Integrated analysis of genomics, longitudinal metabolomics, and Alzheimer's risk factors among 1,111 cohort participants

Burcu F. Darst,[1,2]* Qiongshi Lu,[1,3] Sterling C. Johnson,[1,4,5,6] Corinne D. Engelman[1,2,4,6]

[1] University of Wisconsin, Madison, WI, 53706, USA

[2] Department of Population Health Sciences, University of Wisconsin School of Medicine and Public Health, Madison, WI, 53726, USA

[3] Department of Biostatistics & Medical Informatics, Madison, WI, 53792, USA

[4] Wisconsin Alzheimer's Institute, University of Wisconsin School of Medicine and Public Health, Madison, WI, 53719, USA

[5] Geriatric Research Education and Clinical Center, Wm. S. Middleton Memorial VA Hospital, Madison, WI, 53705, USA

[6] Alzheimer's Disease Research Center, University of Wisconsin School of Medicine and Public Health, Madison, WI, 53792, USA

* Correspondence:

Burcu F. Darst

323-442-0078

bdarst@wisc.edu

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 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. Further, 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.

Introduction

Genome-wide association studies (GWAS) have identified tens of thousands of single nucleotide polymorphism (SNP)-trait associations(1). However, these variants tend to have very small effect sizes and typically explain a small portion of trait heritability. Late onset 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(2). 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(3-8). However, studies have been able to show that the use of multiple omics data types is more predictive than single data types(5, 9). A recent study with dense longitudinal omics data displayed the utility of integrating such data with regards to personalized medicine(10). 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(11), obesity(12), and cardiovascular disease(13), 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(14) and many metabolites have a low heritability(15, 16), 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(17). 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.

Materials and 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(18, 19). Participants included in this analysis had genetic ancestry that was primarily of European descent, had both genomic and metabolomic data available, and up to seventeen 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(20), 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 $A\beta_{42}$) were 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)(21); details are outlined in the Supplemental Note. 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

134 missing a median of 11.7% plasma metabolites, while plasma metabolites were missing in a
135 median of 1.2% of samples.

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

140 Missing plasma and CSF metabolite values were imputed to the lowest level of detection
141 for each metabolite(22). Metabolite values were median-scaled and log-transformed to normalize
142 metabolite distributions(23). If a participant reported that they did not fast or withhold
143 medications and caffeine for at least eight hours prior to the blood draw, the plasma sample was
144 excluded from analyses (159 plasma samples), leaving 1,097 plasma metabolites among 2,189
145 plasma samples (1,111 individuals) for analyses. Similarly, if a participant reported that they did
146 not fast for at least eight hours prior to the LP, the CSF sample was excluded from analyses (4
147 CSF samples), leaving 364 CSF metabolites among 346 CSF samples (155 individuals) for
148 analyses.

149 **CSF and plasma metabolite correlations**

150 A total of 326 metabolites were captured in both CSF and plasma. The correlations of
151 these metabolites between tissue types were calculated using the Pearson correlation coefficient.
152 In order to reduce variability due to the time interval between plasma and CSF sample collection,
153 correlations were based on 141 pairs of plasma and CSF samples that were collected within a
154 timespan of four months of each other. After removing these samples, plasma and CSF samples
155 were collected a median of 27 days apart.

156 **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 (Figure S1). 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(24). 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(25). First, kinship coefficients (KCs) were calculated between all pairs of individuals using genomic data with the Kinship-based Inference for Gwas (KING)-robust method(26). 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(27). 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(28), using the Haplotype Reference Consortium (HRC) v. r1.1 2016(29) as the reference panel and Eagle2 v2.3(30) for phasing. Prior to imputation, the HRC Imputation Checking Tool(31) 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(32) and VCFtools v0.1.14(33). 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(34) with the Depression Genes and Networks reference dataset(35), 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 analytic approach we used for our network analysis was similar to that of Price et al., 2017(10). A total of 12,856 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(36), were used to adjust each variable for age and sex and included a random intercept for individual to account for repeated measures and family to account for sibling relationships. 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(37). 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(38).

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 gene-metabolite correlations with the most consistent biological support from the GWAS catalog(1) (www.ebi.ac.uk/gwas, date accessed: May 9, 2018), 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(36), 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 an 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(39). To determine whether this difference was significant, standard errors and *P*-values were estimated using the quasi-Bayesian Monte Carlo method with 1,000 simulations. Because the mediation package can only handle mixed models with one random effect, the mediation analysis was run with models 1 and 4

249 excluding the random effect for family. As a sensitivity analysis, the mediation analysis was
250 rerun limiting models 1 and 4 to unrelated individuals (n=898 with 1,774 observations). A fifth
251 linear mixed model was used to assess interactions by adding a gene*metabolite interaction term
252 to model 4. Model 5 did not use the mediation package and was thus able to include random
253 intercepts for both within-individual repeated measures and within-family relationships. All
254 models including a gene had covariates for age, sex, and the first four PCs, while models
255 including a metabolite had covariates for age, sex, cholesterol lowering medication use, and
256 sample storage time.
257
258 Results
259 **Participants**
260 A total of 1,111 WRAP participants had both genomic and plasma metabolomic data. At
261 baseline, 68.9% of participants were female and participants were 61.0 years old with a
262 bachelor's degree, on average (Table 2). Participants each had 1,097 plasma metabolites
263 available for analyses, 347 (31.6%) of which were of unknown chemical structure, whole blood
264 gene expression for 11,376 genes, and up to 17 AD risk factors. A subset of 155 individuals also
265 had 364 CSF metabolites available for analyses, 56 (15.4%) of which were of unknown chemical
266 structure. Participants with CSF metabolomic data had similar characteristics as the full sample
267 (Table 2). Properties of each plasma and CSF metabolite, such as biochemical name, super
268 pathway, and sub pathway are described in Table S1, and numbers of metabolites within each
269 super pathway are summarized in Table S2.
270 **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 Table S3.

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 (Figure S2). 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 Figure S3), and its corresponding community partitions are shown in Figure S4. 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 Table S4.

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

294 equivalents (METs), diastolic blood pressure (DBP), and systolic blood pressure (SBP) (Figure
295 S5). Expression levels of these eight genes were all indirectly linked to AD risk factors within
296 this community through plasma metabolites. *CPS1* expression levels were negatively correlated
297 with plasma gamma-glutamylglycine, proprionylglycine, and glycine levels, all of which were
298 negatively correlated with BMI, WHR, IL-6, and/or HOMA-IR (Figure 3). *TMEM229B* and
299 *PLEKHH1* were both negatively correlated with two glycerophosphatidylcholines (1-(1-enyl-
300 palmitoyl)-2-palmitoleoyl-GPC (P-16:0/16:1) and 1-(1-enyl-palmitoyl)-2-palmitoyl-GPC (P-
301 16:0/16:0)), which were also negatively correlated with BMI, WHR, and/or HOMA-IR.
302 *NAALAD2* was negatively correlated with an amino acid beta-citrylglutamate, which was
303 positively correlated with BMI, WHR, IL-6, and HOMA-IR. *ZNF655* and *ZKSCAN1* were both
304 positively correlated with X-12063, which was also positively correlated with BMI, WHR, and
305 HOMA-IR. *CHRNA5* was positively correlated with 5-hydroxylysine, which was positively
306 correlated with BMI, WHR, IL-6, and HOMA-IR, and negatively correlated with METs. *ARVCF*
307 was negatively correlated with X-11593, which was positively correlated with BMI, IL-6, and
308 HOMA-IR.

