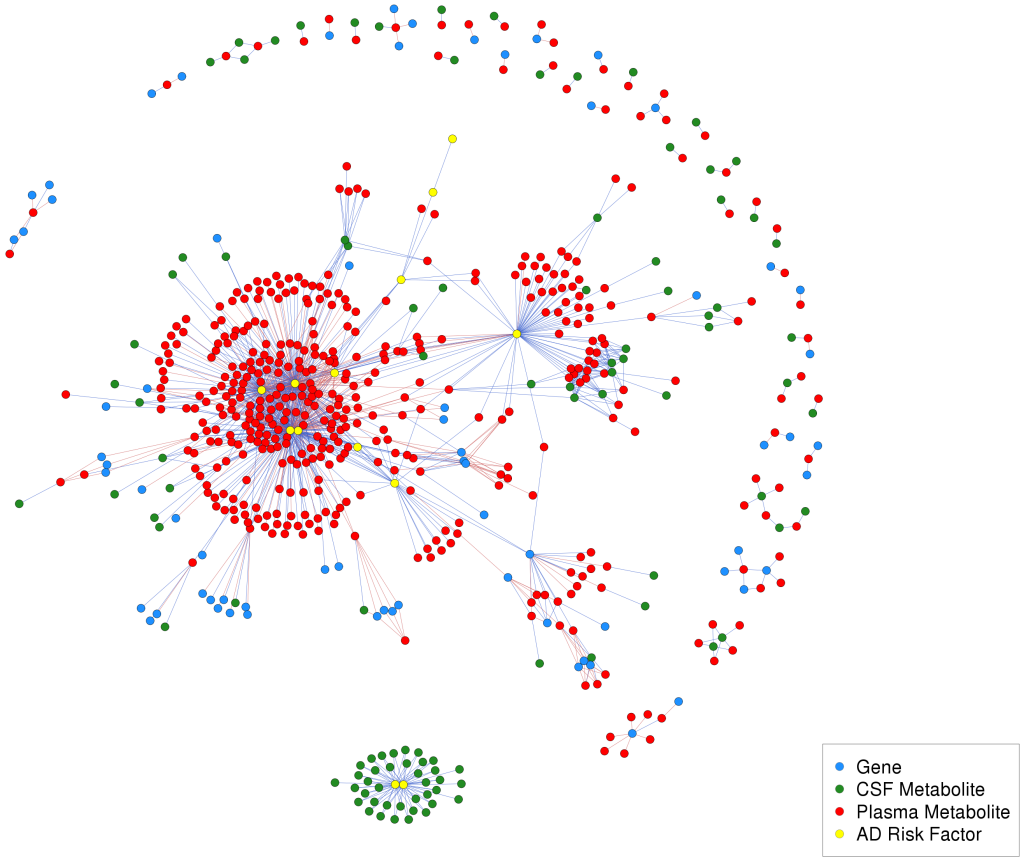


Figure 2. Inter-omic network.

This network has 1,224 edges and 635 nodes, which included 171 metabolite-gene edges, 833 metabolite-AD risk factor edges. Of these, 73 were CSF metabolite-AD risk factor edges (CSF T-tau and P-tau, exclusively) and 4 were CSF metabolite-gene edges. Red edges indicate negative correlations and blue edges indicate positive correlations.

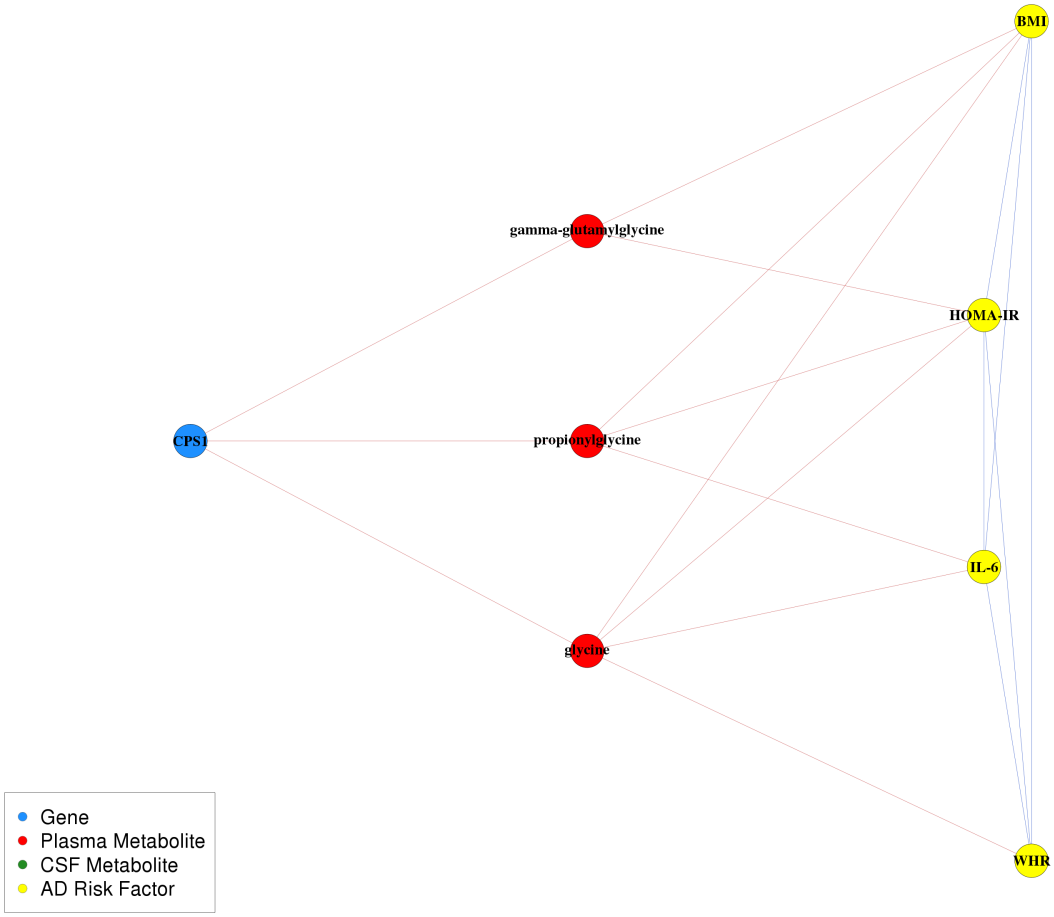


Figure 3. *CPS1*, glycine, and cardiovascular and diabetes sub-network.

CPS1, glycine, and cardiovascular and diabetes sub-network. Relationships within this pathway are highly cited; however, the pathway as a whole is not understood as well. Red edges indicate negative correlations and blue edges indicate positive correlations.

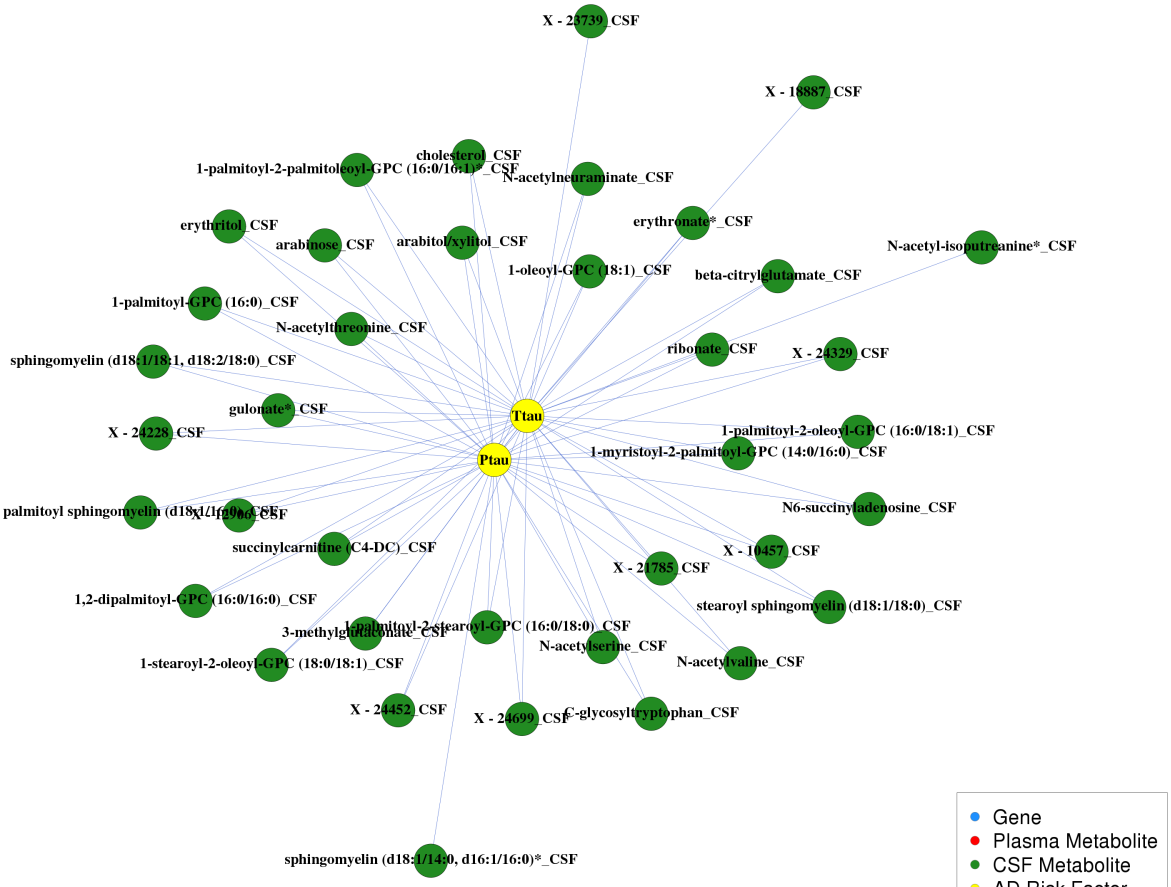


Figure 4. CSF biomarker community.

This network has 73 edges among 38 CSF metabolites and CSF biomarkers T-tau and P-tau. Red edges indicate negative correlations and blue edges indicate positive correlations.

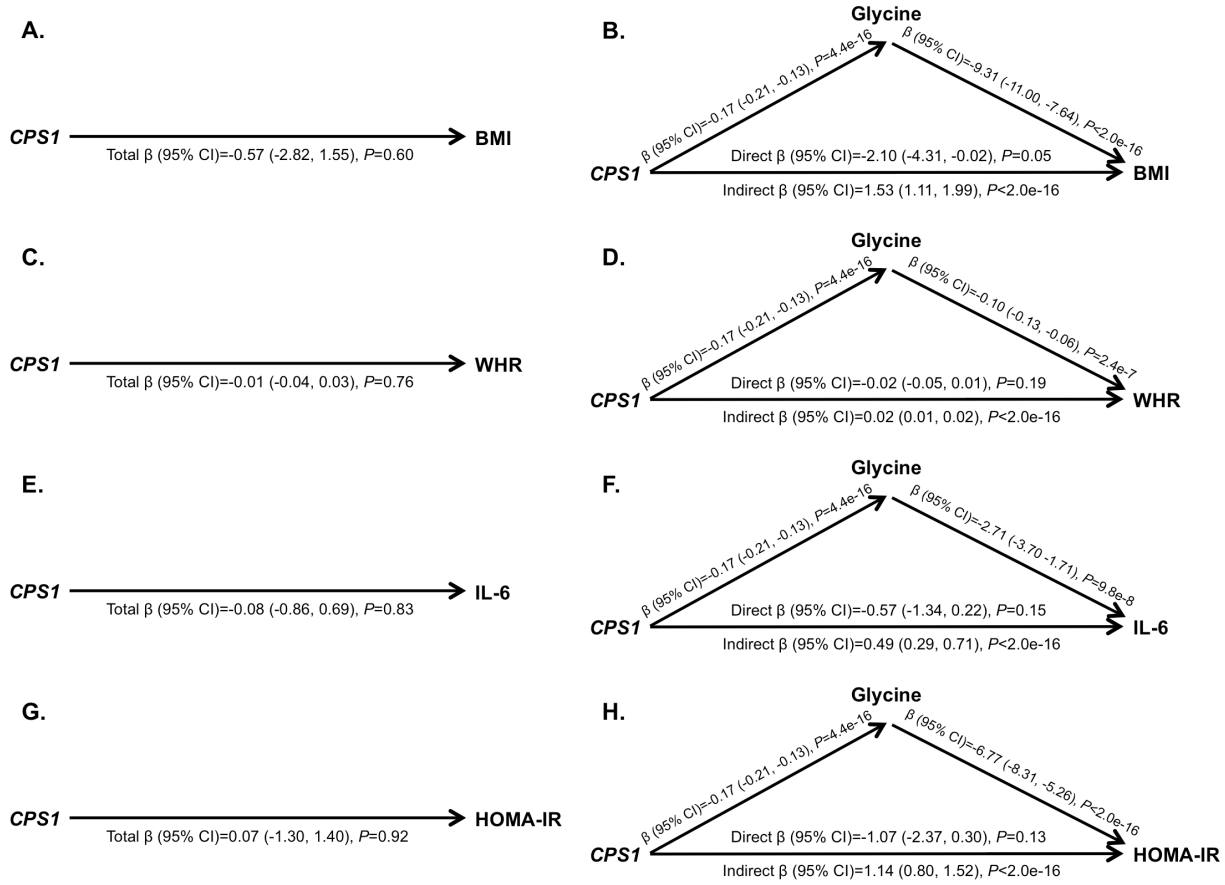


Figure 5. Mediation analyses to assess whether plasma glycine mediates the relationships between imputed *CPS1* expression, BMI, WHR, IL-6, and HOMA-IR. A. Total effect of *CPS1* on BMI. B. Direct and indirect effects of *CPS1* on BMI. C. Total effect of *CPS1* on WHR. D. Direct and indirect effects of *CPS1* on WHR. E. Total effect of *CPS1* on IL-6. F. Direct and indirect effects of *CPS1* on IL-6. F. Total effect of *CPS1* on HOMA-IR. G. Direct and indirect effects of *CPS1* on HOMA-IR. All models adjusted for age and sex; models including *CPS1* additionally adjusted for the first four PCs; models that included glycine additionally adjusted for cholesterol lowering medication use and sample storage time.

Conclusion

The research conducted in this dissertation offers an important improvement regarding our knowledge of risk factors of Alzheimer's disease (AD) in several key ways. First, it offers a deeper knowledge about the impact of aging and sex, two known risk factors for AD, on a panel of plasma metabolite levels, while also describing how heritable each of those metabolites are. Second, it has identified several metabolites and metabolic pathways, particularly those related to fatty acid metabolism and cysteine metabolism, to be associated with an aspect of cognition that declines early in the AD trajectory. Third, using a data-driven approach, it identifies many potential gene-metabolite relationships that could be associated with AD risk factors, and shows that most plasma and cerebral spinal fluid (CSF) metabolite levels are weakly correlated with each other. This information could be used to guide further investigations of specific biological mechanisms that could lead to the development of AD, and ultimately, could be used to improve predictive models of this devastating disease.

While this work is geared towards better understanding AD risk, having knowledge of how aging and sex influence metabolites has implications for diseases of aging beyond AD. Using a large panel of longitudinal metabolomics data, we conducted a comprehensive investigation of the influence of aging and sex on metabolomics. Our findings suggest that levels of most metabolites are highly influenced by sex and age, and that sex differentially influences levels and trajectories of many metabolites. Many metabolites that we identified to be associated with age and sex were previously reported and had the same effect directions (Krumsiek et al., 2015; Menni et al., 2013), which strengthens the replicated and novel findings of our investigations. However, our

investigation provides important advances to what was previously known on this subject. First, our results are based on a large number of longitudinal samples, which facilitates the examination of metabolite trajectories and is able to capture age-related phenomena that could introduce variation to metabolite levels. Second, our study utilizes a larger panel of metabolites than previously investigated, which enabled us to identify a broader set of metabolites that are influenced by age and sex. Third, we offer a more comprehensive investigation than previously reported by assessing the influences of both age and sex on metabolite levels, as well as metabolite trajectories that differ by sex.

Our findings underscore the importance of incorporating age and sex in the design and analysis of metabolomics investigations and offer a deeper understanding of the aging process. These results could inform many novel hypotheses regarding the role of metabolites in healthy and accelerated aging. For instance, one could identify a profile of metabolite trajectories for participants who go on to develop AD and compare it to trajectories identified here. These results may also have clinical utility, as they could be used to inform predictive models of accelerated aging. We also report that many metabolite levels are influenced by a complex combination of both genomic and environmental influences. This information could help us to understand how much metabolite variation can be controlled by non-genetic factors and could guide the use of metabolites to identify gene-environment interactions in future studies. However, our heritability estimates were not highly correlated with previous estimates (Long et al., 2017; Shin et al., 2014), which also were not highly correlated with each other, despite some participants being common to both studies. This is likely a reflection of the high

variability of metabolites and suggests that these estimates may vary widely between different populations.

Using the same panel of longitudinal metabolomics data, we also found that certain metabolites, including cysteine S-sulfate, which is an amino acid, and erucate, which is an omega-9 fatty acid, were associated with age-related change in executive function, one of the earliest aspects of cognitive function to change in the course of AD development. Results from this investigation offer novel insights into cognitive function and cognitive decline, as we have not identified previous metabolomics studies of cognitive function. However, fatty acids, particularly of omega-3 fatty acids, have long been suspected to influence cognitive performance, although studies have had mixed findings about their involvement (Cederholm et al., 2013; Mazereeuw et al., 2012). Our results suggest that it may be difficult to define relationships between particular fatty acids and cognitive function because they may change as individuals age. Specifically, while investigating how metabolites change with age, we found that the majority of fatty acids tested (82/126) significantly increased with age; however, our cognitive findings suggest that higher levels of erucate in midlife and lower levels later in life (*i.e.*, levels that decrease with age) were associated with better executive function. This is further supported by consistent relationships seen with two other partially characterized fatty acids (fatty acid 8:1 acyl glutamine conjugate and fatty acid 6:1 acyl glutamine conjugate). More information about these two particular metabolites could prove useful in understanding the relationship between fatty acids and cognitive function. Beyond cognitive performance, omega fatty acids have also been shown to be dysregulated in certain brain regions of patients with AD pathology (Snowden et al., 2017), further

strengthening the potential relevance of fatty acids. As more individuals in the WRAP cohort develop cognitive impairment, it will be better powered to assess the relationship between metabolite levels and cognitive function. As such, future assessments could repeat the analyses conducted here to assess the validity of our findings.