309 Several genes outside of the cardiovascular and diabetes community were indirectly
310 linked to AD risk factors within this community. Gene expression of *FOSL2*, *KRTCAP3*, and
311 *ZNF513* were positively correlated, while *IFT172*, *NRBP1*, *PPM1G*, and *ZNF512* were
312 negatively correlated, with levels of plasma mannose, a carbohydrate that was positively
313 correlated with BMI, WHR, IL-6, and HOMA-IR (Figure S6A). *CABP1*, *SPPL3*, and *UNC119B*
314 expression levels were negatively correlated with plasma butyrylcarnitine (C4), which was
315 positively correlated with BMI, WHR, IL-6, and HOMA-IR (Figure S6B). *SLC27A4*, *PHYHD1*,
316 *ENDOG*, and *SH3GLB2* expression levels were negatively correlated with plasma 2'-O-

317 methyluridine and 2'-O-methylcytidine levels, both nucleotides involved in pyrimidine
318 metabolism, and the latter nucleotide is also negatively correlated with BMI and WHR (Figure
319 S6C). *PHYHD1* was also negatively correlated with CSF levels of 2'-O-methylcytidine.
320 The only correlations identified among the CSF biomarkers (*i.e.*, amyloid and tau) are
321 shown in Figure 4. Higher CSF T-tau and P-tau levels were correlated with higher levels of 38
322 CSF metabolites, collectively. These metabolites included 13 lipids (six phosphatidylcholines,
323 two lysophosphatidylcholines, five sphingolipids, and cholesterol), seven amino acids, five
324 carbohydrates, one nucleotide, one energy metabolite, one cofactor and vitamin metabolite, one
325 xenobiotic, and nine unknown metabolites. However, none of the CSF amyloid biomarkers were
326 correlated with CSF metabolites. We investigated how much of the variance of T-tau and P-tau
327 could be explained by these metabolites with linear mixed models that included random
328 intercepts for within-subject repeated measures and within-family relationships, using the R^2
329 statistic for mixed models as defined by Edwards et al., 2008(40) and implemented in the
330 r2glmm R package. After removing the variation explained by age and sex, the 37 metabolites
331 correlated with T-tau explained 60.7% of the variation of T-tau, while the 35 metabolites
332 correlated with P-tau explained 64.0% of the variation of P-tau.
333 **Targeted mediation and interaction analyses**
334 Targeted mediation and interaction analyses were focused on a particular pathway
335 identified within the large cardiovascular and diabetes community involving *CPS1*, glycine
336 plasma metabolites (glycine, proprionylglycine, and gamma-glutamylglycine), BMI, WHR, IL-6,
337 and HOMA-IR. Associations between *CPS1* variants and glycine have been reported in at least
338 nine studies(15, 16, 41-47), more than any of the other gene-metabolite associations identified in
339 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(45, 48-55). 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 (Figure 5A and Figure5B), WHR (Figure 5C and Figure 5D), IL-6 (Figure 5E and Figure 5F), and HOMA-IR (Figure 5G and Figure 5H). The total effect of *CPS1* was null for each of these three 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(56). 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 Figure S7 and Figure S8. We did not identify any interactions between *CPS1* and the three glycine metabolites 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

363 levels were directly correlated with AD risk factors, there were many instances of genes being
364 indirectly correlated with AD risk factors though metabolites. Building on one such instance, we
365 found that glycine mediated the pathway between *CPS1* expression and cardiovascular and
366 diabetes risk factors. This suggests that our results may have generated many valid hypotheses
367 that warrant further investigation. We also found that correlations between plasma and CSF
368 metabolites ranged widely but typically had low correlations. This could suggest that most
369 plasma metabolites are not representative of certain metabolic changes occurring in the brain,
370 although we cannot rule out the possibility that the low average correlation is, at least partially,
371 due to the time difference between the plasma and CSF sample collection.
372 The low correlation we observed between plasma and CSF metabolite levels could be
373 related to ~98% of small molecules not being able to pass the blood-brain barrier (BBB)(57).
374 Cholesterol is an example of a lipid metabolite that typically cannot pass the BBB(58), and was
375 not correlated between tissues (r=-0.07). On the other hand, caffeine (a xenobiotic) readily
376 crosses the BBB(59) and it was highly correlated between tissues (r=0.81), as was 5-
377 acetylamino-6-amino-3-methyluracil (r=0.82), which is a caffeine metabolite, and theophylline
378 (r=0.82), which is structurally and pharmacologically similar to caffeine. This could contribute to
379 lipids having the weakest average correlation and xenobiotics having the strongest average
380 correlation between plasma and CSF tissues. However, it is important to note that metabolites
381 within a given pathway can vary widely from each other and should be considered on an
382 individual basis, accordingly, as the averages presented here may not reflect a particular
383 metabolite's unique properties. The hypothesis about plasma and CSF differing due to the BBB
384 is also supported by the only correlations in the network analysis involving CSF biomarkers (*i.e.*,
385 tau) being with CSF metabolites, although we cannot rule out the possibility that this correlation

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

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

398 One advantage of using imputed gene expression data is that it only represents the
399 genetically regulated component of gene expression, reducing the risk of confounding due to
400 environmental factors and reverse causality in mediation analyses. We found that glycine
401 mediated the relationship between *CPS1* and BMI, WHR, IL-6, and HOMA-IR, such that lower
402 *CPS1* expression was associated with higher levels of glycine, which were associated with lower
403 BMI, WHR, IL-6, and HOMA-IR. Relationships between *CPS1*, glycine, and cardiovascular risk
404 factors have been hypothesized recently, but not clearly defined(43, 60). The *CPS1* (Carbamoyl-
405 Phosphate Synthase 1) gene encodes for a mitochondrial enzyme that catalyzes the first step of
406 the hepatic urea cycle by synthesizing carbamoyl phosphate from ammonia, bicarbonate, and two
407 molecules of ATP, and is important for removal of urea from cells(61). Notably, all genes
408 encoding enzymes involved in the urea cycle are expressed in the brain, including *CPS1*(62), and

409 levels of enzymes and metabolic intermediates involved in the urea cycle are altered in AD
410 patients(63). *CPS1* variants have been linked to CPS1 deficiency(61), neonatal pulmonary
411 hypertension(64), vascular function(65), traits related to blood clotting, such as fibrinogen levels
412 and platelet count(66-69), homocysteine levels(70-73), HDL cholesterol(74), kidney function
413 and disease(75-78), AD(79), and BMI(80, 81). Higher adipose tissue expression of *CPS1* has
414 been associated with detrimental traits, including weight gain(60). At least nine studies have
415 reported associations between *CPS1* variants and glycine(15, 16, 41-47) and others have reported
416 associations with betaine, a derivative of glycine(15, 16, 82). Glycine is a common amino acid
417 involved in the production of DNA, phospholipids, and collagen, and in the release of energy.
418 Previous studies have identified negative correlations between glycine and cardiovascular and
419 diabetes risk factors such as BMI, waist circumference, HOMA-IR, obesity and visceral obesity,
420 subcutaneous and visceral fat area, hypertension, and acute myocardial infarction(45, 48-55).
421 These previous findings are in the same direction as our findings and are highly supportive of the
422 biological relevance of our results, which lead us to hypothesize that the *CPS1*-cardiovascular
423 risk pathway is linked through the mediation of glycine.

424 One particular *CPS1* variant, rs715, has been linked to urine and blood glycine levels(15,
425 16, 43-45), blood levels of betaine(15, 16, 82), blood levels of fibrinogen(66, 67), and BMI(80).
426 This is a common variant, with a MAF=0.27 based on 62,784 whole genome sequences from
427 Trans-Omics for Precision Medicine (TOPMed)(83). The minor C allele of rs715 decreases
428 *CPS1* expression(82). To further test our findings, we conducted additional mediation analyses
429 using this variant and found highly consistent results, suggesting that having one or two minor
430 alleles of rs715 (which decreases *CPS1* expression) increases levels of the three glycine plasma

metabolites, which decreases BMI, WHR, IL-6, and HOMA-IR (Figures S9-S11). 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(84, 85). 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(86). 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. However, it had limited success for traits that follow complex inheritance patterns(86). 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 offer insight into the onset of AD and risk factors associated with its onset. Our research has generated many promising hypotheses that could drive subsequent experimental investigations and potentially offer clinicians and researchers new insights regarding the development of tau tangles. As the generation of omics data accelerates

477 across investigations of a variety of research fields, continued efforts to navigate statistical and
478 computational issues will be critical. The work presented here represents early efforts to integrate
479 omics data, but much more research is needed to identify the most effective means of doing so
480 and thereby maximize the utility of such rich sources of data. The success of precision medicine
481 is heavily reliant on the advancement of computational biology and the ability to translate
482 millions of biological data points into individual clinical implications.
483
484 Funding
485 BFD was supported by a National Library of Medicine training grant to the Computation and
486 Informatics in Biology and Medicine Training Program [grant number NLM 5T15LM007359].
487 This research was also supported by the National Institutes of Health [grant numbers
488 R01AG054047, R01AG27161, UL1TR000427, and P2C HD047873]; Helen Bader Foundation;
489 Northwestern Mutual Foundation; Extendicare Foundation; and State of Wisconsin. QL was
490 supported by the Clinical and Translational Science Award program, through the NIH National
491 Center for Advancing Translational Sciences [grant number UL1TR000427].
492
493 Acknowledgements
494 The authors thank the University of Wisconsin Madison Biotechnology Center Gene Expression
495 Center for providing Illumina Infinium genotyping services. We especially thank the WRAP
496 participants.
497
498 Conflict of Interests
499 The authors declare no competing interests.

500

501 References

502 1 MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H.,
503 McMahon, A., Milano, A., Morales, J. *et al.* (2017) The new NHGRI-EBI Catalog of published
504 genome-wide association studies (GWAS Catalog). *Nucleic acids research*, **45**, D896-D901.
505 2 Ridge, P.G., Hoyt, K.B., Boehme, K., Mukherjee, S., Crane, P.K., Haines, J.L., Mayeux,
506 R., Farrer, L.A., Pericak-Vance, M.A., Schellenberg, G.D. *et al.* (2016) Assessment of the
507 genetic variance of late-onset Alzheimer's disease. *Neurobiology of aging*, **41**, 200 e213-200
508 e220.
509 3 Ritchie, M.D., Holzinger, E.R., Li, R., Pendergrass, S.A. and Kim, D. (2015) Methods of
510 integrating data to uncover genotype-phenotype interactions. *Nature reviews. Genetics*, **16**, 85-
511 97.
512 4 Huang, S., Chaudhary, K. and Garmire, L.X. (2017) More Is Better: Recent Progress in
513 Multi-Omics Data Integration Methods. *Frontiers in genetics*, **8**, 84.
514 5 Buescher, J.M. and Driggers, E.M. (2016) Integration of omics: more than the sum of its
515 parts. *Cancer & metabolism*, **4**, 4.
516 6 Bersanelli, M., Mosca, E., Remondini, D., Giampieri, E., Sala, C., Castellani, G. and
517 Milanesi, L. (2016) Methods for the integration of multi-omics data: mathematical aspects. *BMC*
518 *bioinformatics*, **17 Suppl 2**, 15.
519 7 Lopez de Maturana, E., Pineda, S., Brand, A., Van Steen, K. and Malats, N. (2016)
520 Toward the integration of Omics data in epidemiological studies: still a "long and winding road".
521 *Genetic epidemiology*, **40**, 558-569.

8 Gligorijevic, V. and Przulj, N. (2015) Methods for biological data integration: perspectives and challenges. *Journal of the Royal Society, Interface*, **12**.

9 Mankoo, P.K., Shen, R., Schultz, N., Levine, D.A. and Sander, C. (2011) Time to recurrence and survival in serous ovarian tumors predicted from integrated genomic profiles. *PloS one*, **6**, e24709.

10 Price, N.D., Magis, A.T., Earls, J.C., Glusman, G., Levy, R., Lausted, C., McDonald, D.T., Kusebauch, U., Moss, C.L., Zhou, Y. *et al.* (2017) A wellness study of 108 individuals using personal, dense, dynamic data clouds. *Nat Biotechnol*, **35**, 747-756.

11 Enche Ady, C.N.A., Lim, S.M., Teh, L.K., Salleh, M.Z., Chin, A.V., Tan, M.P., Poi, P.J.H., Kamaruzzaman, S.B., Abdul Majeed, A.B. and Ramasamy, K. (2017) Metabolomic-guided discovery of Alzheimer's disease biomarkers from body fluid. *Journal of neuroscience research*, **95**, 2005-2024.

12 Zhang, A.H., Sun, H. and Wang, X.J. (2017) Emerging role and recent applications of metabolomics biomarkers in obesity disease research. *Rsc Adv*, **7**, 14966-14973.

13 Ussher, J.R., Elmariah, S., Gerszten, R.E. and Dyck, J.R. (2016) The Emerging Role of Metabolomics in the Diagnosis and Prognosis of Cardiovascular Disease. *Journal of the American College of Cardiology*, **68**, 2850-2870.

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

15 Long, T., Hicks, M., Yu, H.C., Biggs, W.H., Kirkness, E.F., Menni, C., Zierer, J., Small, K.S., Mangino, M., Messier, H. *et al.* (2017) Whole-genome sequencing identifies common-to-rare variants associated with human blood metabolites. *Nature genetics*, **49**, 568-578.

544 16 Shin, S.Y., Fauman, E.B., Petersen, A.K., Krumsiek, J., Santos, R., Huang, J., Arnold,
545 M., Erte, I., Forgetta, V., Yang, T.P. *et al.* (2014) An atlas of genetic influences on human blood
546 metabolites. *Nature genetics*, **46**, 543-550.
547 17 Jack, C.R., Jr., Lowe, V.J., Weigand, S.D., Wiste, H.J., Senjem, M.L., Knopman, D.S.,
548 Shiung, M.M., Gunter, J.L., Boeve, B.F., Kemp, B.J. *et al.* (2009) Serial PIB and MRI in normal,
549 mild cognitive impairment and Alzheimer's disease: implications for sequence of pathological
550 events in Alzheimer's disease. *Brain : a journal of neurology*, **132**, 1355-1365.
551 18 Sager, M.A., Hermann, B. and La Rue, A. (2005) Middle-aged children of persons with
552 Alzheimer's disease: APOE genotypes and cognitive function in the Wisconsin Registry for
553 Alzheimer's Prevention. *Journal of geriatric psychiatry and neurology*, **18**, 245-249.
554 19 Johnson, S.C., Koscik, R.L., Jonaitis, E.M., Clark, L.R., Mueller, K.D., Berman, S.E.,
555 Bendlin, B.B., Engelman, C.D., Okonkwo, O.C., Hogan, K.J. *et al.* (2018) The Wisconsin
556 Registry for Alzheimer's Prevention: A review of findings and current directions. *Alzheimers*
557 *Dement (Amst)*, **10**, 130-142.
558 20 Darst, B.F., Koscik, R.L., Racine, A.M., Oh, J.M., Krause, R.A., Carlsson, C.M.,
559 Zetterberg, H., Blennow, K., Christian, B.T., Bendlin, B.B. *et al.* (2017) Pathway-Specific
560 Polygenic Risk Scores as Predictors of Amyloid-beta Deposition and Cognitive Function in a
561 Sample at Increased Risk for Alzheimer's Disease. *Journal of Alzheimer's disease : JAD*, **55**,
562 473-484.
563 21 Evans, A.M., Bridgewater, B.R., Liu, Q., Mitchell, M.W., Robinson, R.J., Dai, H.,
564 Stewart, S.J., DeHaven, C.D. and Miller, L.A.D. (2014) High Resolution Mass Spectrometry
565 Improves Data Quantity and Quality as Compared to Unit Mass Resolution Mass Spectrometry