We also developed an integrative network to investigate relationships between plasma metabolomics, CSF metabolomics, genomics, and AD risk factors. Although no gene expression levels were directly correlated with AD risk factors, there were many instances of gene expression being indirectly correlated with AD risk factors through metabolites. One such relationship was further investigated using mediation and moderation analyses. We found that glycine mediated the relationship between whole blood *CPS1* expression and body mass index (BMI), waist-hip ratio (WHR), and insulin resistance, such that higher *CPS1* expression was associated with lower plasma glycine levels, which was associated with higher BMI, WHR, and insulin resistance. Without considering glycine, the relationships between *CPS1* and BMI, WHR, and insulin resistance were undetectable. This suggests that our results may have generated many valid hypotheses that warrant further investigation. As such, these hypotheses could motivate many future research ideas.

Network analyses also revealed that higher CSF total tau (t-tau) and phosphorylated tau (p-tau) levels were correlated with higher levels of 22 CSF metabolites, collectively, which included five lipids (four phosphatidylcholines and one lysophosphatidylcholine), five amino acids, four carbohydrates, one nucleotide, one energy metabolite, and six unknown metabolites. However, CSF amyloid biomarkers were not correlated with any CSF metabolites, which could have implications about the

biological function of amyloid versus tau. For example, the amyloid hypothesis suggests that the accumulation of amyloid occurs decades prior to the diagnosis of AD and leads to downstream events, including the accumulation of tau (Jack et al., 2010). The broader influence of tau that we observed could be in agreement with this timeline, if our mid-to-late life cohort is transitioning to the tau accumulation stage, and could explain the observed metabolic associations with tau. It is also possible that we did not capture small molecules that amyloid may be associated with, or that amyloid is generally not associated with small molecules. The latter hypothesis is supported by a previous investigation of amyloid burden (as measured by PET Pittsburgh compound B) and plasma metabolites, which did not identify any metabolites that were independently associated with amyloid burden (Voyle et al., 2016). Although our CSF findings were limited by their small sample size, they offer potentially novel information regarding the interface between CSF biomarkers and CSF metabolites, as we have not identified previous studies investigating these relationships. Future CSF metabolomics investigations utilizing larger sample sizes could offer important information about the potential metabolic implications of neural amyloid and tau accumulations during AD progression.

We also found that correlations between plasma and CSF metabolites ranged widely, but the low average correlation suggested that plasma metabolites typically may not be representative of certain metabolic changes occurring in the brain. Plasma and CSF xenobiotics had the strongest correlations, while lipids had the weakest correlations. However, it is important to note that metabolites within a given pathway can vary widely from each other. As such, metabolites should be considered on an

individual basis, as the averages presented here may not reflect a particular metabolite's unique properties. The low correlation observed between plasma and CSF metabolite levels could be related to ~98% of small molecules not being able to pass the blood-brain barrier (BBB) (Pardridge, 2005). Cholesterol is an example of a lipid that typically cannot pass the BBB (Bjorkhem & Meaney, 2004), and was not correlated between tissues (r=-0.07). On the other hand, caffeine readily crosses the BBB (McCall et al., 1982) and it was highly correlated between tissues (r=0.81), as was 5-acetylamino-6-amino-3-methyluracil (r=0.82), which is a caffeine metabolite, and theophylline (r=0.82), which is structurally and pharmacologically similar to caffeine. This hypothesis is further supported by the observed correlations between CSF biomarkers and CSF metabolites, but absence of correlations between CSF biomarkers and plasma metabolites.

By integrating genomics, metabolomics, and clinical risk factors for AD, we were able to identify potentially complex relationships that could lead to a better understanding of AD and risk factors associated with its onset. It is becoming more common for cohorts to collect such datasets—to this effect, the National Institutes of Health is sponsoring the new Trans-Omics for Precision Medicine (TOPMed) nation-wide consortium that aims to deeply phenotype its participants utilizing omics technologies (www.nhlbiwgs.org)—and these data are expected to greatly advance our knowledge of many complex diseases and traits. To fully utilize these rich data, it will be crucial to identify effective means of integrating them in order to maximize their potential to provide a more holistic understanding of the disease process. While there is still a great need for such methods, the process that we have outlined here, which involved

building an integrated network based on correlations between inter-omic data and further exploring the results with mediation and interaction analyses, is one effective way to integrate omics data.

Although genome-wide association studies (GWAS) have identified thousands of single nucleotide polymorphism (SNP)-trait associations (MacArthur et al., 2017), these variants tend to have very small effect sizes and typically explain a small portion of trait heritability. Alzheimer's disease (AD) is an example of such a trait: 53% of its phenotypic variance can be explained by genomic variants, collectively (*i.e.*, SNP heritability); yet, the 21 GWAS variants known to be associated with AD only account for 31% of its genetic variance, leaving 69% unaccounted for (Ridge et al., 2016). In order to more comprehensively understand the disease risk conveyed by genetic factors, it is crucial to consider genomics in combination with other omics data types and to use integrative multi-omics approaches that can capture intricate relationships.

Although there has been great interest recently in the integration of multi-omics datasets, progress in this field has been fairly limited and faces many challenges, most notably including the lack of appropriate statistical methods and the computation demands of such data (Bersanelli et al., 2016; Buescher & Driggers, 2016; Gligorijevic & Przulj, 2015; Huang et al., 2017; Lopez de Maturana et al., 2016; Ritchie et al., 2015). Despite these limitations, studies have been able to show that the use of multiple omics data types is more predictive than single data types (Buescher & Driggers, 2016; Mankoo et al., 2011). The integrative approach used by our study was similar to that of a recent study that integrated dense longitudinal omics data to assess various outcomes (Price et al., 2017). Although limited by its sample size of 108 participants,

this investigation identified meaningful systems biology relationships that were able to improve the health of its participants, displaying the utility of integrating such data with regards to personalized medicine. For example, they were able to use clinical laboratory and genomic testing results to identify participants who could be at risk for hemochromatosis and refer them to the appropriate health care specialist. Our study considerably adds to the design of this previous investigation, as it is based on a much larger sample size, utilizes a larger panel of metabolites, and uses genomics in a manner that could be more effective (the previous study focused on polygenic scores for 127 traits and diseases, whereas we focused on imputed gene expression data for ~11,000 genes). As it is becoming more feasible and common to acquire multiple omics data types, it is essential that we move towards system biology approaches of understanding complex diseases, rather than focusing on single data types that are unable to capture the intricacies imposed by biology.

Opponents of the “big data” era have criticized omics approaches because they are not hypothesis driven and do not follow the standard scientific method (Duncan, 2007; Stieb et al., 2017). However, we know biology to be complex far beyond our current understanding. To believe that we currently have the ability to generate valid biological hypotheses to understand complex conditions without data would be a fallacy. This was a lesson learned in the years preceding the completion of the human genome sequence in 2001 when research efforts were heavily invested into targeted genetic loci and genome-wide linkage screens of ~500 loci (Fallin et al., 2016). While this approach was successful for genes that follow Mendelian patterns, such as highly penetrant variants in the *BRCA1* and *BRCA2* genes that are responsible for inherited forms of

breast cancer and the *APP*, *PSEN1*, and *PSEN2* genes that cause the inherited early onset form of AD, it had limited success for traits that follow complex inheritance patterns (Fallin et al., 2016). The utility of omics data, and particularly integrated omics approaches, is the ability to generate data driven hypotheses. Our knowledge of biology has been evolving for centuries; however, with the data we are able to generate due to recent biotechnological advances, we now have the opportunity to advance our knowledge of biology at an unprecedented rate. Such data could lead to dramatic improvements in the state of preventative and therapeutic medicine, particularly for complex diseases such as AD, for which few such preventative or therapeutic methods exist and little is known about the underlying biological mechanisms.

Bibliography

. (2016) *Use of Metabolomics to Advance Research on Environmental Exposures and the Human Exposome: Workshop in Brief*. Washington (DC).

Albert, M. S. (1996). Cognitive and neurobiologic markers of early Alzheimer disease. *Proc Natl Acad Sci U S A, 93*(24), 13547-13551.

Astle, W. J., Elding, H., Jiang, T., Allen, D., Ruklisa, D., Mann, A. L., . . . Soranzo, N. (2016). The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell, 167*(5), 1415-1429 e1419. doi:10.1016/j.cell.2016.10.042

Auro, K., Joensuu, A., Fischer, K., Kettunen, J., Salo, P., Mattsson, H., . . . Perola, M. (2014). A metabolic view on menopause and ageing. *Nat Commun, 5*, 4708. doi:10.1038/ncomms5708

Bates, D., Machler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1-48.

Battle, A., Mostafavi, S., Zhu, X., Potash, J. B., Weissman, M. M., McCormick, C., . . . Koller, D. (2014). Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res, 24*(1), 14-24. doi:10.1101/gr.155192.113

Baudic, S., Barba, G. D., Thibaudet, M. C., Smagghe, A., Remy, P., & Traykov, L. (2006). Executive function deficits in early Alzheimer's disease and their relations with episodic memory. *Arch Clin Neuropsychol, 21*(1), 15-21. doi:10.1016/j.acn.2005.07.002

Benedict, R. H. (1997). Brief Visuospatial Memory Test-Revised. Odessa, FL: Psychological Assessment Resources, Inc.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B-Methodological, 57*(1), 289-300.

Bersanelli, M., Mosca, E., Remondini, D., Giampieri, E., Sala, C., Castellani, G., & Milanesi, L. (2016). Methods for the integration of multi-omics data: mathematical aspects. *BMC Bioinformatics, 17 Suppl 2*, 15. doi:10.1186/s12859-015-0857-9

Bjorkhem, I., & Meaney, S. (2004). Brain cholesterol: long secret life behind a barrier. *Arterioscler Thromb Vasc Biol, 24*(5), 806-815. doi:10.1161/01.ATV.0000120374.59826.1b

Blennow, K. (2004). Cerebrospinal fluid protein biomarkers for Alzheimer's disease. *NeuroRx, 1*(2), 213-225. doi:10.1602/neurorx.1.2.213

Buescher, J. M., & Driggers, E. M. (2016). Integration of omics: more than the sum of its parts. *Cancer Metab, 4*, 4. doi:10.1186/s40170-016-0143-y

Cederholm, T., Salem, N., Jr., & Palmblad, J. (2013). omega-3 fatty acids in the prevention of cognitive decline in humans. *Adv Nutr, 4*(6), 672-676. doi:10.3945/an.113.004556

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience, 4*, 7. doi:10.1186/s13742-015-0047-8

Cheng, S., Rhee, E. P., Larson, M. G., Lewis, G. D., McCabe, E. L., Shen, D., . . . Wang, T. J. (2012). Metabolite profiling identifies pathways associated with

metabolic risk in humans. *Circulation, 125*(18), 2222-2231. doi:10.1161/CIRCULATIONAHA.111.067827

Clark, L. R., Racine, A. M., Koscik, R. L., Okonkwo, O. C., Engelman, C. D., Carlsson, C. M., . . . Johnson, S. C. (2016). Beta-amyloid and cognitive decline in late middle age: Findings from the Wisconsin Registry for Alzheimer's Prevention study. *Alzheimers Dement, 12*(7), 805-814. doi:10.1016/j.jalz.2015.12.009

Clark, L. R., Schiehser, D. M., Weissberger, G. H., Salmon, D. P., Delis, D. C., & Bondi, M. W. (2012). Specific measures of executive function predict cognitive decline in older adults. *J Int Neuropsychol Soc, 18*(1), 118-127. doi:10.1017/S1355617711001524

Clauset, A., Newman, M. E. J., & Moore, C. (2004). Finding community structure in very large networks. *Physical Review E, 70*(6). doi:Artn 066111 10.1103/Physreve.70.066111

Conomos, M. P., Miller, M. B., & Thornton, T. A. (2015). Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. *Genet Epidemiol, 39*(4), 276-293. doi:10.1002/gepi.21896

Conomos, M. P., Reiner, A. P., Weir, B. S., & Thornton, T. A. (2016). Model-free Estimation of Recent Genetic Relatedness. *Am J Hum Genet, 98*(1), 127-148. doi:10.1016/j.ajhg.2015.11.022

Cruchaga, C., Karch, C. M., Jin, S. C., Benitez, B. A., Cai, Y., Guerreiro, R., . . . Goate, A. M. (2013). Rare coding variants in the phospholipase D3 gene confer risk for Alzheimer's disease. *Nature*. doi:10.1038/nature12825

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., . . . Genomes Project Analysis, G. (2011). The variant call format and VCFtools. *Bioinformatics, 27*(15), 2156-2158. doi:10.1093/bioinformatics/btr330

Danik, J. S., Pare, G., Chasman, D. I., Zee, R. Y., Kwiatkowski, D. J., Parker, A., . . . Ridker, P. M. (2009). Novel loci, including those related to Crohn disease, psoriasis, and inflammation, identified in a genome-wide association study of fibrinogen in 17 686 women: the Women's Genome Health Study. *Circ Cardiovasc Genet, 2*(2), 134-141. doi:10.1161/CIRCGENETICS.108.825273

Darst, B. F., & Engelman, C. D. (2016). Transmission and decorrelation methods for detecting rare variants using sequencing data from related individuals. *BMC Proc, 10*(Suppl 7), 203-207. doi:10.1186/s12919-016-0031-z

Darst, B. F., Koscik, R. L., Hermann, B. P., La Rue, A., Sager, M. A., Johnson, S. C., & Engelman, C. D. (2015). Heritability of Cognitive Traits Among Siblings with a Parental History of Alzheimer's Disease. *J Alzheimers Dis*. doi:10.3233/JAD-142658

Darst, B. F., Koscik, R. L., Racine, A. M., Oh, J. M., Krause, R. A., Carlsson, C. M., . . . Engelman, C. D. (2017). Pathway-Specific Polygenic Risk Scores as Predictors of Amyloid-beta Deposition and Cognitive Function in a Sample at Increased Risk for Alzheimer's Disease. *J Alzheimers Dis, 55*(2), 473-484. doi:10.3233/JAD-160195

Das, S., Forer, L., Schonherr, S., Sidore, C., Locke, A. E., Kwong, A., . . . Fuchsberger, C. (2016). Next-generation genotype imputation service and methods. *Nat Genet, 48*(10), 1284-1287. doi:10.1038/ng.3656

Davison, S. L., Bell, R., Donath, S., Montalto, J. G., & Davis, S. R. (2005). Androgen levels in adult females: changes with age, menopause, and oophorectomy. *J Clin Endocrinol Metab, 90*(7), 3847-3853. doi:10.1210/jc.2005-0212

De Meyer, G., Shapiro, F., Vanderstichele, H., Vanmechelen, E., Engelborghs, S., De Deyn, P. P., . . . Trojanowski, J. Q. (2010). Diagnosis-independent Alzheimer disease biomarker signature in cognitively normal elderly people. *Archives of Neurology, 67*(8), 949-956. doi:10.1001/archneurol.2010.179

de Vries, P. S., Chasman, D. I., Sabater-Lleal, M., Chen, M. H., Huffman, J. E., Steri, M., . . . Dehghan, A. (2016). A meta-analysis of 120 246 individuals identifies 18 new loci for fibrinogen concentration. *Hum Mol Genet, 25*(2), 358-370. doi:10.1093/hmg/ddv454

Deidda, M., Piras, C., Bassareo, P. P., Dessalvi, C. C., & Mercuro, G. (2015). Metabolomics, a promising approach to translational research in cardiology. *Ijc Metabolic & Endocrine, 9*, 31-38. doi:10.1016/j.ijcme.2015.10.001

Demirkan, A., Henneman, P., Verhoeven, A., Dharuri, H., Amin, N., van Klinken, J. B., . . . van Dijk, K. W. (2015). Insight in genome-wide association of metabolite quantitative traits by exome sequence analyses. *PLoS Genet, 11*(1), e1004835. doi:10.1371/journal.pgen.1004835

Ding, Y., Svingen, G. F., Pedersen, E. R., Gregory, J. F., Ueland, P. M., Tell, G. S., & Nygard, O. K. (2015). Plasma Glycine and Risk of Acute Myocardial Infarction in

Patients With Suspected Stable Angina Pectoris. *J Am Heart Assoc, 5*(1). doi:10.1161/JAHA.115.002621

Draisma, H. H. M., Pool, R., Kobl, M., Jansen, R., Petersen, A. K., Vaarhorst, A. A. M., . . . Boomsma, D. I. (2015). Genome-wide association study identifies novel genetic variants contributing to variation in blood metabolite levels. *Nat Commun, 6*, 7208. doi:10.1038/ncomms8208

Duits, F. H., Teunissen, C. E., Bouwman, F. H., Visser, P. J., Mattsson, N., Zetterberg, H., . . . van der Flier, W. M. (2014). The cerebrospinal fluid "Alzheimer profile": Easily said, but what does it mean? *Alzheimer's & dementia : the journal of the Alzheimer's Association*. doi:10.1016/j.jalz.2013.12.023

Duncan, M. W. (2007). Omics and its 15 minutes. *Exp Biol Med (Maywood), 232*(4), 471-472.

Dunn, W. B., Lin, W., Broadhurst, D., Begley, P., Brown, M., Zelena, E., . . . Kell, D. B. (2015). Molecular phenotyping of a UK population: defining the human serum metabolome. *Metabolomics, 11*, 9-26. doi:10.1007/s11306-014-0707-1

Edwards, L. J., Muller, K. E., Wolfinger, R. D., Qaqish, B. F., & Schabenberger, O. (2008). An R2 statistic for fixed effects in the linear mixed model. *Stat Med, 27*(29), 6137-6157. doi:10.1002/sim.3429

El Hafidi, M., Perez, I., & Banos, G. (2006). Is glycine effective against elevated blood pressure? *Curr Opin Clin Nutr Metab Care, 9*(1), 26-31.

Enche Ady, C. N. A., Lim, S. M., Teh, L. K., Salleh, M. Z., Chin, A. V., Tan, M. P., . . . Ramasamy, K. (2017). Metabolomic-guided discovery of Alzheimer's disease

biomarkers from body fluid. *J Neurosci Res, 95*(10), 2005-2024. doi:10.1002/jnr.24048

Engelman, C. D., Darst, B. F., Bilgel, M., Vasiljevic, E., Koscik, R. L., Jedynak, B. M., & Johnson, S. C. (2017). The effect of rare variants in TREM2 and PLD3 on longitudinal cognitive function in the Wisconsin Registry for Alzheimer's Prevention. *Neurobiol Aging*. doi:10.1016/j.neurobiolaging.2017.12.025

Evans, A. M., Bridgewater, B. R., Liu, Q., Mitchell, M. W., Robinson, R. J., Dai, H., . . . Miller, L. A. D. (2014). High Resolution Mass Spectrometry Improves Data Quantity and Quality as Compared to Unit Mass Resolution Mass Spectrometry in High-Throughput Profiling Metabolomics. *Metabolomics, 4*(1), 1-7. doi:10.4172/2153-0769.1000132

Fallin, M. D., Duggal, P., & Beaty, T. H. (2016). Genetic Epidemiology and Public Health: The Evolution From Theory to Technology. *Am J Epidemiol, 183*(5), 387-393. doi:10.1093/aje/kww001

Felig, P., Marliss, E., & Cahill, G. F., Jr. (1969). Plasma amino acid levels and insulin secretion in obesity. *N Engl J Med, 281*(15), 811-816. doi:10.1056/NEJM196910092811503

Ferrini, R. L., & Barrett-Connor, E. (1998). Sex hormones and age: a cross-sectional study of testosterone and estradiol and their bioavailable fractions in community-dwelling men. *Am J Epidemiol, 147*(8), 750-754.

Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., . . . Im, H. K. (2015). A gene-based association method for

mapping traits using reference transcriptome data. *Nat Genet, 47*(9), 1091-1098. doi:10.1038/ng.3367

Geidenstam, N., Magnusson, M., Danielsson, A. P. H., Gerszten, R. E., Wang, T. J., Reinius, L. E., . . . Ridderstrale, M. (2017). Amino Acid Signatures to Evaluate the Beneficial Effects of Weight Loss. *Int J Endocrinol, 2017*, 6490473. doi:10.1155/2017/6490473

Genomes Project, C., Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., . . . Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature, 526*(7571), 68-74. doi:10.1038/nature15393

Gligorijevic, V., & Przulj, N. (2015). Methods for biological data integration: perspectives and challenges. *J R Soc Interface, 12*(112). doi:10.1098/rsif.2015.0571

Golovko, M. Y., & Murphy, E. J. (2006). Uptake and metabolism of plasma-derived erucic acid by rat brain. *J Lipid Res, 47*(6), 1289-1297. doi:10.1194/jlr.M600029-JLR200

Gonzalez-Covarrubias, V., Beekman, M., Uh, H. W., Dane, A., Troost, J., Paliukhovich, I., . . . Slagboom, E. P. (2013). Lipidomics of familial longevity. *Aging Cell, 12*(3), 426-434. doi:10.1111/acel.12064

Gonzalez-Dominguez, R., Sayago, A., & Fernandez-Recamales, A. (2017). Metabolomics in Alzheimer's disease: The need of complementary analytical platforms for the identification of biomarkers to unravel the underlying pathology. *J Chromatogr B Analyt Technol Biomed Life Sci*. doi:10.1016/j.jchromb.2017.02.008

Goodman-Gruen, D., & Barrett-Connor, E. (2000). Sex differences in the association of endogenous sex hormone levels and glucose tolerance status in older men and women. *Diabetes Care, 23*(7), 912-918.

Gorski, M., van der Most, P. J., Teumer, A., Chu, A. Y., Li, M., Mijatovic, V., . . . Fuchsberger, C. (2017). 1000 Genomes-based meta-analysis identifies 10 novel loci for kidney function. *Sci Rep, 7*, 45040. doi:10.1038/srep45040

Griffin, J. W., & Bradshaw, P. C. (2017). Amino Acid Catabolism in Alzheimer's Disease Brain: Friend or Foe? *Oxid Med Cell Longev, 2017*, 5472792. doi:10.1155/2017/5472792

Haberle, J., Shchelochkov, O. A., Wang, J., Katsonis, P., Hall, L., Reiss, S., . . . Summar, M. (2011). Molecular defects in human carbamoy phosphate synthetase I: mutational spectrum, diagnostic and protein structure considerations. *Hum Mutat, 32*(6), 579-589. doi:10.1002/humu.21406

Hansmannel, F., Sillaire, A., Kamboh, M. I., Lendon, C., Pasquier, F., Hannequin, D., . . . Lambert, J. C. (2010). Is the urea cycle involved in Alzheimer's disease? *J Alzheimers Dis, 21*(3), 1013-1021. doi:10.3233/JAD-2010-100630

Harman, S. M., Metter, E. J., Tobin, J. D., Pearson, J., Blackman, M. R., & Baltimore Longitudinal Study of, A. (2001). Longitudinal effects of aging on serum total and free testosterone levels in healthy men. Baltimore Longitudinal Study of Aging. *J Clin Endocrinol Metab, 86*(2), 724-731. doi:10.1210/jcem.86.2.7219

Hartiala, J. A., Tang, W. H., Wang, Z., Crow, A. L., Stewart, A. F., Roberts, R., . . . Allayee, H. (2016). Genome-wide association study and targeted metabolomics

identifies sex-specific association of CPS1 with coronary artery disease. *Nat Commun, 7*, 10558. doi:10.1038/ncomms10558

Horgan, R. P., & Kenny, L. C. (2011). 'Omic' technologies: genomics, transcriptomics, proteomics and metabolomics. *The Obstetrician & Gynaecologist, 13*, 189-195. doi:10.1576/toag.13.3.189.27672

Huang, S., Chaudhary, K., & Garmire, L. X. (2017). More Is Better: Recent Progress in Multi-Omics Data Integration Methods. *Front Genet, 8*, 84. doi:10.3389/fgene.2017.00084

Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychol Methods, 15*(4), 309-334. doi:10.1037/a0020761

Jack, C. R., Jr., Knopman, D. S., Jagust, W. J., Shaw, L. M., Aisen, P. S., Weiner, M. W., . . . Trojanowski, J. Q. (2010). Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *Lancet Neurol, 9*(1), 119-128. doi:10.1016/S1474-4422(09)70299-6

Jack, C. R., Jr., Lowe, V. J., Weigand, S. D., Wiste, H. J., Senjem, M. L., Knopman, D. S., . . . Alzheimer's Disease Neuroimaging, I. (2009). Serial PIB and MRI in normal, mild cognitive impairment and Alzheimer's disease: implications for sequence of pathological events in Alzheimer's disease. *Brain, 132*(Pt 5), 1355-1365. doi:10.1093/brain/awp062

Johnson, S. C., Christian, B. T., Okonkwo, O. C., Oh, J. M., Harding, S., Xu, G., . . . Sager, M. A. (2014). Amyloid burden and neural function in people at risk for Alzheimer's Disease. *Neurobiology of aging, 35*(3), 576-584. doi:10.1016/j.neurobiolaging.2013.09.028

Johnson, S. C., Koscik, R. L., Jonaitis, E. M., Clark, L. R., Mueller, K. D., Berman, S. E., . . . Sager, M. A. (2018). The Wisconsin Registry for Alzheimer's Prevention: A review of findings and current directions. *Alzheimers Dement (Amst), 10*, 130-142. doi:10.1016/j.dadm.2017.11.007

Jun, G., Ibrahim-Verbaas, C. A., Vronskaya, M., Lambert, J. C., Chung, J., Naj, A. C., . . . Farrer, L. A. (2016). A novel Alzheimer disease locus located near the gene encoding tau protein. *Mol Psychiatry, 21*(1), 108-117. doi:10.1038/mp.2015.23

Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res, 28*(1), 27-30.

Ke, C., Hou, Y., Zhang, H., Yang, K., Wang, J., Guo, B., . . . Li, K. (2015). Plasma Metabolic Profiles in Women are Menopause Dependent. *PLoS One, 10*(11), e0141743. doi:10.1371/journal.pone.0141743

Kettunen, J., Demirkan, A., Wurtz, P., Draisma, H. H., Haller, T., Rawal, R., . . . Ala-Korpela, M. (2016). Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. *Nat Commun, 7*, 11122. doi:10.1038/ncomms11122

Kim, E., Ko, H. J., Jeon, S. J., Lee, S., Lee, H. E., Kim, H. N., . . . Ryu, J. H. (2016). The memory-enhancing effect of erucic acid on scopolamine-induced cognitive impairment in mice. *Pharmacol Biochem Behav, 142*, 85-90. doi:10.1016/j.pbb.2016.01.006

Kottgen, A., Pattaro, C., Boger, C. A., Fuchsberger, C., Olden, M., Glazer, N. L., . . . Fox, C. S. (2010). New loci associated with kidney function and chronic kidney disease. *Nat Genet, 42*(5), 376-384. doi:10.1038/ng.568

Kraus, W. E., Pieper, C. F., Huffman, K. M., Thompson, D. K., Kraus, V. B., Morey, M. C., . . . Newgard, C. B. (2016). Association of Plasma Small-Molecule Intermediate Metabolites With Age and Body Mass Index Across Six Diverse Study Populations. *J Gerontol A Biol Sci Med Sci, 71*(11), 1507-1513. doi:10.1093/gerona/glw031

Krumsiek, J., Mittelstrass, K., Do, K. T., Stuckler, F., Ried, J., Adamski, J., . . . Kastenmuller, G. (2015). Gender-specific pathway differences in the human serum metabolome. *Metabolomics, 11*(6), 1815-1833. doi:10.1007/s11306-015-0829-0

Kumar, A., Dejanovic, B., Hetsch, F., Semtner, M., Fusca, D., Arjune, S., . . . Belaidi, A. A. (2017). S-sulfocysteine/NMDA receptor-dependent signaling underlies neurodegeneration in molybdenum cofactor deficiency. *J Clin Invest, 127*(12), 4365-4378. doi:10.1172/JCI89885

Lafleche, G., & Albert, M. S. (1995). Executive function deficits in mild Alzheimer's disease. *Neuropsychology, 9*(3), 313-320. doi:10.1037/0894-4105.9.3.313

Lange, L. A., Croteau-Chonka, D. C., Marvelle, A. F., Qin, L., Gaulton, K. J., Kuzawa, C. W., . . . Mohlke, K. L. (2010). Genome-wide association study of homocysteine levels in Filipinos provides evidence for CPS1 in women and a stronger MTHFR effect in young adults. *Hum Mol Genet, 19*(10), 2050-2058. doi:10.1093/hmg/ddq062

Lewczuk, P., Lelental, N., Spitzer, P., Maler, J. M., & Kornhuber, J. (2015). Amyloid-beta 42/40 cerebrospinal fluid concentration ratio in the diagnostics of

Alzheimer's disease: validation of two novel assays. *Journal of Alzheimer's disease : JAD, 43*(1), 183-191. doi:10.3233/JAD-140771

Locke, A. E., Kahali, B., Berndt, S. I., Justice, A. E., Pers, T. H., Day, F. R., . . . Speliotes, E. K. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature, 518*(7538), 197-206. doi:10.1038/nature14177

Loh, P. R., Danecek, P., Palamara, P. F., Fuchsberger, C., Y, A. R., H, K. F., . . . A, L. P. (2016). Reference-based phasing using the Haplotype Reference Consortium panel. *Nat Genet, 48*(11), 1443-1448. doi:10.1038/ng.3679

Long, T., Hicks, M., Yu, H. C., Biggs, W. H., Kirkness, E. F., Menni, C., . . . Telenti, A. (2017). Whole-genome sequencing identifies common-to-rare variants associated with human blood metabolites. *Nat Genet, 49*(4), 568-578. doi:10.1038/ng.3809

Lopez de Maturana, E., Pineda, S., Brand, A., Van Steen, K., & Malats, N. (2016). Toward the integration of Omics data in epidemiological studies: still a "long and winding road". *Genet Epidemiol, 40*(7), 558-569. doi:10.1002/gepi.21992

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., . . . Parkinson, H. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res, 45*(D1), D896-D901. doi:10.1093/nar/gkw1133

Mahajan, A., Rodan, A. R., Le, T. H., Gaulton, K. J., Haessler, J., Stilp, A. M., . . . Franceschini, N. (2016). Trans-ethnic Fine Mapping Highlights Kidney-Function Genes Linked to Salt Sensitivity. *Am J Hum Genet, 99*(3), 636-646. doi:10.1016/j.ajhg.2016.07.012

Makinen, V. P., & Ala-Korpela, M. (2016). Metabolomics of aging requires large-scale longitudinal studies with replication. *Proc Natl Acad Sci U S A, 113*(25), E3470. doi:10.1073/pnas.1607062113

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., & Chen, W. M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics, 26*(22), 2867-2873. doi:10.1093/bioinformatics/btq559

Mankoo, P. K., Shen, R., Schultz, N., Levine, D. A., & Sander, C. (2011). Time to recurrence and survival in serous ovarian tumors predicted from integrated genomic profiles. *PLoS One, 6*(11), e24709. doi:10.1371/journal.pone.0024709

Matone, A., Scott-Boyer, M. P., Carayol, J., Fazelzadeh, P., Lefebvre, G., Valsesia, A., . . . Hager, J. (2016). Network Analysis of Metabolite GWAS Hits: Implication of CPS1 and the Urea Cycle in Weight Maintenance. *PLoS One, 11*(3), e0150495. doi:10.1371/journal.pone.0150495

Mazereeuw, G., Lanctot, K. L., Chau, S. A., Swardfager, W., & Herrmann, N. (2012). Effects of omega-3 fatty acids on cognitive performance: a meta-analysis. *Neurobiol Aging, 33*(7), 1482 e1417-1429. doi:10.1016/j.neurobiolaging.2011.12.014

McCall, A. L., Millington, W. R., & Wurtman, R. J. (1982). Blood-brain barrier transport of caffeine: dose-related restriction of adenine transport. *Life Sci, 31*(24), 2709-2715.

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., . . . Haplotype Reference, C. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet, 48*(10), 1279-1283. doi:10.1038/ng.3643

Melen, E., Granell, R., Kogevinas, M., Strachan, D., Gonzalez, J. R., Wjst, M., . . . Lasky-Su, J. (2013). Genome-wide association study of body mass index in 23 000 individuals with and without asthma. *Clin Exp Allergy, 43*(4), 463-474. doi:10.1111/cea.12054

Menni, C., Kastenmuller, G., Petersen, A. K., Bell, J. T., Psatha, M., Tsai, P. C., . . . Valdes, A. M. (2013). Metabolomic markers reveal novel pathways of ageing and early development in human populations. *Int J Epidemiol, 42*(4), 1111-1119. doi:10.1093/ije/dyt094

Mielke, M. M., Bandaru, V. V., Han, D., An, Y., Resnick, S. M., Ferrucci, L., & Haughey, N. J. (2015). Factors affecting longitudinal trajectories of plasma sphingomyelins: the Baltimore Longitudinal Study of Aging. *Aging Cell, 14*(1), 112-121. doi:10.1111/acel.12275

Mittelstrass, K., Ried, J. S., Yu, Z., Krumsiek, J., Gieger, C., Prehn, C., . . . Illig, T. (2011). Discovery of sexual dimorphisms in metabolic and genetic biomarkers. *PLoS Genet, 7*(8), e1002215. doi:10.1371/journal.pgen.1002215

Niccoli, T., & Partridge, L. (2012). Ageing as a risk factor for disease. *Curr Biol, 22*(17), R741-752. doi:10.1016/j.cub.2012.07.024

Noyan, V., Yucel, A., & Sagsoz, N. (2004). The association of androgenic sex steroids with serum lipid levels in postmenopausal women. *Acta Obstet Gynecol Scand, 83*(5), 487-490. doi:10.1111/j.0001-6349.2004.00417.x

Olney, J. W., Misra, C. H., & de Gubareff, T. (1975). Cysteine-S-sulfate: brain damaging metabolite in sulfite oxidase deficiency. *J Neuropathol Exp Neurol, 34*(2), 167-177.

Pardridge, W. M. (2005). The blood-brain barrier: bottleneck in brain drug development. *NeuroRx, 2*(1), 3-14. doi:10.1602/neurorx.2.1.3

Pare, G., Chasman, D. I., Parker, A. N., Zee, R. R., Malarstig, A., Seedorf, U., . . . Ridker, P. M. (2009). Novel associations of CPS1, MUT, NOX4, and DPEP1 with plasma homocysteine in a healthy population: a genome-wide evaluation of 13 974 participants in the Women's Genome Health Study. *Circ Cardiovasc Genet, 2*(2), 142-150. doi:10.1161/CIRCGENETICS.108.829804

Pattaro, C., Teumer, A., Gorski, M., Chu, A. Y., Li, M., Mijatovic, V., . . . Fox, C. S. (2016). Genetic associations at 53 loci highlight cell types and biological pathways relevant for kidney function. *Nat Commun, 7*, 10023. doi:10.1038/ncomms10023

Pearson, D. L., Dawling, S., Walsh, W. F., Haines, J. L., Christman, B. W., Bazyk, A., . . . Summar, M. L. (2001). Neonatal pulmonary hypertension--urea-cycle intermediates, nitric oxide production, and carbamoyl-phosphate synthetase function. *N Engl J Med, 344*(24), 1832-1838. doi:10.1056/NEJM200106143442404

Price, N. D., Magis, A. T., Earls, J. C., Glusman, G., Levy, R., Lausted, C., . . . Hood, L. (2017). A wellness study of 108 individuals using personal, dense, dynamic data clouds. *Nature Biotechnology, 35*(8), 747-756. doi:10.1038/nbt.3870

Raffler, J., Friedrich, N., Arnold, M., Kacprowski, T., Rueedi, R., Altmaier, E., . . . Suhre, K. (2015). Genome-Wide Association Study with Targeted and Non-targeted NMR Metabolomics Identifies 15 Novel Loci of Urinary Human Metabolic Individuality. *PLoS Genet, 11*(9), e1005487. doi:10.1371/journal.pgen.1005487

Rayner, N. W., Robertson, N., Mahajan, A., & McCarthy, M. I. (2016). *A suite of programs for pre- and post-imputation data checking*. Paper presented at the The American Society of Human Genetics, Vancouver, Canada. http://www.ashg.org/2016meeting/

Reitan, R. M., & Wolfson, D. (1985). *The Halstead–Reitan Neuropsychological Test Battery: Therapy and clinical interpretation*. Tucson, AZ: Neuropsychological Press.