566 in High-Throughput Profiling Metabolomics. *Metabolomics : Official journal of the*
567 *Metabolomic Society*, **4**, 1-7.
568 22 Menni, C., Kastenmuller, G., Petersen, A.K., Bell, J.T., Psatha, M., Tsai, P.C., Gieger, C.,
569 Schulz, H., Erte, I., John, S. *et al.* (2013) Metabolomic markers reveal novel pathways of ageing
570 and early development in human populations. *International journal of epidemiology*, **42**, 1111-
571 1119.
572 23 van den Berg, R.A., Hoefsloot, H.C., Westerhuis, J.A., Smilde, A.K. and van der Werf,
573 M.J. (2006) Centering, scaling, and transformations: improving the biological information
574 content of metabolomics data. *BMC genomics*, **7**, 142.
575 24 Conomos, M.P., Miller, M.B. and Thornton, T.A. (2015) Robust inference of population
576 structure for ancestry prediction and correction of stratification in the presence of relatedness.
577 *Genetic epidemiology*, **39**, 276-293.
578 25 Genomes Project, C., Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang,
579 H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A. *et al.* (2015) A global reference
580 for human genetic variation. *Nature*, **526**, 68-74.
581 26 Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M. and Chen, W.M.
582 (2010) Robust relationship inference in genome-wide association studies. *Bioinformatics*, **26**,
583 2867-2873.
584 27 Conomos, M.P., Reiner, A.P., Weir, B.S. and Thornton, T.A. (2016) Model-free
585 Estimation of Recent Genetic Relatedness. *American journal of human genetics*, **98**, 127-148.
586 28 Das, S., Forer, L., Schonherr, S., Sidore, C., Locke, A.E., Kwong, A., Vrieze, S.I., Chew,
587 E.Y., Levy, S., McGue, M. *et al.* (2016) Next-generation genotype imputation service and
588 methods. *Nature genetics*, **48**, 1284-1287.

589 29 McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang,
590 H.M., Fuchsberger, C., Danecek, P., Sharp, K. *et al.* (2016) A reference panel of 64,976
591 haplotypes for genotype imputation. *Nature genetics*, **48**, 1279-1283.
592 30 Loh, P.R., Danecek, P., Palamara, P.F., Fuchsberger, C., Y, A.R., H, K.F., Schoenherr,
593 S., Forer, L., McCarthy, S., Abecasis, G.R. *et al.* (2016) Reference-based phasing using the
594 Haplotype Reference Consortium panel. *Nature genetics*, **48**, 1443-1448.
595 31 Rayner, N.W., Robertson, N., Mahajan, A. and McCarthy, M.I. (2016), In *The American*
596 *Society of Human Genetics*, Vancouver, Canada, in press.
597 32 Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M. and Lee, J.J. (2015)
598 Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*, **4**,
599 7.
600 33 Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A.,
601 Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T. *et al.* (2011) The variant call format and
602 VCFtools. *Bioinformatics*, **27**, 2156-2158.
603 34 Gamazon, E.R., Wheeler, H.E., Shah, K.P., Mozaffari, S.V., Aquino-Michaels, K.,
604 Carroll, R.J., Eyler, A.E., Denny, J.C., Consortium, G.T., Nicolae, D.L. *et al.* (2015) A gene-
605 based association method for mapping traits using reference transcriptome data. *Nature genetics*,
606 **47**, 1091-1098.
607 35 Battle, A., Mostafavi, S., Zhu, X., Potash, J.B., Weissman, M.M., McCormick, C.,
608 Haudenschild, C.D., Beckman, K.B., Shi, J., Mei, R. *et al.* (2014) Characterizing the genetic
609 basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome research*,
610 **24**, 14-24.

611 36 Bates, D., Machler, M., Bolker, B.M. and Walker, S.C. (2015) Fitting Linear Mixed-
612 Effects Models Using lme4. *Journal of Statistical Software*, **67**, 1-48.
613 37 Csardi, G. and Nepusz, T. (2006) The igraph software package for complex network
614 research. *InterJournal*, **Complex Systems**, 1695.
615 38 Clauset, A., Newman, M.E.J. and Moore, C. (2004) Finding community structure in very
616 large networks. *Phys Rev E*, **70**.
617 39 Imai, K., Keele, L. and Tingley, D. (2010) A general approach to causal mediation
618 analysis. *Psychological methods*, **15**, 309-334.
619 40 Edwards, L.J., Muller, K.E., Wolfinger, R.D., Qaqish, B.F. and Schabenberger, O. (2008)
620 An R2 statistic for fixed effects in the linear mixed model. *Statistics in medicine*, **27**, 6137-6157.
621 41 Kettunen, J., Demirkan, A., Wurtz, P., Draisma, H.H., Haller, T., Rawal, R., Vaarhorst,
622 A., Kangas, A.J., Lyytikainen, L.P., Pirinen, M. *et al.* (2016) Genome-wide study for circulating
623 metabolites identifies 62 loci and reveals novel systemic effects of LPA. *Nature*
624 *communications*, **7**, 11122.
625 42 Draisma, H.H.M., Pool, R., Kobl, M., Jansen, R., Petersen, A.K., Vaarhorst, A.A.M., Yet,
626 I., Haller, T., Demirkan, A., Esko, T. *et al.* (2015) Genome-wide association study identifies
627 novel genetic variants contributing to variation in blood metabolite levels. *Nature*
628 *communications*, **6**, 7208.
629 43 Xie, W., Wood, A.R., Lyssenko, V., Weedon, M.N., Knowles, J.W., Alkayyali, S.,
630 Assimes, T.L., Quertermous, T., Abbasi, F., Paananen, J. *et al.* (2013) Genetic variants
631 associated with glycine metabolism and their role in insulin sensitivity and type 2 diabetes.
632 *Diabetes*, **62**, 2141-2150.

633 44 Raffler, J., Friedrich, N., Arnold, M., Kacprowski, T., Rueedi, R., Altmaier, E.,
634 Bergmann, S., Budde, K., Gieger, C., Homuth, G. *et al.* (2015) Genome-Wide Association Study
635 with Targeted and Non-targeted NMR Metabolomics Identifies 15 Novel Loci of Urinary Human
636 Metabolic Individuality. *PLoS genetics*, **11**, e1005487.
637 45 Demirkan, A., Henneman, P., Verhoeven, A., Dharuri, H., Amin, N., van Klinken, J.B.,
638 Karssen, L.C., de Vries, B., Meissner, A., Goraler, S. *et al.* (2015) Insight in genome-wide
639 association of metabolite quantitative traits by exome sequence analyses. *PLoS genetics*, **11**,
640 e1004835.
641 46 Suhre, K., Shin, S.Y., Petersen, A.K., Mohney, R.P., Meredith, D., Wagele, B., Altmaier,
642 E., CardioGram, Deloukas, P., Erdmann, J. *et al.* (2011) Human metabolic individuality in
643 biomedical and pharmaceutical research. *Nature*, **477**, 54-60.
644 47 Yu, B., Zheng, Y., Alexander, D., Morrison, A.C., Coresh, J. and Boerwinkle, E. (2014)
645 Genetic determinants influencing human serum metabolome among African Americans. *PLoS*
646 *genetics*, **10**, e1004212.
647 48 Takashina, C., Tsujino, I., Watanabe, T., Sakaue, S., Ikeda, D., Yamada, A., Sato, T.,
648 Ohira, H., Otsuka, Y., Oyama-Manabe, N. *et al.* (2016) Associations among the plasma amino
649 acid profile, obesity, and glucose metabolism in Japanese adults with normal glucose tolerance.
650 *Nutrition & metabolism*, **13**, 5.
651 49 Kraus, W.E., Pieper, C.F., Huffman, K.M., Thompson, D.K., Kraus, V.B., Morey, M.C.,
652 Cohen, H.J., Ravussin, E., Redman, L.M., Bain, J.R. *et al.* (2016) Association of Plasma Small-
653 Molecule Intermediate Metabolites With Age and Body Mass Index Across Six Diverse Study
654 Populations. *The journals of gerontology. Series A, Biological sciences and medical sciences*, **71**,
655 1507-1513.