Richiardi, L., Bellocco, R., & Zugna, D. (2013). Mediation analysis in epidemiology: methods, interpretation and bias. *Int J Epidemiol, 42*(5), 1511-1519. doi:10.1093/ije/dyt127

Ridge, P. G., Hoyt, K. B., Boehme, K., Mukherjee, S., Crane, P. K., Haines, J. L., . . . Alzheimer's Disease Genetics, C. (2016). Assessment of the genetic variance of late-onset Alzheimer's disease. *Neurobiol Aging, 41*, 200 e213-200 e220. doi:10.1016/j.neurobiolaging.2016.02.024

Rist, M. J., Roth, A., Frommherz, L., Weinert, C. H., Kruger, R., Merz, B., . . . Watzl, B. (2017). Metabolite patterns predicting sex and age in participants of the Karlsruhe Metabolomics and Nutrition (KarMeN) study. *PLoS One, 12*(8), e0183228. doi:10.1371/journal.pone.0183228

Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A., & Kim, D. (2015). Methods of integrating data to uncover genotype-phenotype interactions. *Nat Rev Genet, 16*(2), 85-97. doi:10.1038/nrg3868

Sabater-Lleal, M., Huang, J., Chasman, D., Naitza, S., Dehghan, A., Johnson, A. D., . . . O'Donnell, C. J. (2013). Multiethnic meta-analysis of genome-wide association

studies in >100 000 subjects identifies 23 fibrinogen-associated Loci but no strong evidence of a causal association between circulating fibrinogen and cardiovascular disease. *Circulation, 128*(12), 1310-1324. doi:10.1161/CIRCULATIONAHA.113.002251

Sager, M. A., Hermann, B., & La Rue, A. (2005). Middle-aged children of persons with Alzheimer's disease: APOE genotypes and cognitive function in the Wisconsin Registry for Alzheimer's Prevention. *J Geriatr Psychiatry Neurol, 18*(4), 245-249. doi:10.1177/0891988705281882

Schmidt, M. (1996). *Rey Auditory Verbal Learning Test: A Handbook*. Los Angeles, California: Western Psychological Services.

Serrano-Pozo, A., Frosch, M. P., Masliah, E., & Hyman, B. T. (2011). Neuropathological alterations in Alzheimer disease. *Cold Spring Harb Perspect Med, 1*(1), a006189. doi:10.1101/cshperspect.a006189

Shah, S. H., & Newgard, C. B. (2015). Integrated metabolomics and genomics: systems approaches to biomarkers and mechanisms of cardiovascular disease. *Circ Cardiovasc Genet, 8*(2), 410-419. doi:10.1161/CIRCGENETICS.114.000223

Shin, S. Y., Fauman, E. B., Petersen, A. K., Krumsiek, J., Santos, R., Huang, J., . . . Soranzo, N. (2014). An atlas of genetic influences on human blood metabolites. *Nat Genet, 46*(6), 543-550. doi:10.1038/ng.2982

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics, 23*(4), 323-355.

Skoog, I., Davidsson, P., Aevarsson, O., Vanderstichele, H., Vanmechelen, E., & Blennow, K. (2003). Cerebrospinal fluid beta-amyloid 42 is reduced before the onset of sporadic dementia: a population-based study in 85-year-olds. *Dementia and geriatric cognitive disorders, 15*(3), 169-176. doi:68478

Smith, G. D., & Ebrahim, S. (2003). 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol, 32*(1), 1-22.

Snowden, S. G., Ebshiana, A. A., Hye, A., An, Y., Pletnikova, O., O'Brien, R., . . . Thambisetty, M. (2017). Association between fatty acid metabolism in the brain and Alzheimer disease neuropathology and cognitive performance: A nontargeted metabolomic study. *PLoS Med, 14*(3), e1002266. doi:10.1371/journal.pmed.1002266

Spencer, J. B., Klein, M., Kumar, A., & Azziz, R. (2007). The age-associated decline of androgens in reproductive age and menopausal Black and White women. *J Clin Endocrinol Metab, 92*(12), 4730-4733. doi:10.1210/jc.2006-2365

Stieb, D. M., Boot, C. R., & Turner, M. C. (2017). Promise and pitfalls in the application of big data to occupational and environmental health. *BMC Public Health, 17*(1), 372. doi:10.1186/s12889-017-4286-8

Stock, J. H., Wright, J. H., & Yogo, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics, 20*(4), 518-529. doi:10.1198/073500102288618658

Stomrud, E., Hansson, O., Blennow, K., Minthon, L., & Londos, E. (2007). Cerebrospinal fluid biomarkers predict decline in subjective cognitive function

over 3 years in healthy elderly. *Dementia and geriatric cognitive disorders, 24*(2), 118-124. doi:10.1159/000105017

Suhre, K., Shin, S. Y., Petersen, A. K., Mohney, R. P., Meredith, D., Wagele, B., . . . Gieger, C. (2011). Human metabolic individuality in biomedical and pharmaceutical research. *Nature, 477*(7362), 54-60. doi:10.1038/nature10354

Summar, M. L., Gainer, J. V., Pretorius, M., Malave, H., Harris, S., Hall, L. D., . . . Brown, N. J. (2004). Relationship between carbamoyl-phosphate synthetase genotype and systemic vascular function. *Hypertension, 43*(2), 186-191. doi:10.1161/01.HYP.0000112424.06921.52

Takashina, C., Tsujino, I., Watanabe, T., Sakaue, S., Ikeda, D., Yamada, A., . . . Nishimura, M. (2016). Associations among the plasma amino acid profile, obesity, and glucose metabolism in Japanese adults with normal glucose tolerance. *Nutr Metab (Lond), 13*, 5. doi:10.1186/s12986-015-0059-5

Thalacker-Mercer, A. E., Ingram, K. H., Guo, F., Ilkayeva, O., Newgard, C. B., & Garvey, W. T. (2014). BMI, RQ, diabetes, and sex affect the relationships between amino acids and clamp measures of insulin action in humans. *Diabetes, 63*(2), 791-800. doi:10.2337/db13-0396

TNT, C. (2017). Trans-Omics for Precision Medicine (TOPMed). Retrieved April 14, 2018 https://bravo.sph.umich.edu

Trenerry, M. R., Crosson, B., DeBoe, J., & Leber, W. R. (1989). *Stroop Neuropsychological Screening Test Manual*: Psychological Assessment Resources.

Trushina, E., & Mielke, M. M. (2014). Recent advances in the application of metabolomics to Alzheimer's Disease. *Biochim Biophys Acta, 1842*(8), 1232-1239. doi:10.1016/j.bbadis.2013.06.014

Tzoulaki, I., Ebbels, T. M., Valdes, A., Elliott, P., & Ioannidis, J. P. (2014). Design and analysis of metabolomics studies in epidemiologic research: a primer on -omic technologies. *Am J Epidemiol, 180*(2), 129-139. doi:10.1093/aje/kwu143

Ussher, J. R., Elmariah, S., Gerszten, R. E., & Dyck, J. R. (2016). The Emerging Role of Metabolomics in the Diagnosis and Prognosis of Cardiovascular Disease. *J Am Coll Cardiol, 68*(25), 2850-2870. doi:10.1016/j.jacc.2016.09.972

van den Berg, R. A., Hoefsloot, H. C., Westerhuis, J. A., Smilde, A. K., & van der Werf, M. J. (2006). Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics, 7*, 142. doi:10.1186/1471-2164-7-142

van Meurs, J. B., Pare, G., Schwartz, S. M., Hazra, A., Tanaka, T., Vermeulen, S. H., . . . Ahmadi, K. R. (2013). Common genetic loci influencing plasma homocysteine concentrations and their effect on risk of coronary artery disease. *Am J Clin Nutr, 98*(3), 668-676. doi:10.3945/ajcn.112.044545

Vermeulen, A. (2000). Andropause. *Maturitas, 34*(1), 5-15.

Voyle, N., Kim, M., Proitsi, P., Ashton, N. J., Baird, A. L., Bazenet, C., . . . Alzheimer's Disease Neuroimaging, I. (2016). Blood metabolite markers of neocortical amyloid-beta burden: discovery and enrichment using candidate proteins. *Transl Psychiatry, 6*, e719. doi:10.1038/tp.2015.205

Wechsler, D. (1981). *Wechsler Adult Intelligence Scale - Revised*. San Antonio: The Psychological Corporation.

Wechsler, D. (1987). *Wechsler Memory Scale-Revised*. New York: The Psychological Corporation, Harcourt Brace Jovanovich, inc for Psychological Corp.

Willer, C. J., Schmidt, E. M., Sengupta, S., Peloso, G. M., Gustafsson, S., Kanoni, S., . . . Global Lipids Genetics, C. (2013). Discovery and refinement of loci associated with lipid levels. *Nat Genet, 45*(11), 1274-1283. doi:10.1038/ng.2797

Williams, S. R., Yang, Q., Chen, F., Liu, X., Keene, K. L., Jacques, P., . . . Framingham Heart, S. (2014). Genome-wide meta-analysis of homocysteine and methionine metabolism identifies five one carbon metabolism loci and a novel association of ALDH1L1 with ischemic stroke. *PLoS Genet, 10*(3), e1004214. doi:10.1371/journal.pgen.1004214

Wishart, D. S., Feunang, Y. D., Marcu, A., Guo, A. C., Liang, K., Vazquez-Fresno, R., . . . Scalbert, A. (2018). HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res, 46*(D1), D608-D617. doi:10.1093/nar/gkx1089

Xia, J., & Wishart, D. S. (2016). Using MetaboAnalyst 3.0 for Comprehensive Metabolomics Data Analysis. *Curr Protoc Bioinformatics, 55*, 14 10 11-14 10 91. doi:10.1002/cpbi.11

Xie, W., Wood, A. R., Lyssenko, V., Weedon, M. N., Knowles, J. W., Alkayyali, S., . . . Walker, M. (2013). Genetic variants associated with glycine metabolism and their role in insulin sensitivity and type 2 diabetes. *Diabetes, 62*(6), 2141-2150. doi:10.2337/db12-0876

Yang, J., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2011). GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet, 88*(1), 76-82. doi:10.1016/j.ajhg.2010.11.011

Yu, B., Zheng, Y., Alexander, D., Morrison, A. C., Coresh, J., & Boerwinkle, E. (2014). Genetic determinants influencing human serum metabolome among African Americans. *PLoS Genet, 10*(3), e1004212. doi:10.1371/journal.pgen.1004212

Yu, Z., Zhai, G., Singmann, P., He, Y., Xu, T., Prehn, C., . . . Wang-Sattler, R. (2012). Human serum metabolic profiles are age dependent. *Aging Cell, 11*(6), 960-967. doi:10.1111/j.1474-9726.2012.00865.x

Zaitlen, N., Kraft, P., Patterson, N., Pasaniuc, B., Bhatia, G., Pollack, S., & Price, A. L. (2013). Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. *PLoS Genet, 9*(5), e1003520. doi:10.1371/journal.pgen.1003520

Zhang, A. H., Sun, H., & Wang, X. J. (2017). Emerging role and recent applications of metabolomics biomarkers in obesity disease research. *Rsc Advances, 7*(25), 14966-14973. doi:10.1039/c6ra28715h

Appendices

Appendix A

Appendix A1. Metabolomics quantification

Plasma metabolites were profiled by Metabolon (Durham, NC) using Ultrahigh Performance Liquid Chromatography-Tandom Mass Spectrometry (UPLC-MS/MS). Samples were prepared using the automated MicroLab STAR® system from Hamilton Company. Several recovery standards were added prior to the first step in the extraction process for QC purposes. To remove protein, dissociate small molecules bound to protein or trapped in the precipitated protein matrix, and to recover chemically diverse metabolites, proteins were precipitated with methanol under vigorous shaking for 2 min (Glen Mills GenoGrinder 2000) followed by centrifugation. The resulting extract was divided into five fractions: two for analysis by two separate reverse phase (RP)/UPLC-MS/MS methods with positive ion mode electrospray ionization (ESI), one for analysis by RP/UPLC-MS/MS with negative ion mode ESI, one for analysis by HILIC/UPLC-MS/MS with negative ion mode ESI, and one sample was reserved for backup. Samples were placed briefly on a TurboVap® (Zymark) to remove the organic solvent. The sample extracts were stored overnight under nitrogen before preparation for analysis.

Several types of controls were analyzed in concert with the experimental samples: a pooled matrix sample generated by taking a small volume of each experimental sample (or alternatively, use of a pool of well-characterized human plasma) served as a technical replicate throughout the data set; extracted water samples served as process blanks; and a cocktail of QC standards that were carefully

chosen not to interfere with the measurement of endogenous compounds were spiked into every analyzed sample, allowed instrument performance monitoring and aided chromatographic alignment. Tables 1 and 2 describe these QC samples and standards. Instrument variability was determined by calculating the median relative standard deviation (RSD) for the standards that were added to each sample prior to injection into the mass spectrometers. Overall process variability was determined by calculating the median RSD for all endogenous metabolites (i.e., non-instrument standards) present in 100% of the pooled matrix samples. Experimental samples were randomized across the platform run with QC samples spaced evenly among the injections.

All methods utilized a Waters ACQUITY ultra-performance liquid chromatography (UPLC) and a Thermo Scientific Q-Exactive high resolution/accurate mass spectrometer interfaced with a heated electrospray ionization (HESI-II) source and Orbitrap mass analyzer operated at 35,000 mass resolution. The sample extract was dried then reconstituted in solvents compatible to each of the four methods. Each reconstitution solvent contained a series of standards at fixed concentrations to ensure injection and chromatographic consistency. One aliquot was analyzed using acidic positive ion conditions, chromatographically optimized for more hydrophilic compounds. In this method, the extract was gradient eluted from a C18 column (Waters UPLC BEH C18-2.1x100 mm, 1.7 µm) using water and methanol, containing 0.05% perfluoropentanoic acid (PFPA) and 0.1% formic acid (FA). Another aliquot was also analyzed using acidic positive ion conditions, however it was chromatographically optimized for more hydrophobic compounds. In this method, the extract was gradient eluted from the same afore mentioned C18 column using methanol, acetonitrile, water, 0.05% PFPA and

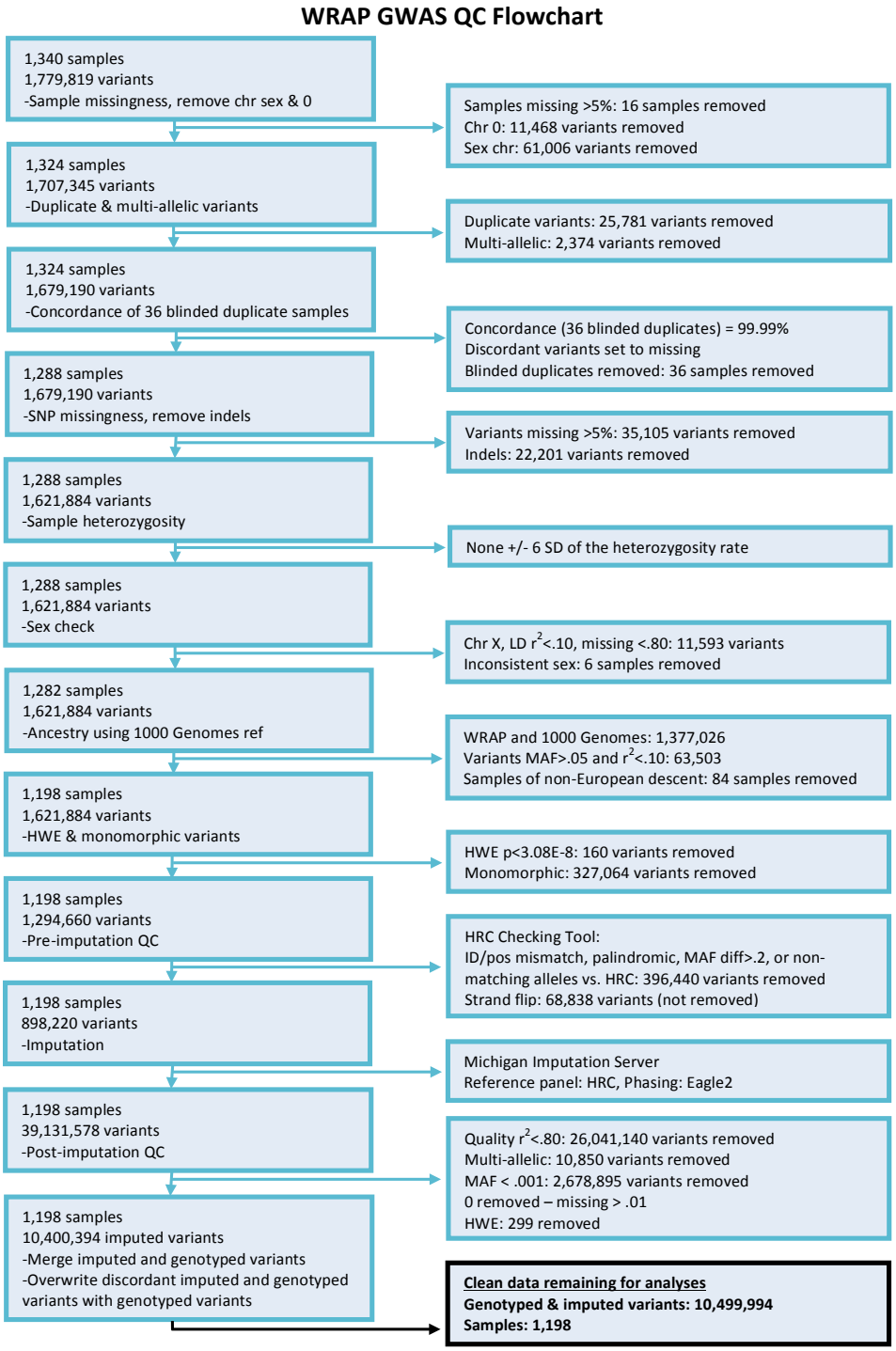
0.01% FA and was operated at an overall higher organic content. Another aliquot was analyzed using basic negative ion optimized conditions using a separate dedicated C18 column. The basic extracts were gradient eluted from the column using methanol and water, however with 6.5mM Ammonium Bicarbonate at pH 8. The fourth aliquot was analyzed via negative ionization following elution from a HILIC column (Waters UPLC BEH Amide 2.1x150 mm, 1.7 µm) using a gradient consisting of water and acetonitrile with 10mM Ammonium Formate, pH 10.8. The MS analysis alternated between MS and data-dependent MS^n scans using dynamic exclusion. The scan range varied slighted between methods but covered 70-1000 m/z. Raw data files are archived and extracted as described below.

Raw data was extracted, peak-identified and QC processed using Metabolon's hardware and software. These systems are built on a web-service platform utilizing Microsoft's .NET technologies, which run on high-performance application servers and fiber-channel storage arrays in clusters to provide active failover and load-balancing. Compounds were identified by comparison to library entries of purified standards or recurrent unknown entities. Metabolon maintains a library based on authenticated standards that contains the retention time/index (RI), mass to charge ratio (*m/z)*, and chromatographic data (including MS/MS spectral data) on all molecules present in the library. Furthermore, biochemical identifications are based on three criteria: retention index within a narrow RI window of the proposed identification, accurate mass match to the library +/- 10 ppm, and the MS/MS forward and reverse scores between the experimental data and authentic standards. The MS/MS scores are based on a comparison of the ions present in the experimental spectrum to the ions present in the

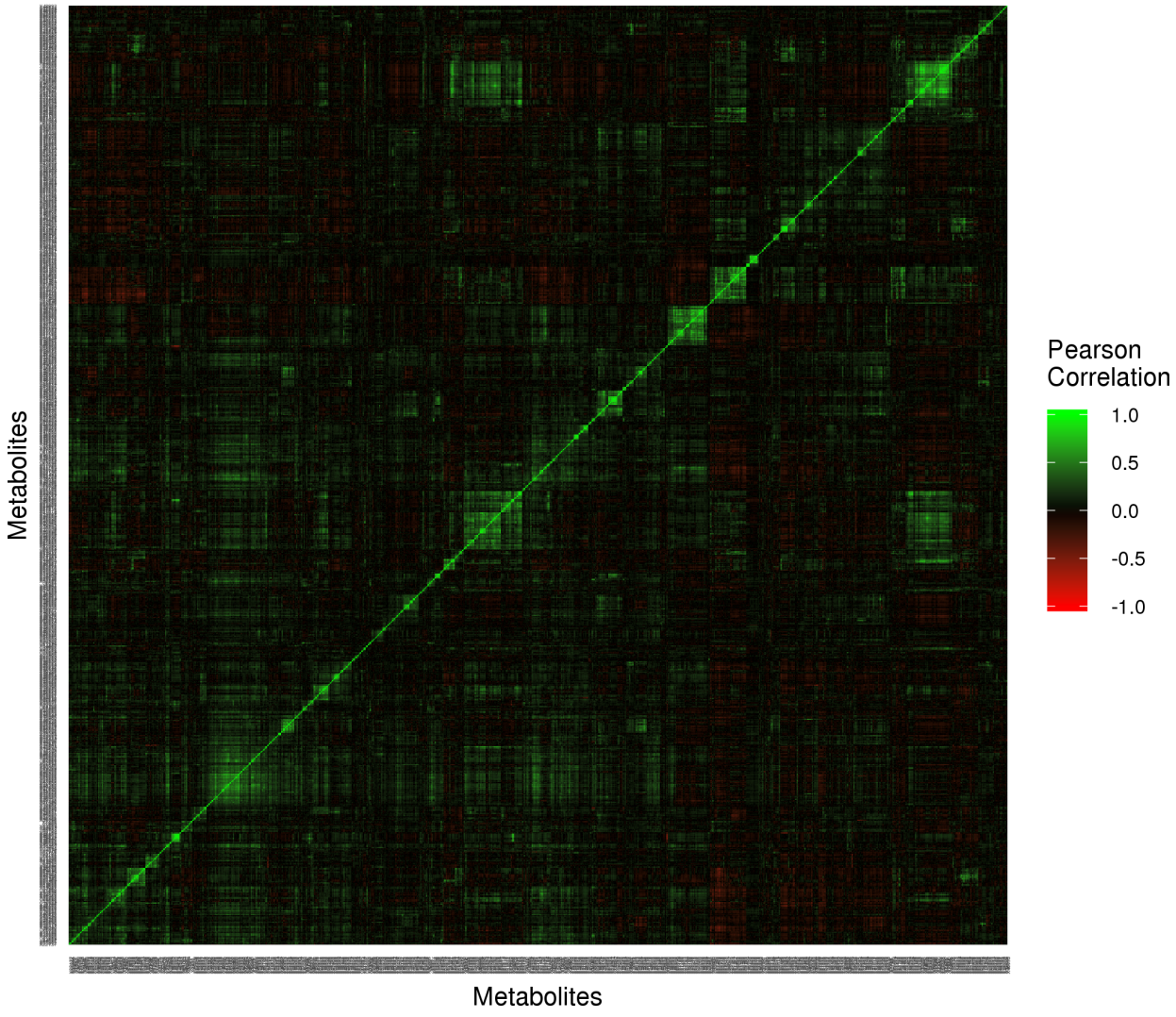
library spectrum. While there may be similarities between these molecules based on one of these factors, the use of all three data points can be utilized to distinguish and differentiate biochemicals. More than 3300 commercially available purified standard compounds have been acquired and registered into LIMS for analysis on all platforms for determination of their analytical characteristics. Additional mass spectral entries have been created for structurally unnamed biochemicals, which have been identified by virtue of their recurrent nature (both chromatographic and mass spectral). These compounds have the potential to be identified by future acquisition of a matching purified standard or by classical structural analysis.

A variety of curation procedures were carried out to ensure that a high quality data set was made available for statistical analysis and data interpretation. The QC and curation processes were designed to ensure accurate and consistent identification of true chemical entities, and to remove those representing system artifacts, mis-assignments, and background noise. Metabolon data analysts use proprietary visualization and interpretation software to confirm the consistency of peak identification among the various samples. Library matches for each compound were checked for each sample and corrected if necessary.

Peaks were quantified using area-under-the-curve. A data normalization setp was performed to correct variation resulting from instrument inter-day tuning differences. Essentially, each compound was corrected in run-day blocks by registering the medians to equal one and normaling each data point proportionately.



Appendix A2. GWAS QC Flowchart.

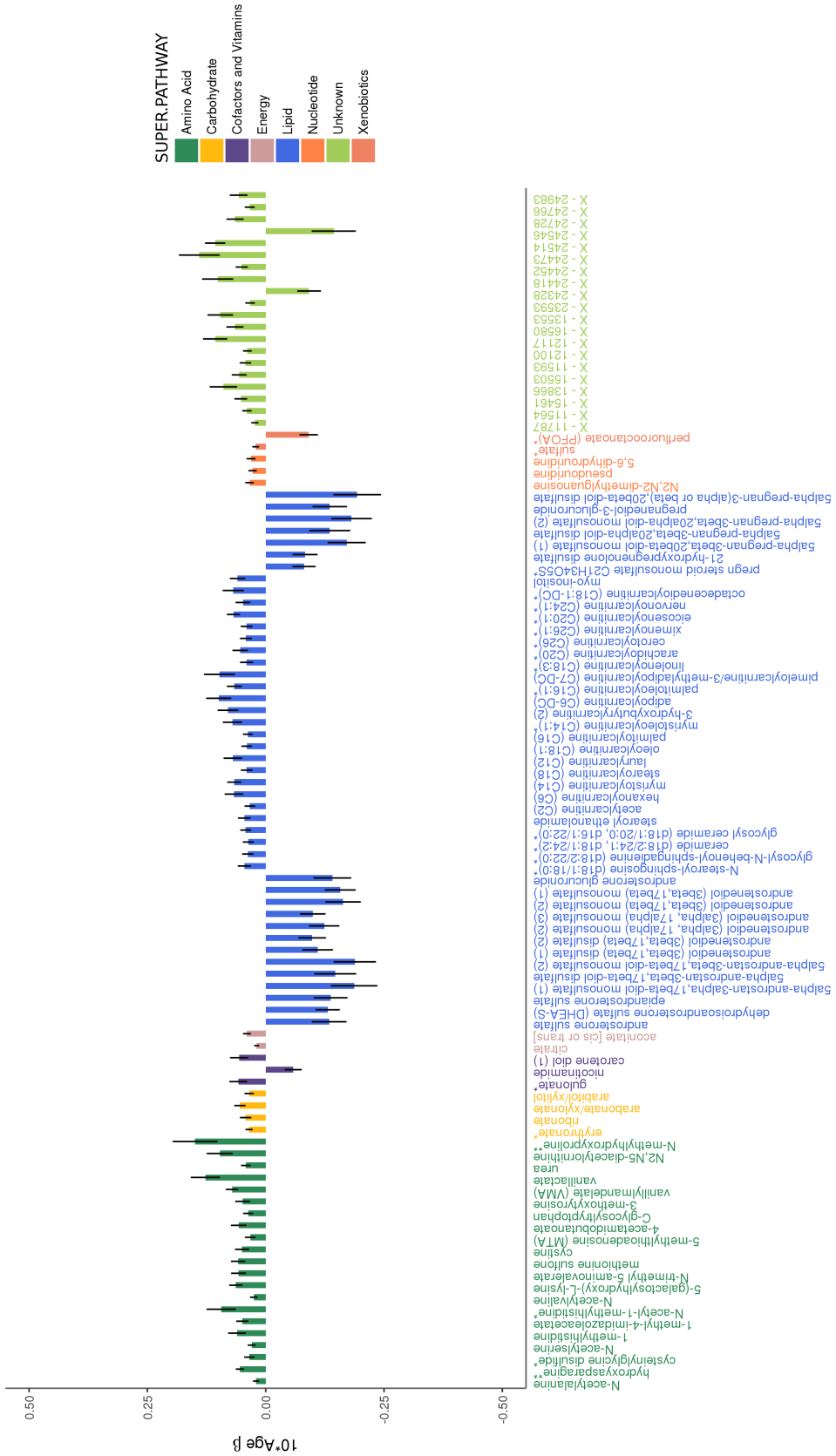


Appendix A3. Heatmap of metabolite correlations.

Pearson correlations (r) are displayed and hierarchical clustering was used to sort metabolites. Clusters of strong positive correlations are largely between amino acid–amino acid and between lipid–lipid (i.e., correlations between super pathway had weaker correlations). Negative correlations are less common and all have an r≥-0.62.

Appendix A4. Metabolome-wide association study results, metabolite properties, and metabolite heritability estimates.

Due to the size of the table, it is provided in an excel spreadsheet.


SUPER.PATHWAY
Amino Acid
Carbohydrate
Cofactors and Vitamins
Energy
Lipid
Nucleotide
Unknown
Xenobiotics
0.50
0.25
0.00
-0.25
-0.50
10*Age β
N-acetylalanine
hydroxyasparagine**
cysteinylglycine disulfide*
N-acetylserine
1-methylhistidine
1-methyl-4-imidazoleacetate
N-acetyl-1-methylhistidine*
N-acetylvaline
5-(galactosylhydroxy)-L-lysine
N-trimethyl 5-aminovalerate
methionine sulfone
cystine
5-methylthioadenosine (MTA)
4-acetamidobutanoate
C-glycosyltryptophan
3-methoxytyrosine
vanillylmandelate (VMA)
vanillactate
urea
N2,N5-diacetylornithine
N-methylhydroxyproline**
erythronate*
arabonate/xylonate
ribonate
arabitol/xylitol
gulonate*
nicotinamide
carotene diol (1)
citrate
aconitate [cis or trans]
androsterone sulfate
dehydroisoandrosterone sulfate (DHEA-S)
epiandrosterone sulfate
5alpha-androstan-3alpha,17beta-diol monosulfate (1)
5alpha-androstan-3beta,17beta-diol disulfate
5alpha-androstan-3beta,17beta-diol monosulfate (2)
androstenediol (3beta,17beta) disulfate (1)
androstenediol (3beta,17beta) disulfate (2)
androstenediol (3alpha, 17alpha) monosulfate (2)
androstenediol (3alpha, 17alpha) monosulfate (3)
androstenediol (3beta,17beta) monosulfate (2)
androstenediol (3beta,17beta) monosulfate (1)
androsterone glucuronide
N-stearoyl-sphingosine (d18:1/18:0)*
glycosyl-N-behenoyl-sphingadienine (d18:2/22:0)*
ceramide (d18:2/24:1, d18:1/24:2)*
glycosyl ceramide (d18:1/20:0, d16:1/22:0)*
stearoyl ethanolamide
acetylcarnitine (C2)
hexanoylcarnitine (C6)
myristoylcarnitine (C14)
stearoylcarnitine (C18)
laurylcarnitine (C12)
oleoylcarnitine (C18:1)
palmitoylcarnitine (C16)
myristoleoylcarnitine (C14:1)*
3-hydroxybutyrylcarnitine (2)
adipoylcarnitine (C6-DC)
palmitoleoylcarnitine (C16:1)*
pimeloylcarnitine/3-methyladipoylcarnitine (C7-DC)
linolenoylcarnitine (C18:3)*
arachidoylcarnitine (C20)*
cerotoylcarnitine (C26)*
ximenoylcarnitine (C26:1)*
eicosenoylcarnitine (C20:1)*
nervonoylcarnitine (C24:1)*
octadecenedioylcarnitine (C18:1-DC)*
myo-inositol
pregn steroid monosulfate C21H34O5S*
21-hydroxypregnenolone disulfate
5alpha-pregnan-3beta,20beta-diol monosulfate (1)
5alpha-pregnan-3beta,20alpha-diol disulfate
5alpha-pregnan-3beta,20alpha-diol monosulfate (2)
pregnanediol-3-glucuronide
5alpha-pregnan-3(alpha or beta),20beta-diol disulfate
N2,N2-dimethylguanosine
pseudouridine
5,6-dihydrouridine
sulfate*
perfluorooctanoate (PFOA)*
X - 11787
X - 11564
X - 13866
X - 15461
X - 15503
X - 11593
X - 12100
X - 12117
X - 16580
X - 13553
X - 23593
X - 24328
X - 24418
X - 24452
X - 24473
X - 24514
X - 24546
X - 24728
X - 24766
X - 24983

Appendix A5. Age stratified by sex: Adjusted effects of a 10-year increase in age on the top 100 metabolites most strongly influenced by age in women.

Positive values indicate the amount a metabolite increased over 10 years, whereas negative values indicate the amount a metabolite decreased over 10 years. Black vertical lines indicate standard errors.


SUPER.PATHWAY
Amino Acid
Carbohydrate
Cofactors and Vitamins
Energy
Lipid
Nucleotide
Peptide
Unknown
Xenobiotics
0.50
0.25
0.00
-0.25
-0.50
10*Age β
N-acetylaspartate (NAA)*
hydroxyasparagine**
N-acetylglutamine
cysteinylglycine
N-acetylserine
1-methyl-4-imidazoleacetate
N-acetyl-1-methylhistidine*
3-methyl-2-oxovalerate
4-methyl-2-oxopentanoate
3-methylglutaconate
5-(galactosylhydroxy)-L-lysine
cystine
hypotaurine
phenylacetate
(N(1) + N(8))-acetylspermidine
C-glycosyltryptophan
3-methoxytyrosine
vanillylmandelate (VMA)
p-cresol-glucuronide*
N2,N5-diacetylornithine
erythronate*
bilirubin (E,E)*
quinolinate
nicotinamide
gamma-tocopherol/beta-tocopherol
citrate
aconitate [cis or trans]
androsterone sulfate
dehydroisoandrosterone sulfate (DHEA-S)
andro steroid monosulfate C19H28O6S (1)*
5alpha-androstan-3alpha,17beta-diol monosulfate (1)*
5alpha-androstan-3beta,17beta-diol monosulfate (2)
androstenediol (3beta,17beta) disulfate (1)
androstenediol (3beta,17beta) disulfate (2)
androstenediol (3alpha, 17alpha) monosulfate (2)
androstenediol (3beta,17beta) monosulfate (2)
androstenediol (3beta,17beta) monosulfate (1)
5alpha-androstan-3alpha,17beta-diol 17-glucuronide
androsterone glucuronide
cortisol
pimeloylcarnitine/3-methyladipoylcarnitine (C7-DC)
eicosenoylcarnitine (C20:1)*
3-hydroxyadipate*
5-hydroxyhexanoate
3-hydroxysebacate
heptanoate (7:0)
3-hydroxy-3-methylglutarate
1-stearoyl-2-arachidonoyl-GPC (18:0/20:4)
1-palmitoyl-2-dihomo-linolenoyl-GPC (16:0/20:3n3 or 6)*
1-linoleoyl-2-arachidonoyl-GPC (18:2/20:4n6)*
1-myristoyl-2-arachidonoyl-GPC (14:0/20:4)*
choline
choline phosphate
1-(1-enyl-palmitoyl)-2-oleoyl-GPC (P-16:0/18:1)*
1-(1-enyl-palmitoyl)-2-palmitoyl-GPC (P-16:0/16:0)*
pregn steroid monosulfate C21H34O5S*
21-hydroxypregnenolone disulfate
tricosanoyl sphingomyelin (d18:1/23:0)*
behenoyl dihydrosphingomyelin (d18:0/22:0)*
N1-methylinosine
N2,N2-dimethylguanosine
orotate
orotidine
3-aminoisobutyrate
3-ureidopropionate
pseudouridine
uridine
phenylacetylglutamine
2-hydroxyhippurate (salicylurate)
p-cresol sulfate
sulfate*
perfluorooctanesulfonic acid (PFOS)
perfluorooctanoate (PFOA)*
salicylic glucuronide*
salicylate
ergothioneine
X - 11381
X - 11843
X - 11850
X - 12729
X - 12798
X - 12851
X - 17690
X - 21736
X - 15503
X - 12100
X - 12117
X - 12688
X - 17351
X - 18935
X - 17612
X - 23196
X - 23787
X - 23997
X - 24514
X - 24544
X - 24546
X - 24728
X - 24766
X - 24948

Appendix A6. Age stratified by sex: Adjusted effects of a 10-year increase in age on the top 100 metabolites most strongly influenced by age in men.