656 50 El Hafidi, M., Perez, I. and Banos, G. (2006) Is glycine effective against elevated blood
657 pressure? *Current opinion in clinical nutrition and metabolic care*, **9**, 26-31.
658 51 Ding, Y., Svingen, G.F., Pedersen, E.R., Gregory, J.F., Ueland, P.M., Tell, G.S. and
659 Nygard, O.K. (2015) Plasma Glycine and Risk of Acute Myocardial Infarction in Patients With
660 Suspected Stable Angina Pectoris. *Journal of the American Heart Association*, **5**.
661 52 Geidenstam, N., Magnusson, M., Danielsson, A.P.H., Gerszten, R.E., Wang, T.J.,
662 Reinius, L.E., Mulder, H., Melander, O. and Ridderstrale, M. (2017) Amino Acid Signatures to
663 Evaluate the Beneficial Effects of Weight Loss. *International journal of endocrinology*, **2017**,
664 6490473.
665 53 Cheng, S., Rhee, E.P., Larson, M.G., Lewis, G.D., McCabe, E.L., Shen, D., Palma, M.J.,
666 Roberts, L.D., Dejam, A., Souza, A.L. *et al.* (2012) Metabolite profiling identifies pathways
667 associated with metabolic risk in humans. *Circulation*, **125**, 2222-2231.
668 54 Felig, P., Marliss, E. and Cahill, G.F., Jr. (1969) Plasma amino acid levels and insulin
669 secretion in obesity. *The New England journal of medicine*, **281**, 811-816.
670 55 Thalacker-Mercer, A.E., Ingram, K.H., Guo, F., Ilkayeva, O., Newgard, C.B. and Garvey,
671 W.T. (2014) BMI, RQ, diabetes, and sex affect the relationships between amino acids and clamp
672 measures of insulin action in humans. *Diabetes*, **63**, 791-800.
673 56 Richiardi, L., Bellocco, R. and Zugna, D. (2013) Mediation analysis in epidemiology:
674 methods, interpretation and bias. *International journal of epidemiology*, **42**, 1511-1519.
675 57 Pardridge, W.M. (2005) The blood-brain barrier: bottleneck in brain drug development.
676 *NeuroRx : the journal of the American Society for Experimental NeuroTherapeutics*, **2**, 3-14.
677 58 Bjorkhem, I. and Meaney, S. (2004) Brain cholesterol: long secret life behind a barrier.
678 *Arteriosclerosis, thrombosis, and vascular biology*, **24**, 806-815.

679 59 McCall, A.L., Millington, W.R. and Wurtman, R.J. (1982) Blood-brain barrier transport
680 of caffeine: dose-related restriction of adenine transport. *Life sciences*, **31**, 2709-2715.
681 60 Matone, A., Scott-Boyer, M.P., Carayol, J., Fazelzadeh, P., Lefebvre, G., Valsesia, A.,
682 Charon, C., Vervoort, J., Astrup, A., Saris, W.H. *et al.* (2016) Network Analysis of Metabolite
683 GWAS Hits: Implication of CPS1 and the Urea Cycle in Weight Maintenance. *PloS one*, **11**,
684 e0150495.
685 61 Haberle, J., Shchelochkov, O.A., Wang, J., Katsonis, P., Hall, L., Reiss, S., Eeds, A.,
686 Willis, A., Yadav, M., Summar, S. *et al.* (2011) Molecular defects in human carbamoy phosphate
687 synthetase I: mutational spectrum, diagnostic and protein structure considerations. *Human*
688 *mutation*, **32**, 579-589.
689 62 Hansmannel, F., Sillaire, A., Kamboh, M.I., Lendon, C., Pasquier, F., Hannequin, D.,
690 Laumet, G., Mounier, A., Ayral, A.M., DeKosky, S.T. *et al.* (2010) Is the urea cycle involved in
691 Alzheimer's disease? *Journal of Alzheimer's disease : JAD*, **21**, 1013-1021.
692 63 Griffin, J.W. and Bradshaw, P.C. (2017) Amino Acid Catabolism in Alzheimer's Disease
693 Brain: Friend or Foe? *Oxidative medicine and cellular longevity*, **2017**, 5472792.
694 64 Pearson, D.L., Dawling, S., Walsh, W.F., Haines, J.L., Christman, B.W., Bazyk, A.,
695 Scott, N. and Summar, M.L. (2001) Neonatal pulmonary hypertension--urea-cycle intermediates,
696 nitric oxide production, and carbamoyl-phosphate synthetase function. *The New England journal*
697 *of medicine*, **344**, 1832-1838.
698 65 Summar, M.L., Gainer, J.V., Pretorius, M., Malave, H., Harris, S., Hall, L.D., Weisberg,
699 A., Vaughan, D.E., Christman, B.W. and Brown, N.J. (2004) Relationship between carbamoyl-
700 phosphate synthetase genotype and systemic vascular function. *Hypertension*, **43**, 186-191.

66 de Vries, P.S., Chasman, D.I., Sabater-Lleal, M., Chen, M.H., Huffman, J.E., Steri, M., Tang, W., Teumer, A., Marioni, R.E., Grossmann, V. *et al.* (2016) A meta-analysis of 120 246 individuals identifies 18 new loci for fibrinogen concentration. *Human molecular genetics*, **25**, 358-370.

67 Sabater-Lleal, M., Huang, J., Chasman, D., Naitza, S., Dehghan, A., Johnson, A.D., Teumer, A., Reiner, A.P., Folkersen, L., Basu, S. *et al.* (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**, 1310-1324.

68 Danik, J.S., Pare, G., Chasman, D.I., Zee, R.Y., Kwiatkowski, D.J., Parker, A., Miletich, J.P. and 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. *Circulation. Cardiovascular genetics*, **2**, 134-141.

69 Astle, W.J., Elding, H., Jiang, T., Allen, D., Ruklisa, D., Mann, A.L., Mead, D., Bouman, H., Riveros-Mckay, F., Kostadima, M.A. *et al.* (2016) The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell*, **167**, 1415-1429 e1419.

70 Pare, G., Chasman, D.I., Parker, A.N., Zee, R.R., Malarstig, A., Seedorf, U., Collins, R., Watkins, H., Hamsten, A., Miletich, J.P. *et al.* (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. *Circulation. Cardiovascular genetics*, **2**, 142-150.

71 van Meurs, J.B., Pare, G., Schwartz, S.M., Hazra, A., Tanaka, T., Vermeulen, S.H., Cotlarciuc, I., Yuan, X., Malarstig, A., Bandinelli, S. *et al.* (2013) Common genetic loci

724 influencing plasma homocysteine concentrations and their effect on risk of coronary artery
725 disease. *The American journal of clinical nutrition*, **98**, 668-676.
726 72 Lange, L.A., Croteau-Chonka, D.C., Marvelle, A.F., Qin, L., Gaulton, K.J., Kuzawa,
727 C.W., McDade, T.W., Wang, Y., Li, Y., Levy, S. *et al.* (2010) Genome-wide association study of
728 homocysteine levels in Filipinos provides evidence for CPS1 in women and a stronger MTHFR
729 effect in young adults. *Human molecular genetics*, **19**, 2050-2058.
730 73 Williams, S.R., Yang, Q., Chen, F., Liu, X., Keene, K.L., Jacques, P., Chen, W.M.,
731 Weinstein, G., Hsu, F.C., Beiser, A. *et al.* (2014) Genome-wide meta-analysis of homocysteine
732 and methionine metabolism identifies five one carbon metabolism loci and a novel association of
733 ALDH1L1 with ischemic stroke. *PLoS genetics*, **10**, e1004214.
734 74 Willer, C.J., Schmidt, E.M., Sengupta, S., Peloso, G.M., Gustafsson, S., Kanoni, S.,
735 Ganna, A., Chen, J., Buchkovich, M.L., Mora, S. *et al.* (2013) Discovery and refinement of loci
736 associated with lipid levels. *Nature genetics*, **45**, 1274-1283.
737 75 Pattaro, C., Teumer, A., Gorski, M., Chu, A.Y., Li, M., Mijatovic, V., Garnaas, M., Tin,
738 A., Sorice, R., Li, Y. *et al.* (2016) Genetic associations at 53 loci highlight cell types and
739 biological pathways relevant for kidney function. *Nature communications*, **7**, 10023.
740 76 Gorski, M., van der Most, P.J., Teumer, A., Chu, A.Y., Li, M., Mijatovic, V., Nolte, I.M.,
741 Cocca, M., Taliun, D., Gomez, F. *et al.* (2017) 1000 Genomes-based meta-analysis identifies 10
742 novel loci for kidney function. *Scientific reports*, **7**, 45040.
743 77 Mahajan, A., Rodan, A.R., Le, T.H., Gaulton, K.J., Haessler, J., Stilp, A.M., Kamatani,
744 Y., Zhu, G., Sofer, T., Puri, S. *et al.* (2016) Trans-ethnic Fine Mapping Highlights Kidney-
745 Function Genes Linked to Salt Sensitivity. *American journal of human genetics*, **99**, 636-646.