Positive values indicate that the metabolite increased with 10 years of age, whereas negative values indicate that the metabolite decreased with 10 years of age. Black vertical lines indicate standard errors.

Appendix A7. Comparison of summary statistics for metabolites that were associated with age in our analyses and in those of Menni et al. 2013.

Due to the size of the table, it is provided as an excel spreadsheet.

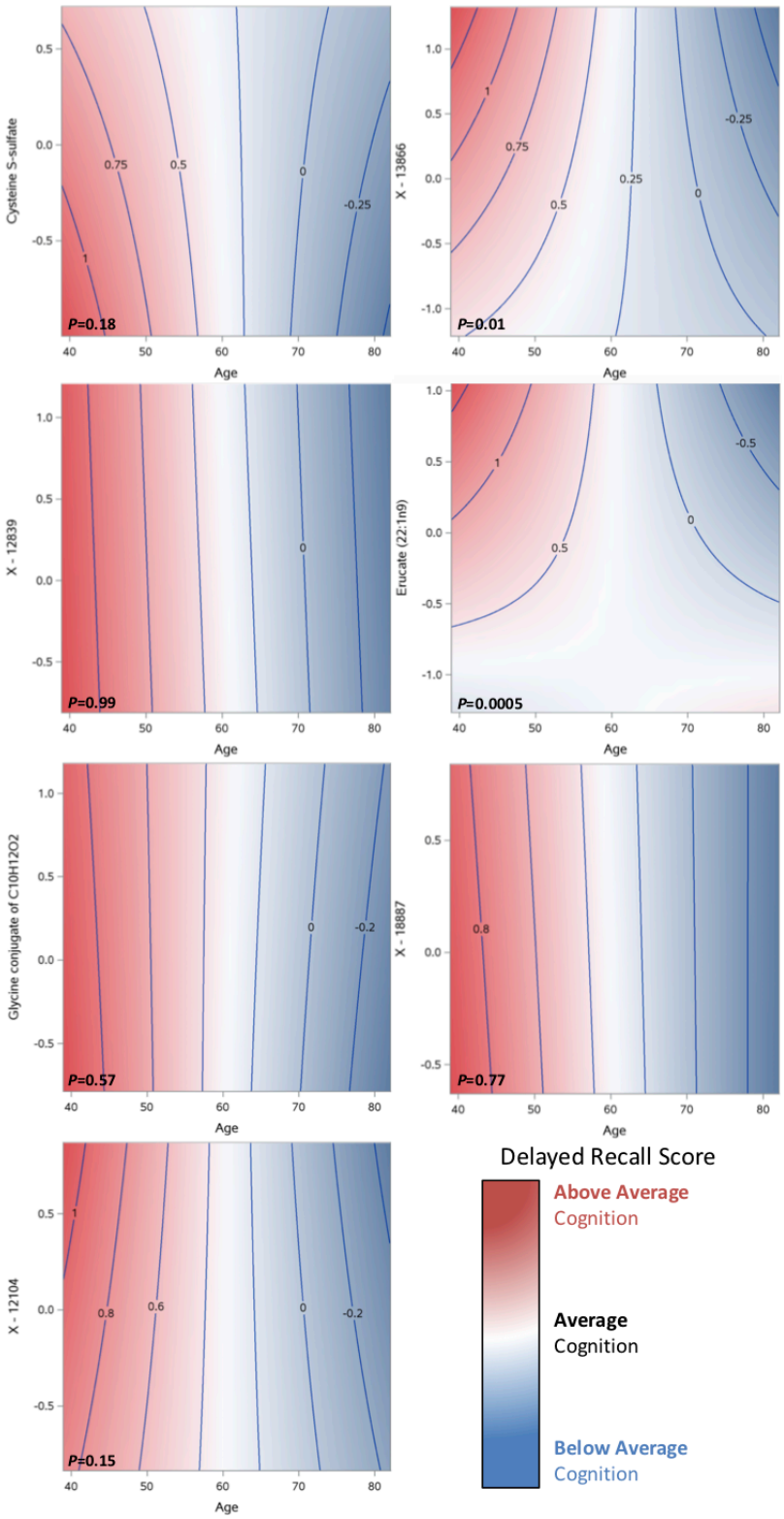
Appendix A8. Comparison of summary statistics for metabolites that were associated with sex in our analyses and in those of Krumsiek et al. 2015.

Due to the size of the table, it is provided as an excel spreadsheet.

Appendix B

Appendix B1. Metabolome-wide association study results and metabolite properties.

Due to the size of the table, it is provided in an excel spreadsheet.


Cysteine S-sulfate
X - 13866
X - 12839
Erucate (22:1n9)
Glycine conjugate of C10H12O2
X - 18887
X - 12104
Age
P=0.18
P=0.01
P=0.99
P=0.0005
P=0.57
P=0.77
P=0.15
Delayed Recall Score
Above Average Cognition
Average Cognition
Below Average Cognition

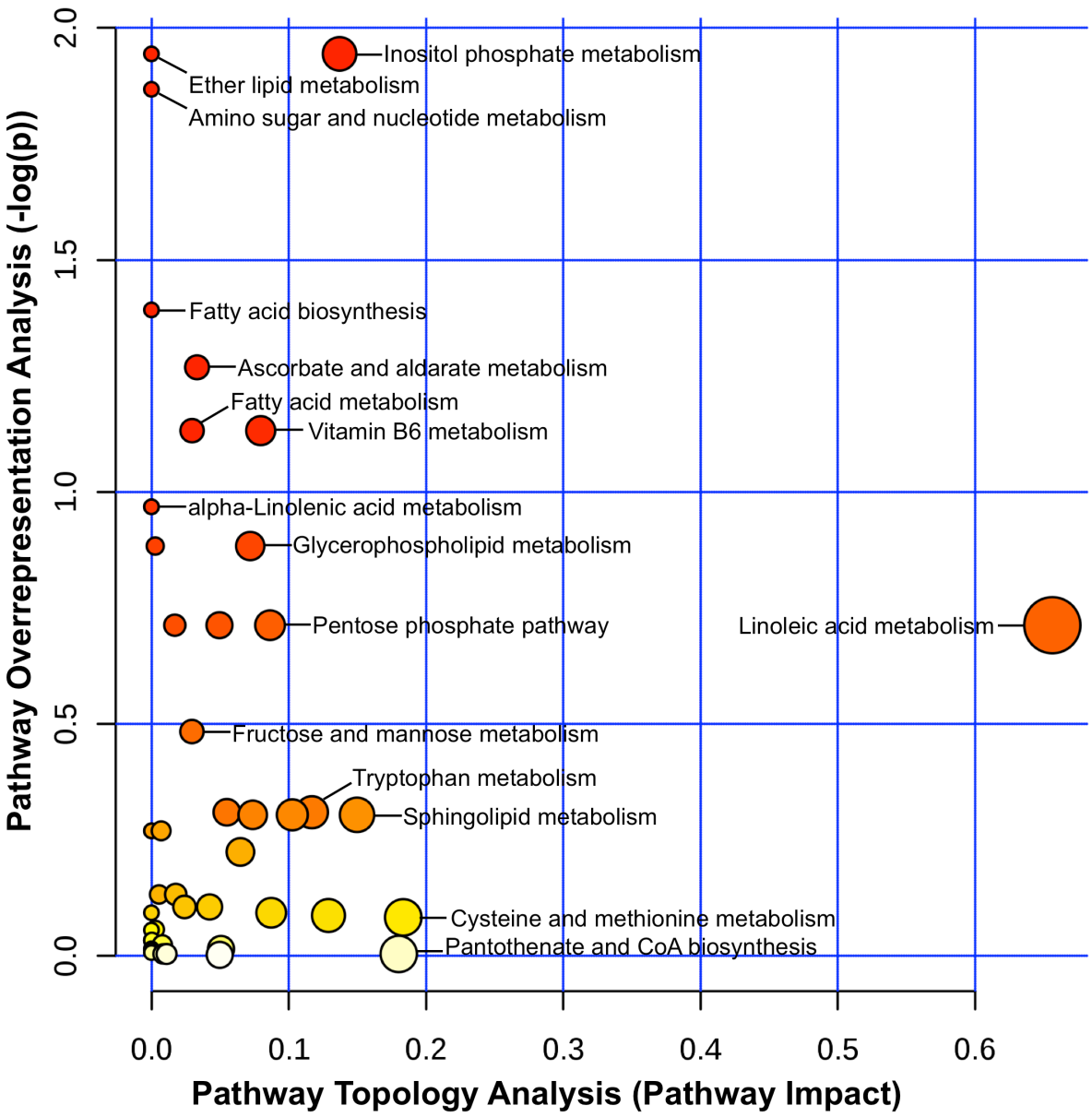
Appendix B2. Longitudinal trajectories of delayed recall by the seven metabolite levels associated with executive function.

The y-axis represents standardized metabolite levels, whereas the z-axis represents the delayed recall composite score. Although none are statistically significant, trajectories are consistent with those of executive function for the four metabolites with a $P<0.20$. Unadjusted *P*-values are indicated for each test.

Appendix B3. Mendelian Randomization Results for Executive Function.

PS		**Correlation Between PS and Metabolite**		**Discovery Results: Metabolite*Age on Executive Function**		**MR Results: PS*Age on Executive Function**	
Metabolite	*#Variants*	*Pearson r*	*F*	*β (SE)*	*P-value*	*β (SE)*	*P-value*
Cysteine S-sulfate	2,601	-0.03	0.77	0.03 (0.01)	5.2e-05	0.0004 (0.001)	0.71
X-13866	2,588	0.004	0.04	-0.02 (0.005)	7.3e-04	-0.00002 (0.001)	0.98
Erucate (22:1n9)	2,533	-0.03	1.71	-0.03 (0.01)	1.5e-04	0.001 (0.001)	0.54

PS: Polygenic score; MR: Mendelian randomization



Appendix B4. Metabolite pathway analysis for cognition.

This included 82 metabolites that were associated with executive function or delayed recall trajectories with an unadjusted $P<0.05$ and that had KEGG compound IDs. Larger circle sizes correspond to higher pathway impacts, while darker circle colors correspond to the strength of the overrepresentation analysis. The linoleic acid metabolism had the highest impact score, but it was not significant in the overrepresentation analysis.

Appendix B5. Cognitive Metabolites Pathway Analysis, Top 10 Pathways.

Pathway Name	**Match Status**	***P*-value**	**Impact**
Ether lipid metabolism	2/2	0.14	0.00
Inositol phosphate metabolism	2/2	0.14	0.14
Amino sugar and nucleotide sugar metabolism	3/4	0.15	0.00
Fatty acid biosynthesis	4/7	0.25	0.00
Ascorbate and aldarate metabolism	3/5	0.28	0.03
Fatty acid metabolism	2/3	0.32	0.03
Vitamin B6 metabolism	2/3	0.32	0.08
Fatty acid elongation in mitochondria	1/1	0.38	0.00
alpha-Linoleic acid metabolism	1/1	0.38	0.00
Galactose metabolism	3/6	0.41	0.00

P-values are unadjusted and based on the pathway enrichment analysis.

Match status indicates how many metabolites were present in our list compared to those known to be in the given pathway.

The pathway impact value is based on the pathway topology analysis.

Appendix C

Appendix C1. Properties of plasma and CSF metabolites.

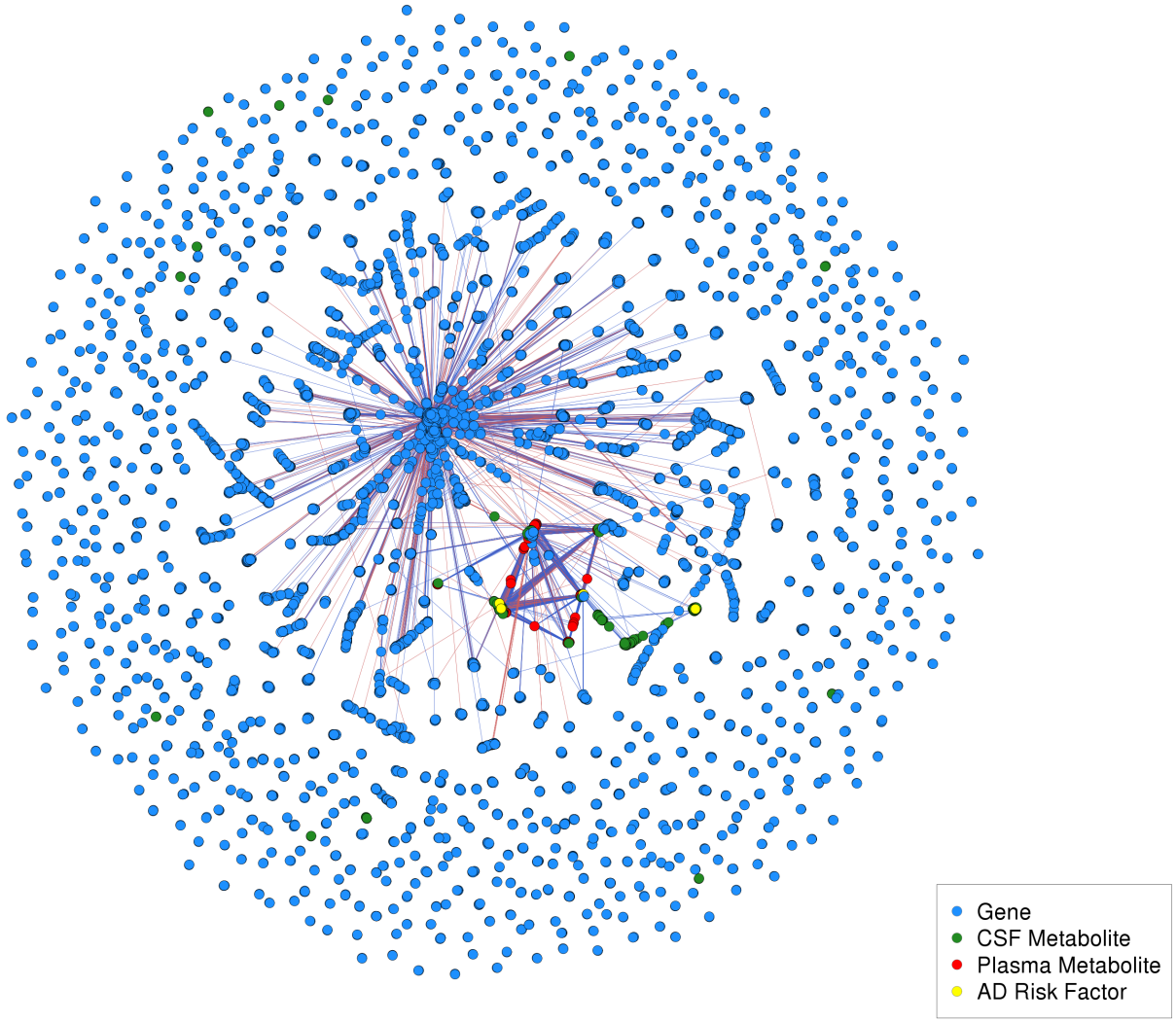
Due to the size of the table, it is provided in an excel spreadsheet.

Appendix C2. Number of plasma and CSF metabolites within each super pathway.

Super pathway	Plasma Metabolites	CSF Metabolites
Amino acids	175 (16.0%)	122 (33.5%)
Carbohydrates	23 (2.1%)	20 (5.5%)
Cofactors and vitamins	28 (2.6%)	14 (3.8%)
Energy	8 (0.7%)	6 (1.6%)
Lipids	353 (32.2%)	69 (19.0%)
Nucleotides	35 (3.2%)	30 (8.2%)
Partially characterized molecules	5 (0.5%)	--
Peptides	22 (2.0%)	8 (2.2%)
Xenobiotics	101 (9.2%)	39 (10.7%)
Unknown	347 (31.6%)	56 (15.4%)
Total	*1,097*	*364*

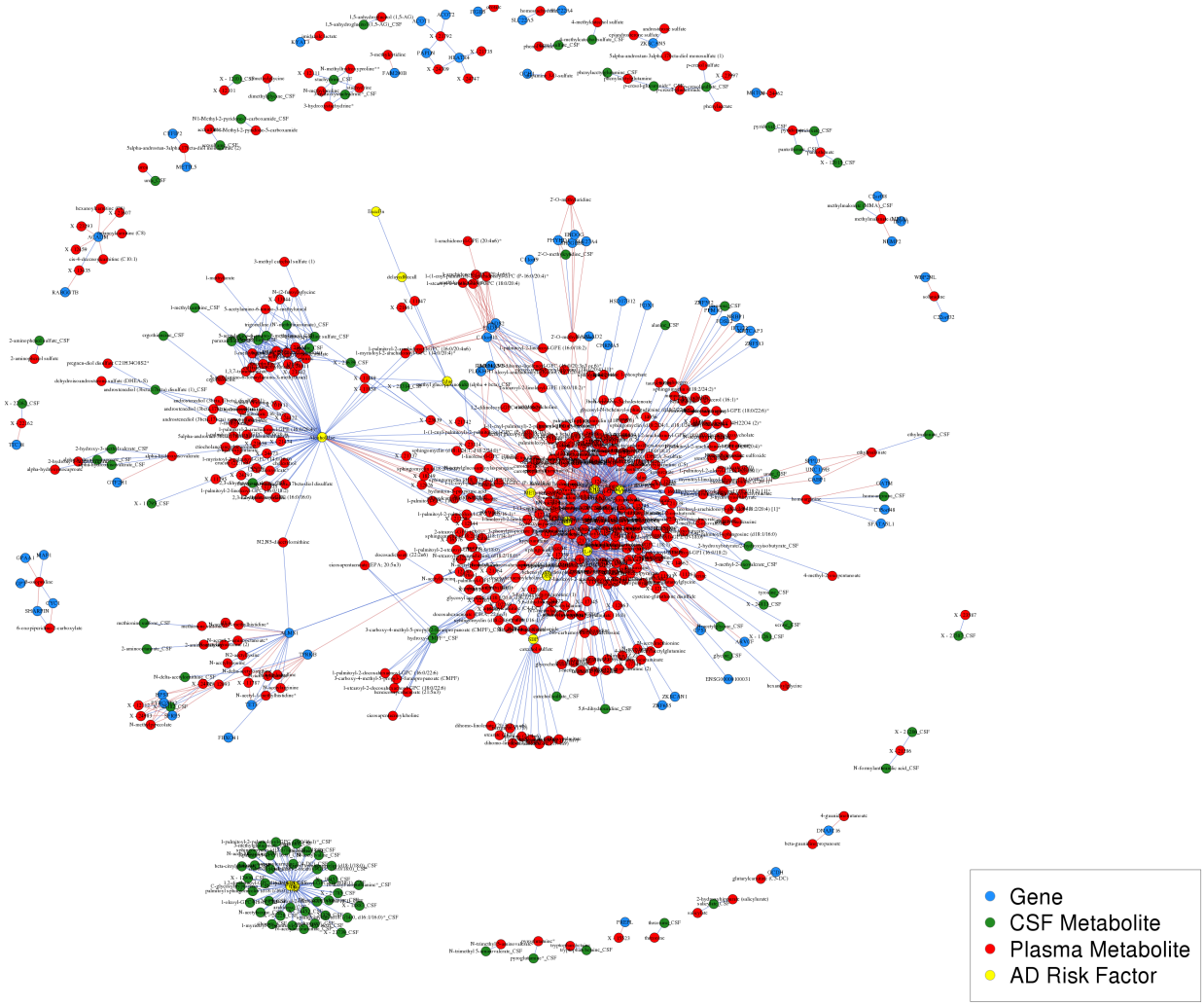
Appendix C3. Correlations between each of the 326 CSF and plasma metabolites.

Due to the size of the table, it is provided in an excel spreadsheet.



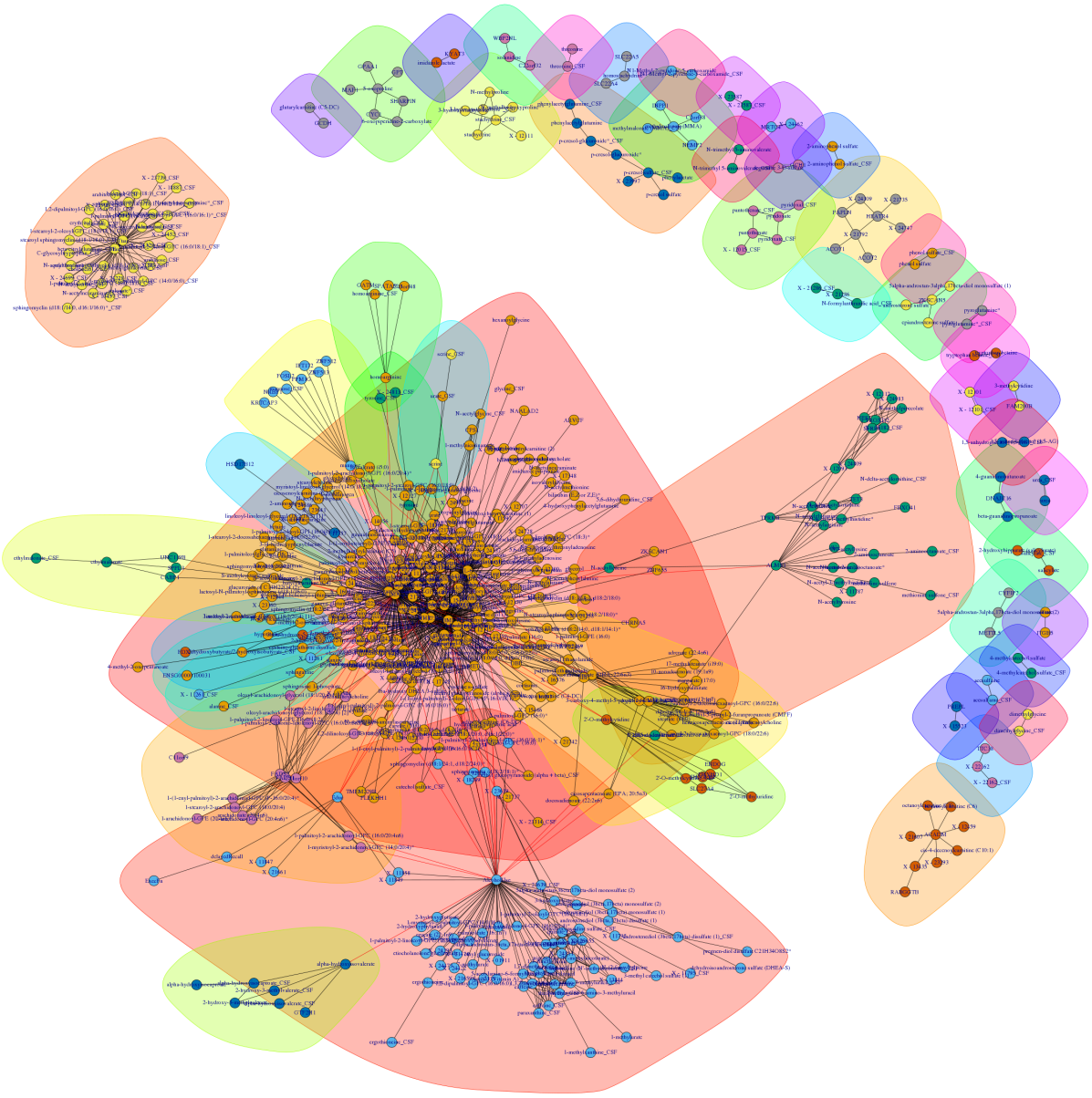
Appendix C4. Overall “hairball” network.

All significant correlations between and within each data type, using a Bonferroni-adjusted *P*-value threshold. This network has 90,308 edges and 10,869 nodes. Due to the large number of nodes in this network, many edges are hard to detect (*i.e.*, blue nodes are actually multiple overlapping genes).



Appendix C5. Labeled inter-omic network.

This figure is identical to Figure 2 but with labeled nodes. This network has 1,224 edges and 635 nodes, which included 171 metabolite-gene edges, 833 metabolite-AD risk factor edges. Of these, 73 were CSF metabolite-AD risk factor edges (CSF T-tau and P-tau, exclusively) and 4 were CSF metabolite-gene edges. Red edges indicate negative correlations and blue edges indicate positive correlations.

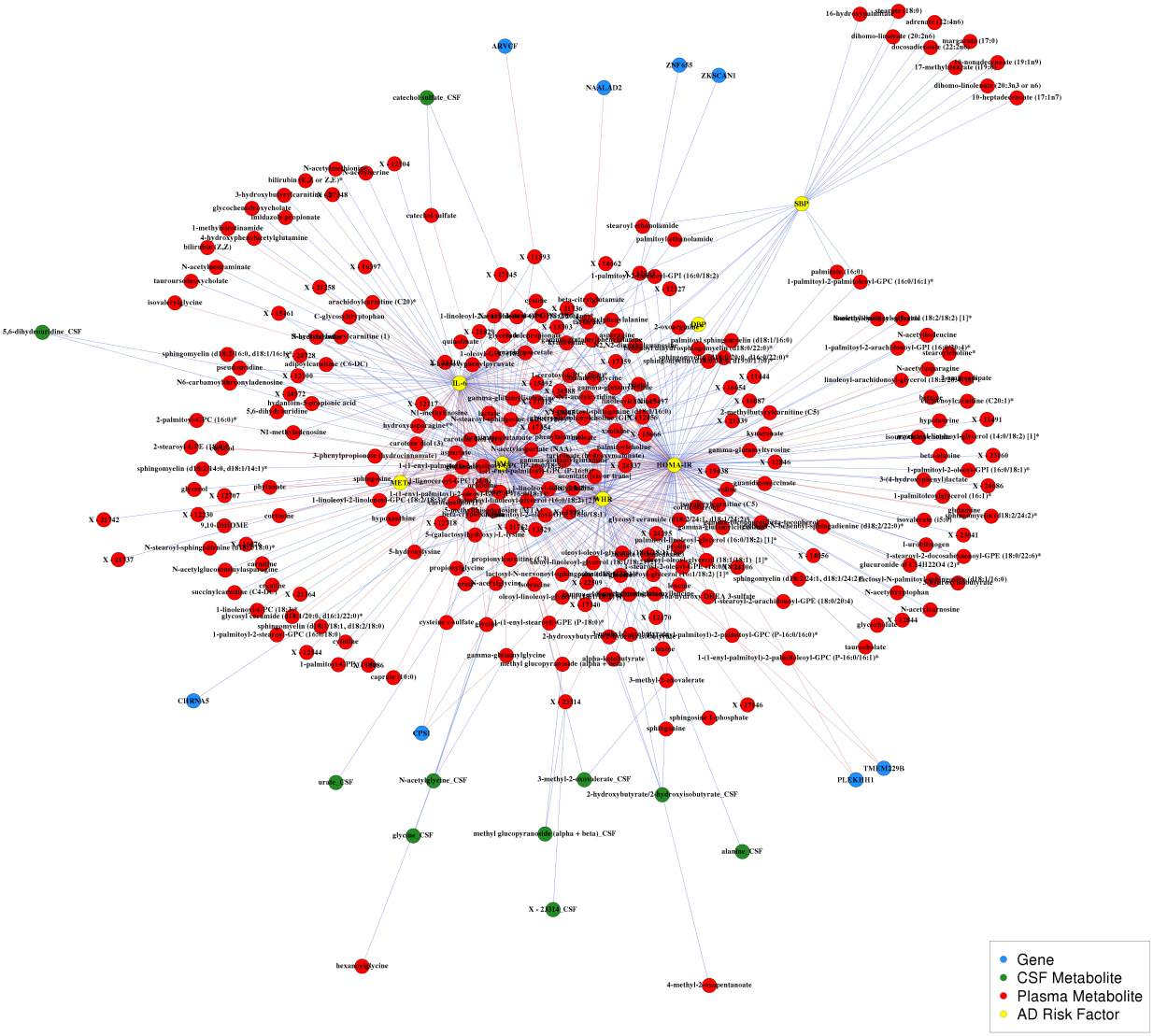


Appendix C6. Community partitions of inter-omic network.

Nodes belonging to a particular community are encompassed. The colors of encompassments only indicate which community nodes belong to (*i.e.*, similar encompassment colors do not indicate similar communities; each encompassment is a unique community). Because some encompassments appear to overlap, nodes are colored to help distinguish which community they belong to.

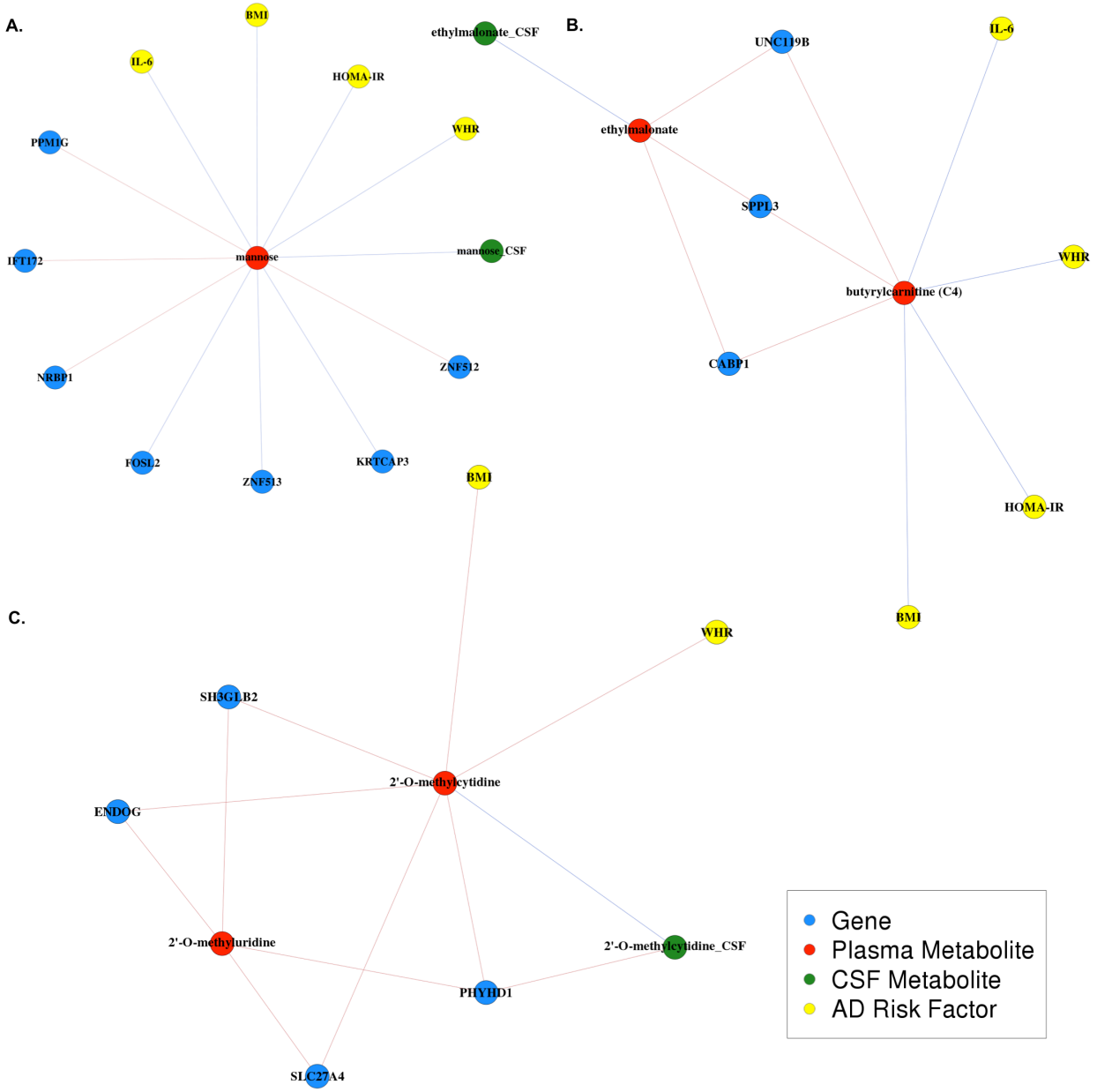
Appendix C7. Description of each of the 908 inter-omic correlations.

Due to the size of the table, it is provided as an excel spreadsheet.



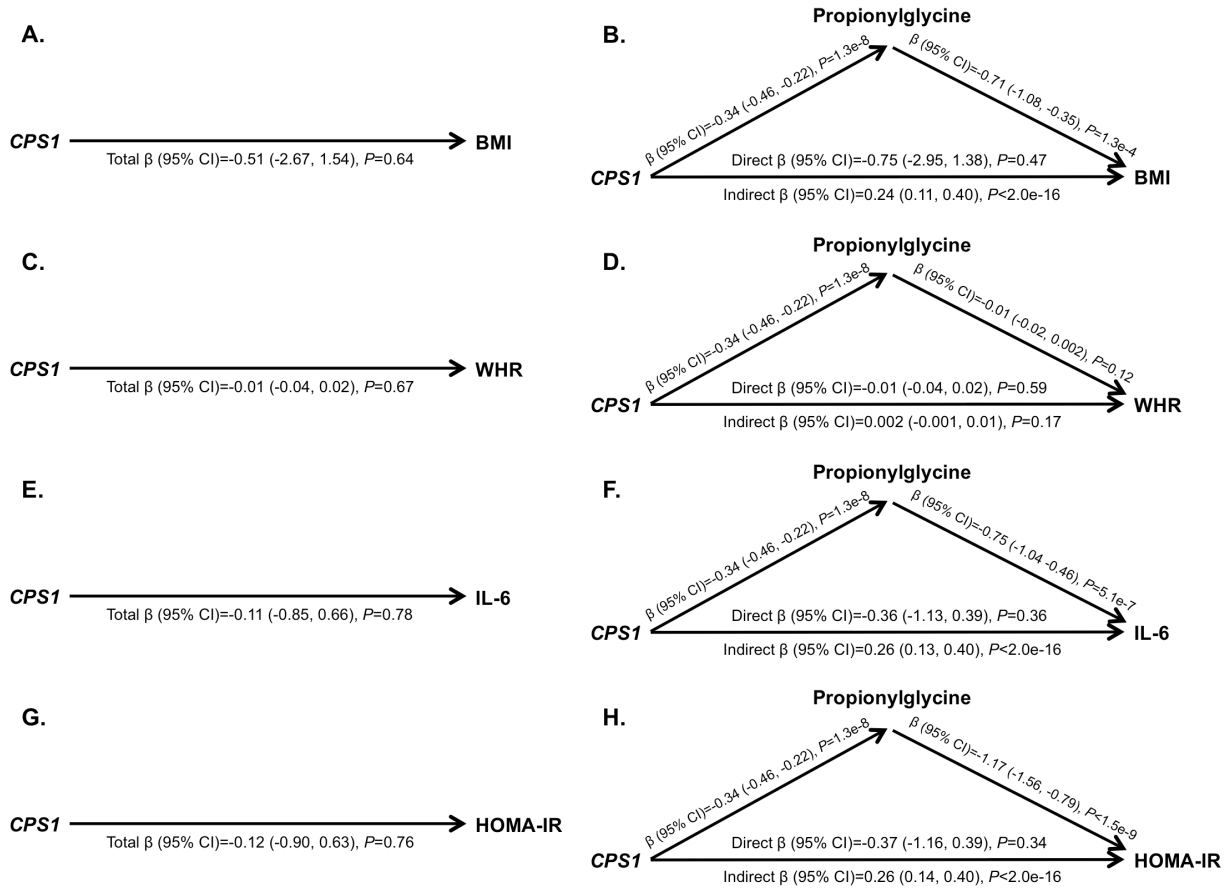
Appendix C8. Cardiovascular disease and diabetes community.