746 78 Kottgen, A., Pattaro, C., Boger, C.A., Fuchsberger, C., Olden, M., Glazer, N.L., Parsa,
747 A., Gao, X., Yang, Q., Smith, A.V. *et al.* (2010) New loci associated with kidney function and
748 chronic kidney disease. *Nature genetics*, **42**, 376-384.
749 79 Jun, G., Ibrahim-Verbaas, C.A., Vronskaya, M., Lambert, J.C., Chung, J., Naj, A.C.,
750 Kunkle, B.W., Wang, L.S., Bis, J.C., Bellenguez, C. *et al.* (2016) A novel Alzheimer disease
751 locus located near the gene encoding tau protein. *Molecular psychiatry*, **21**, 108-117.
752 80 Locke, A.E., Kahali, B., Berndt, S.I., Justice, A.E., Pers, T.H., Day, F.R., Powell, C.,
753 Vedantam, S., Buchkovich, M.L., Yang, J. *et al.* (2015) Genetic studies of body mass index yield
754 new insights for obesity biology. *Nature*, **518**, 197-206.
755 81 Melen, E., Granell, R., Kogevinas, M., Strachan, D., Gonzalez, J.R., Wjst, M., Jarvis, D.,
756 Ege, M., Braun-Fahrlander, C., Genuneit, J. *et al.* (2013) Genome-wide association study of
757 body mass index in 23 000 individuals with and without asthma. *Clinical and experimental*
758 *allergy : journal of the British Society for Allergy and Clinical Immunology*, **43**, 463-474.
759 82 Hartiala, J.A., Tang, W.H., Wang, Z., Crow, A.L., Stewart, A.F., Roberts, R.,
760 McPherson, R., Erdmann, J., Willenborg, C., Hazen, S.L. *et al.* (2016) Genome-wide association
761 study and targeted metabolomics identifies sex-specific association of CPS1 with coronary artery
762 disease. *Nature communications*, **7**, 10558.
763 83 TNT, C. (2017), in press.
764 84 Duncan, M.W. (2007) Omics and its 15 minutes. *Exp Biol Med (Maywood)*, **232**, 471-
765 472.
766 85 Stieb, D.M., Boot, C.R. and Turner, M.C. (2017) Promise and pitfalls in the application
767 of big data to occupational and environmental health. *BMC Public Health*, **17**, 372.

86 Fallin, M.D., Duggal, P. and Beaty, T.H. (2016) Genetic Epidemiology and Public Health: The Evolution From Theory to Technology. *American journal of epidemiology*, **183**, 387-393.

772 Figure Legends

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

777 Figure 2. Inter-omic network. This network has 1,224 edges and 635 nodes, which included 171
778 metabolite-gene edges, 833 metabolite-AD risk factor edges. Of these, 73 were CSF metabolite-
779 AD risk factor edges (CSF T-tau and P-tau, exclusively) and 4 were CSF metabolite-gene edges.
780 Red edges indicate negative correlations and blue edges indicate positive correlations.

781 Figure 3. *CPS1*, glycine, and cardiovascular and diabetes sub-network. Relationships within this
782 pathway are highly cited; however, the pathway as a whole is not understood as well. Red edges
783 indicate negative correlations and blue edges indicate positive correlations.

784 Figure 4. CSF biomarker community. This network has 73 edges among 38 CSF metabolites and
785 CSF biomarkers T-tau and P-tau. Red edges indicate negative correlations and blue edges
786 indicate positive correlations.

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

795 Tables

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

796 AD: Alzheimer's disease, $A\beta_{42}$: β-Amyloid$_{42}$, T-tau: Total-tau, P-tau: Phosphorylated-tau, $A\beta_{40}$:

797 β-Amyloid$_{40}$, BMI: Body-mass index, WHR: Waist-hip ratio, METs: Metabolic equivalents, IL-

798 6: Interleukin 6, SBP: Systolic blood pressure, DBP: Diastolic blood pressure, HOMA-IR:

799 Homeostatic model assessment of insulin resistance

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

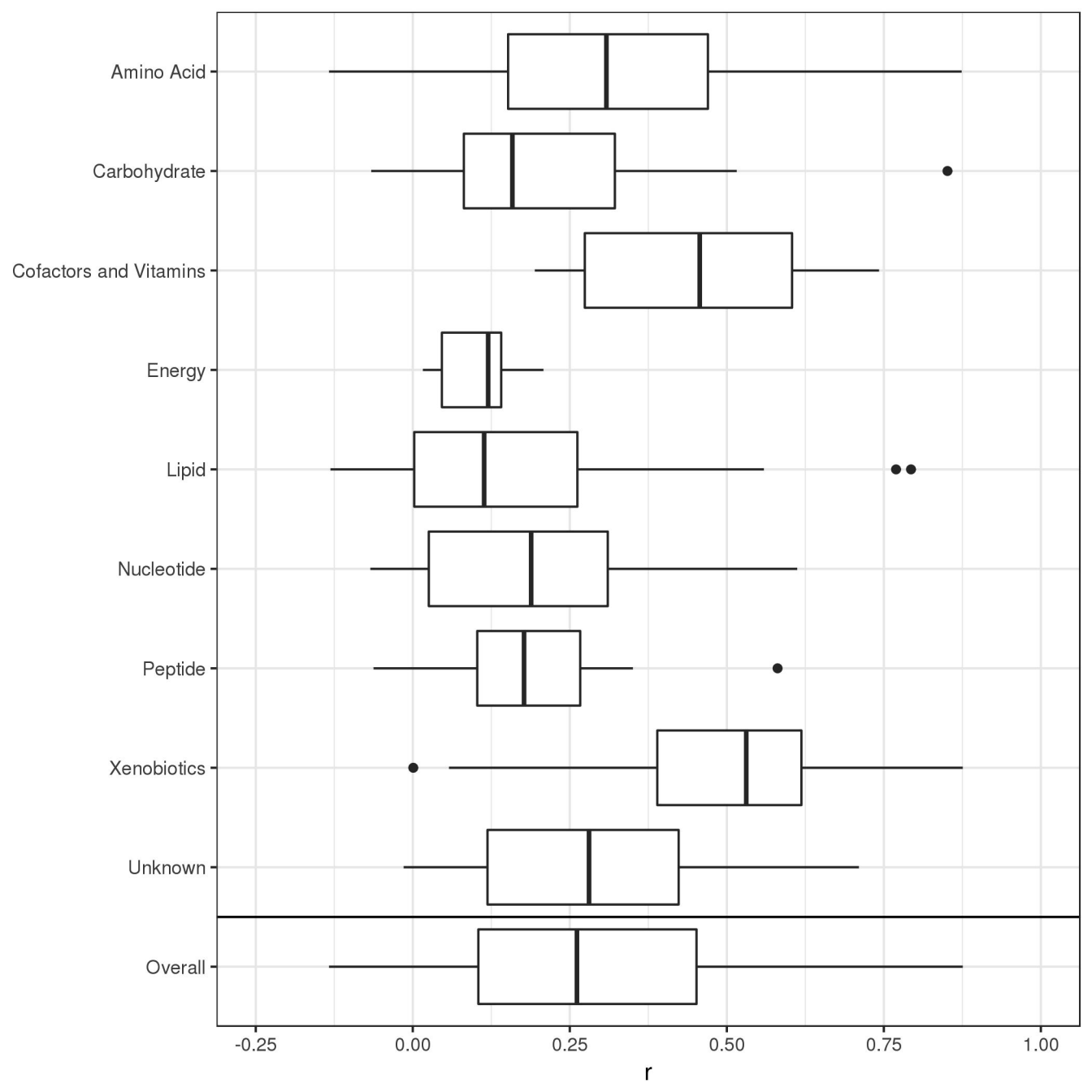
801

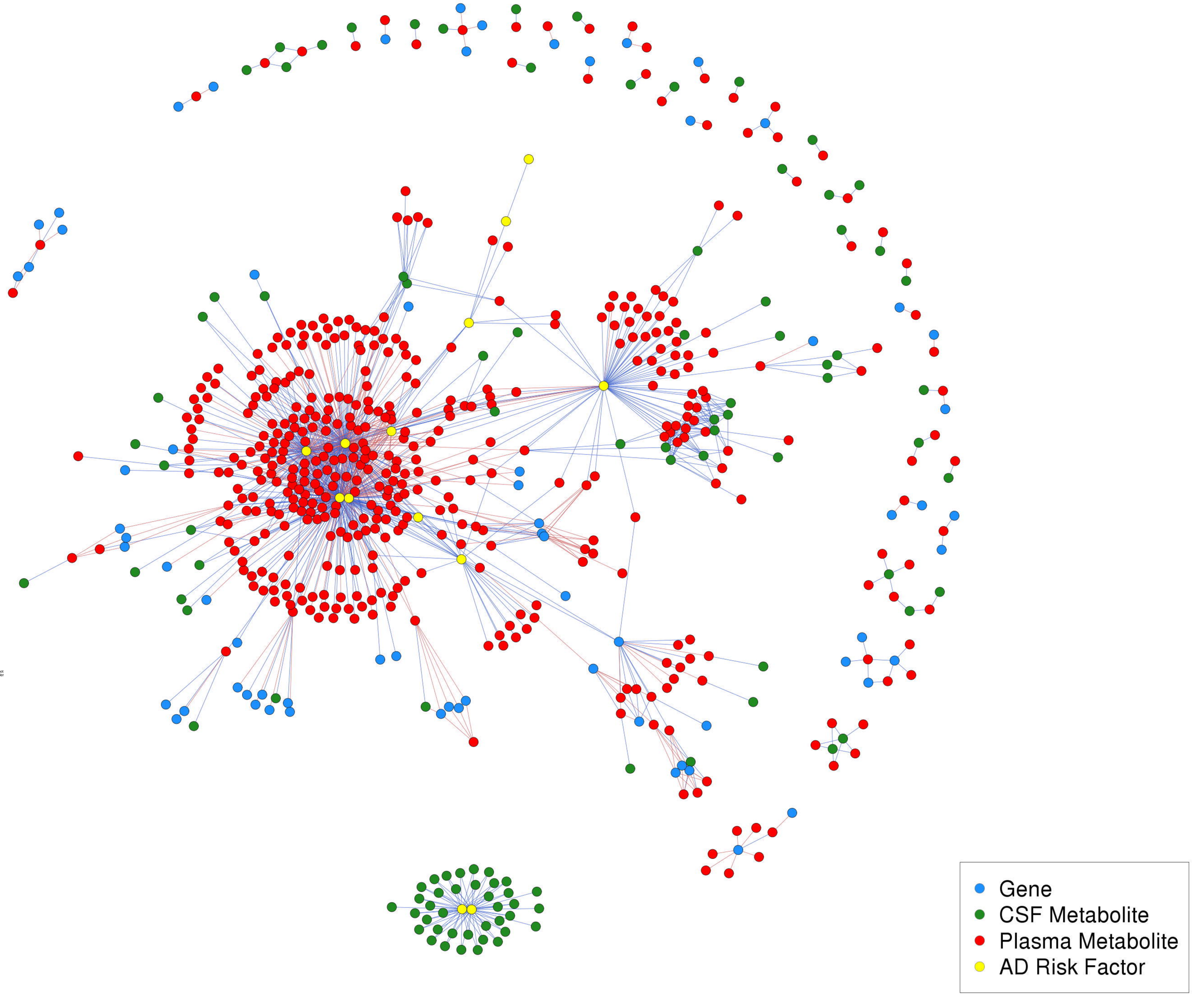
802

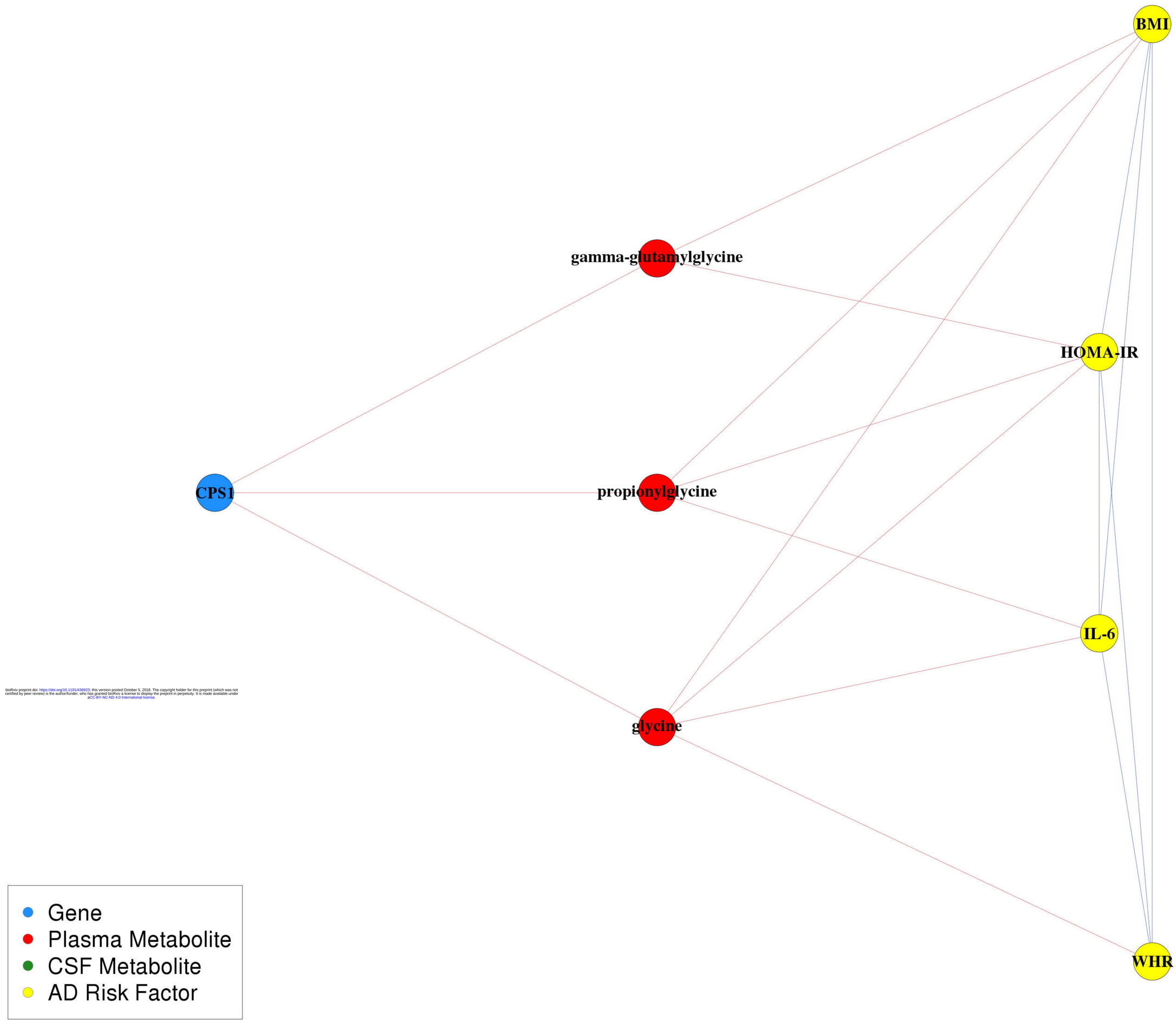
Table 2. WRAP Participant Characteristics at Baseline Sample. Mean (SD) or N (%).