This is the largest inter-omics community, containing 680 edges and 289 nodes. Red edges indicate negative correlations and blue edges indicate positive correlations.



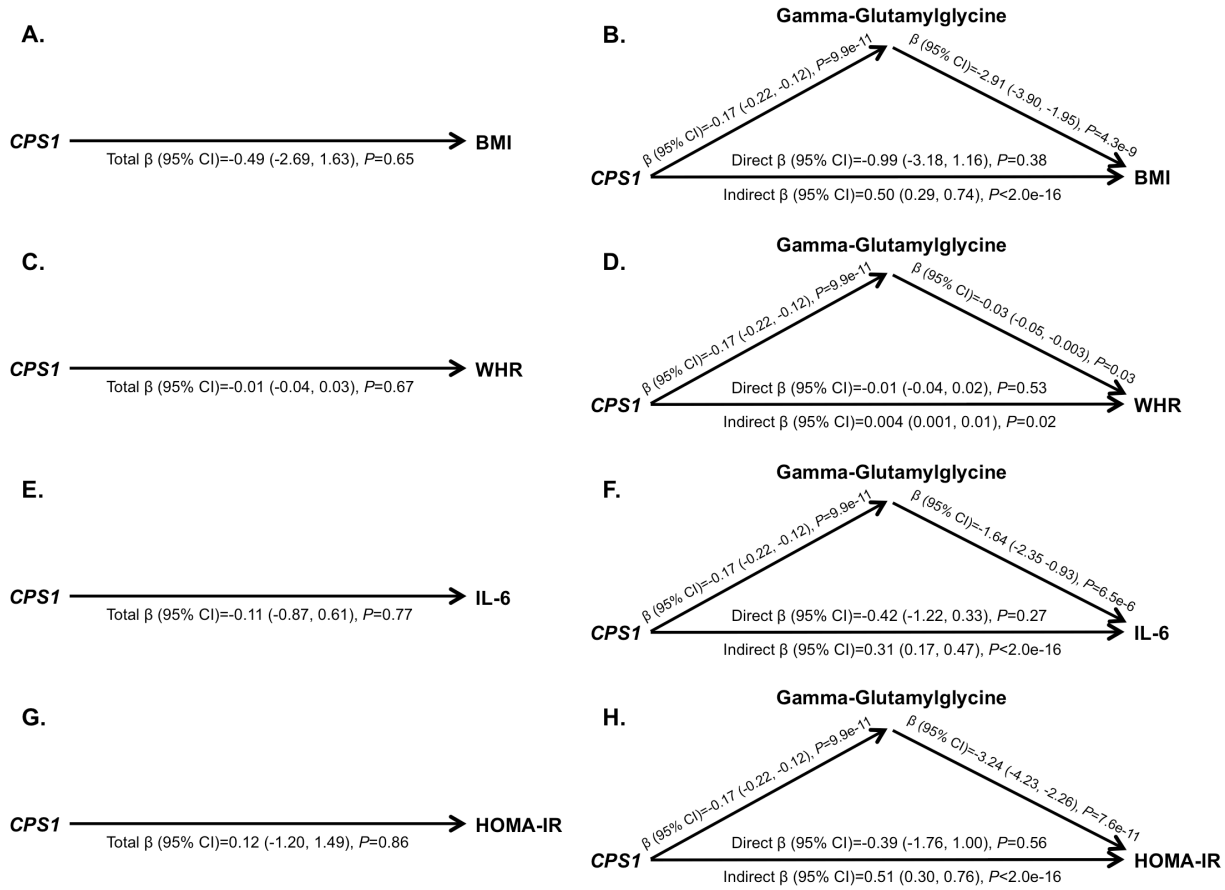
Appendix C9. Intricate sub-networks linking gene expression to cardiovascular and diabetes risk factors.

A. Plasma mannose connected seven genes to BMI, WHR, IL-6, and HOMA. B. Plasma butyrylcarnitine (C4) connected three genes to BMI, WHR, IL-6, and HOMA-IR. C. Plasma 2'O'methylcytidine connected four genes to BMI and WHR. Red edges indicate negative correlations and blue edges indicate positive correlations.



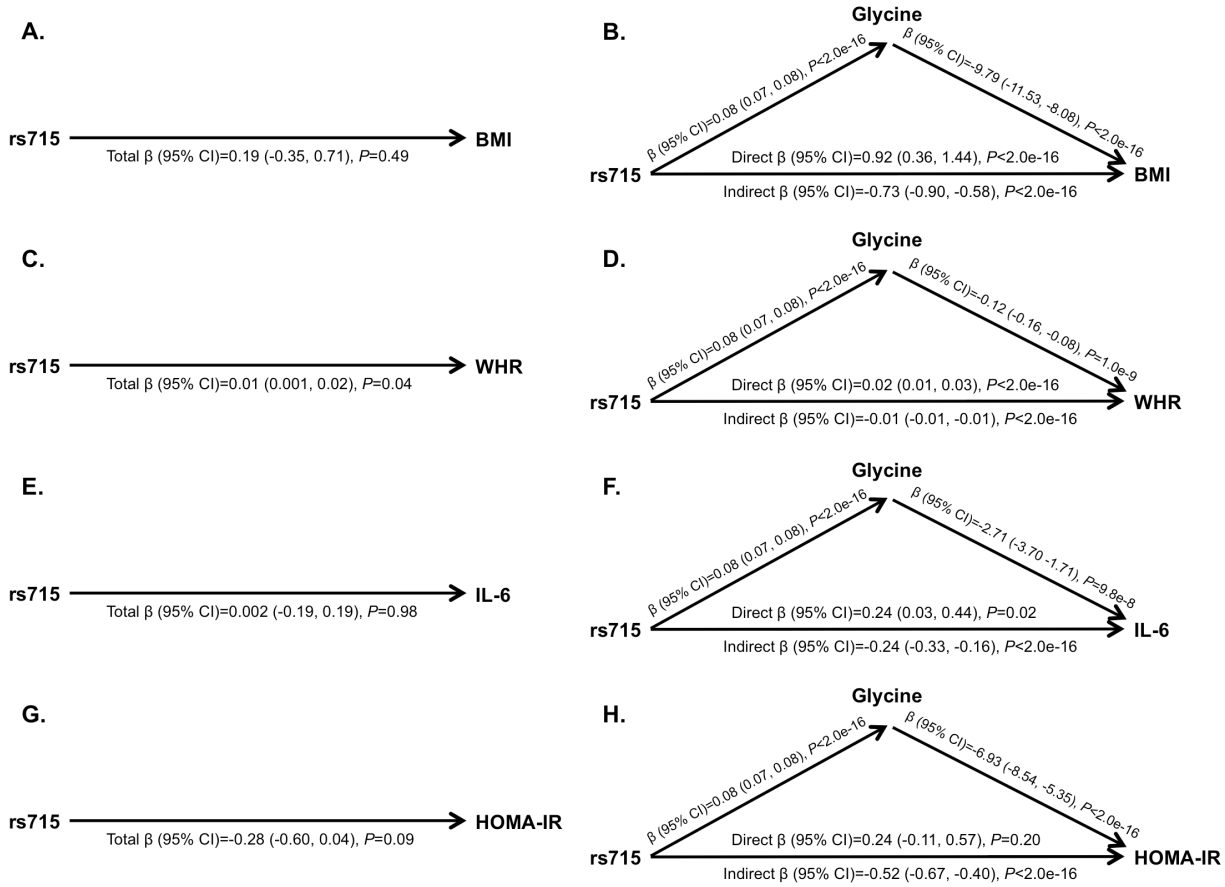
Appendix C10. Mediation analyses to assess whether plasma propionylglycine mediates the relationships between imputed *CPS1* expression, BMI, WHR, IL-6, and HOMA-IR.

A. Total effect of *CPS1* on BMI. B. Direct and indirect effects of *CPS1* on BMI. C. Total effect of *CPS1* on WHR. D. Direct and indirect effects of *CPS1* on WHR. E. Total effect of *CPS1* on IL-6. F. Direct and indirect effects of *CPS1* on IL-6. G. Total effect of *CPS1* on HOMA-IR. H. Direct and indirect effects of *CPS1* on HOMA-IR. All models adjusted for age and sex; models including *CPS1* additionally adjusted for the first four PCs; models that included propionylglycine additionally adjusted for cholesterol lowering medication use and sample storage time.



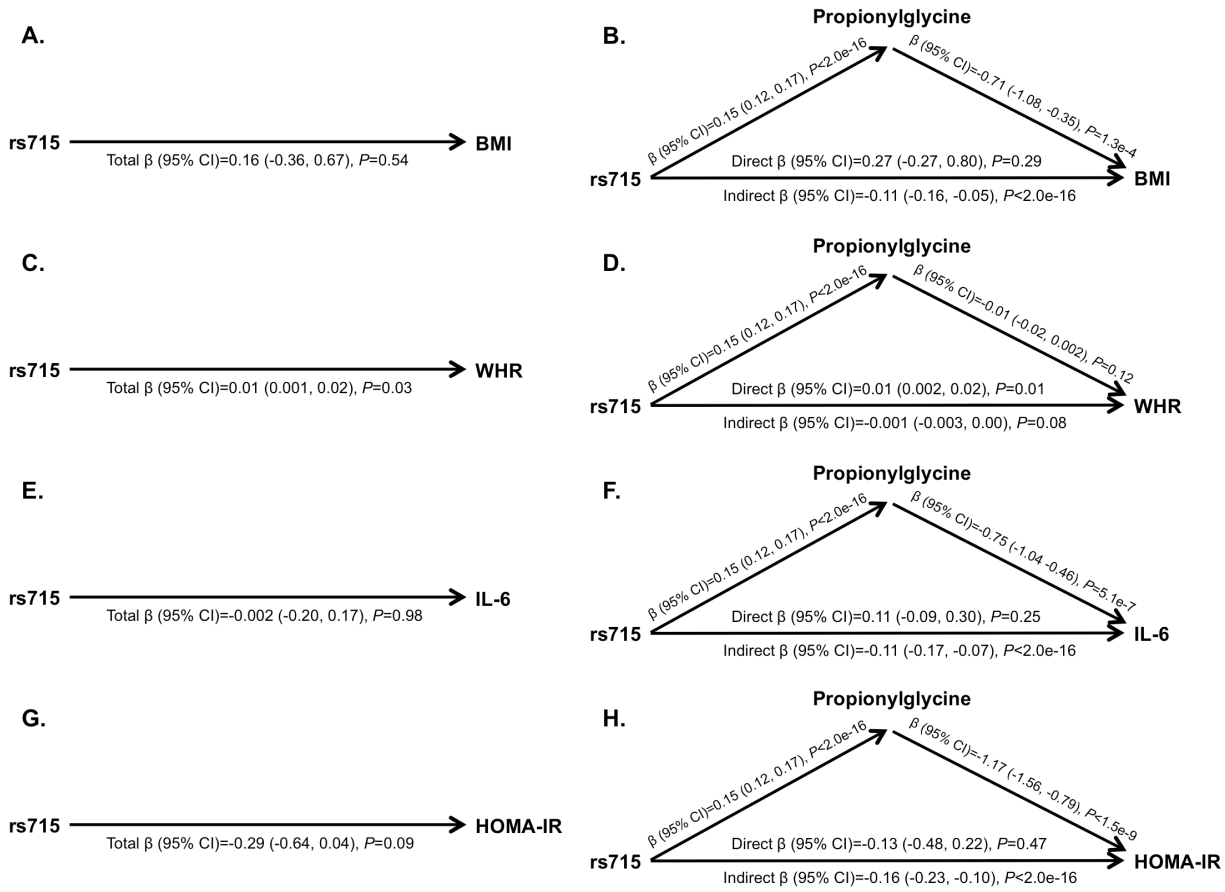
Appendix C11. Mediation analyses to assess whether plasma gamma-glutamylglycine mediates the relationships between imputed *CPS1* expression, BMI, WHR, IL-6, and HOMA-IR.

A. Total effect of *CPS1* on BMI. B. Direct and indirect effects of *CPS1* on BMI. C. Total effect of *CPS1* on WHR. D. Direct and indirect effects of *CPS1* on WHR. E. Total effect of *CPS1* on IL-6. F. Direct and indirect effects of *CPS1* on IL-6. G. Total effect of *CPS1* on HOMA-IR. H. Direct and indirect effects of *CPS1* on HOMA-IR. All models adjusted for age and sex; models including *CPS1* additionally adjusted for the first four PCs; models that included gamma-glutamylglycine additionally adjusted for cholesterol lowering medication use and sample storage time.



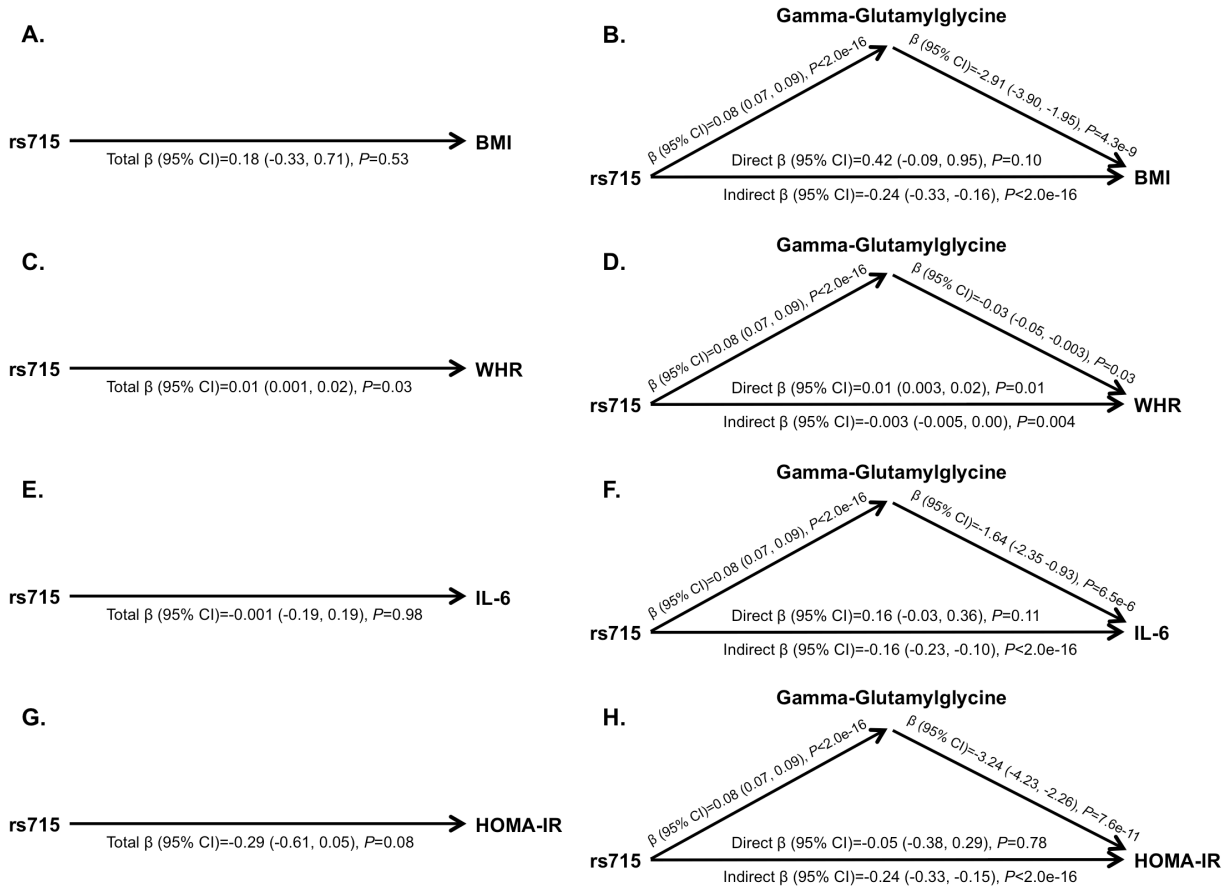
Appendix C12. Mediation analyses to assess whether plasma glycine mediates the relationships between the minor C allele of *CPS1* variant rs715, BMI, WHR, IL-6, and HOMA-IR.

A. Total effect of rs715 on BMI. B. Direct and indirect effects of rs715 on BMI. C. Total effect of rs715 on WHR. D. Direct and indirect effects of rs715 on WHR. E. Total effect of rs715 on IL-6. F. Direct and indirect effects of rs715 on IL-6. G. Total effect of rs715 on HOMA-IR. H. Direct and indirect effects of rs715 on HOMA-IR. All models adjusted for age and sex; models including rs715 additionally adjusted for the first four PCs; models that included glycine additionally adjusted for cholesterol lowering medication use and sample storage time.



Appendix C13. Mediation analyses to assess whether plasma propionylglycine mediates the relationships between the minor C allele of *CPS1* variant rs715, BMI, WHR, IL-6, and HOMA-IR.

A. Total effect of rs715 on BMI. B. Direct and indirect effects of rs715 on BMI. C. Total effect of rs715 on WHR. D. Direct and indirect effects of rs715 on WHR. E. Total effect of rs715 on IL-6. F. Direct and indirect effects of rs715 on IL-6. G. Total effect of rs715 on HOMA-IR. H. Direct and indirect effects of rs715 on HOMA-IR. All models adjusted for age and sex; models including rs715 additionally adjusted for the first four PCs; models that included propionylglycine additionally adjusted for cholesterol lowering medication use and sample storage time.



Appendix C14. Mediation analyses to assess whether plasma gamma-glutamylglycine mediates the relationships between the minor C allele of *CPS1* variant rs715, BMI, WHR, IL-6, and HOMA-IR.

A. Total effect of rs715 on BMI. B. Direct and indirect effects of rs715 on BMI. C. Total effect of rs715 on WHR. D. Direct and indirect effects of rs715 on WHR. E. Total effect of rs715 on IL-6. F. Direct and indirect effects of rs715 on IL-6. G. Total effect of rs715 on HOMA-IR. H. Direct and indirect effects of rs715 on HOMA-IR. All models adjusted for age and sex; models including rs715 additionally adjusted for the first four PCs; models that included gamma-glutamylglycine additionally adjusted for cholesterol lowering medication use and sample storage time.