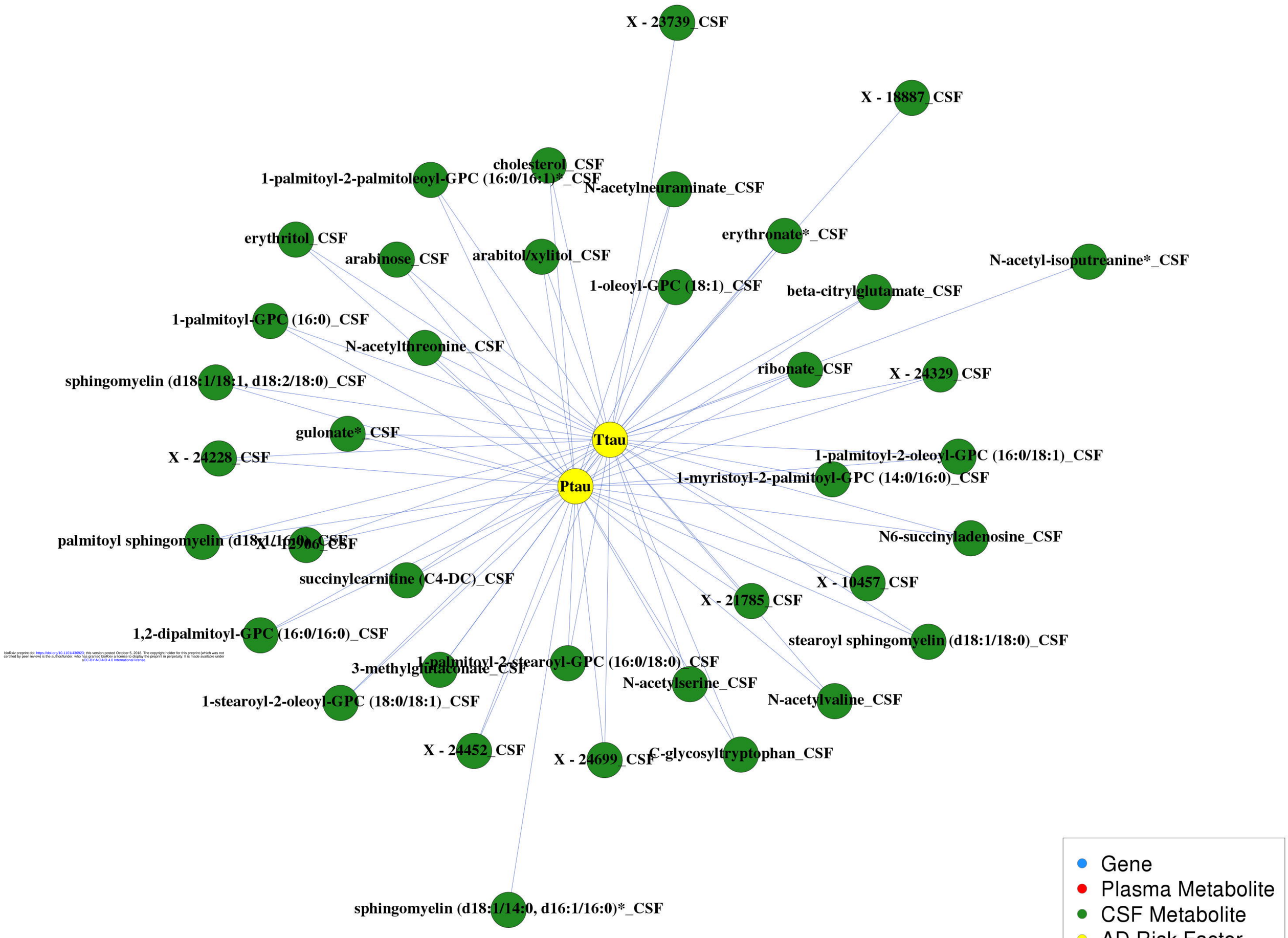
Characteristic	Overall (N=1,111, obs=2,191)	CSF Metabolomics (N=155, obs=346)
Age (years)	61.0 (6.7)	61.2 (6.6)
Female	766 (68.9)	103 (66.5)
Years of education	16.4 (2.8)	16.7 (2.9)
Parental history of AD	803 (72.3)	112 (72.3)
Use of cholesterol-lowering medication	354 (31.9)	45 (29.0)
Number of Visits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

803 obs: observations, CSF: cerebral spinal fluid, AD: Alzheimer's disease

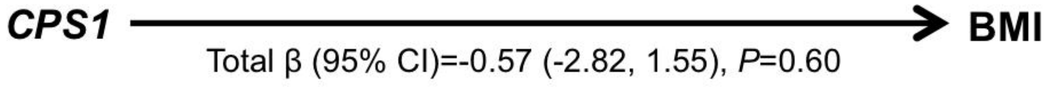

Amino Acid
Carbohydrate
Cofactors and Vitamins
Energy
Lipid
Nucleotide
Peptide
Xenobiotics
Unknown
Overall
-0.25
0.00
0.25
0.50
0.75
1.00
r



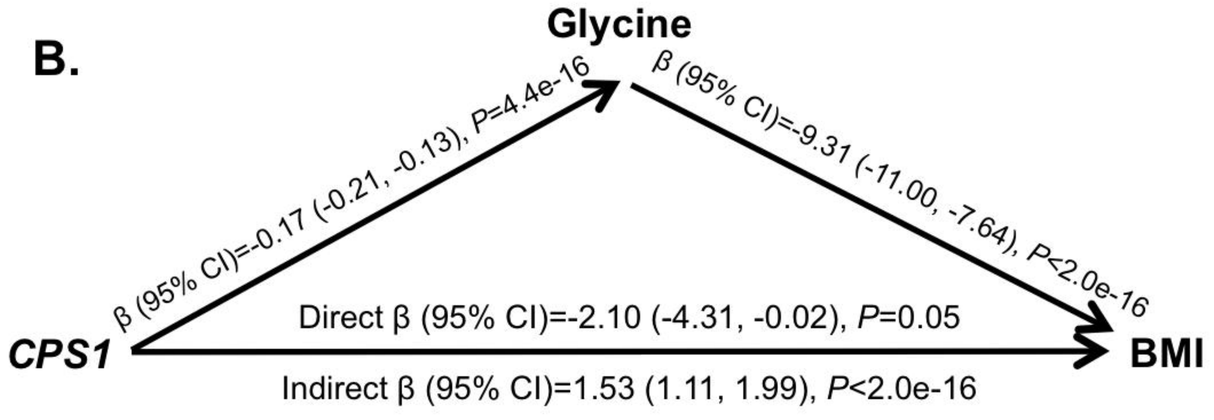




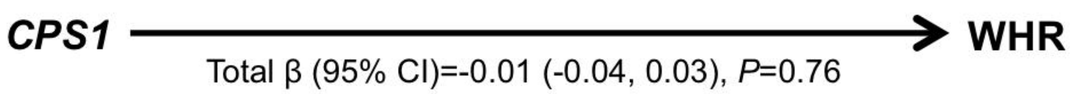
A.



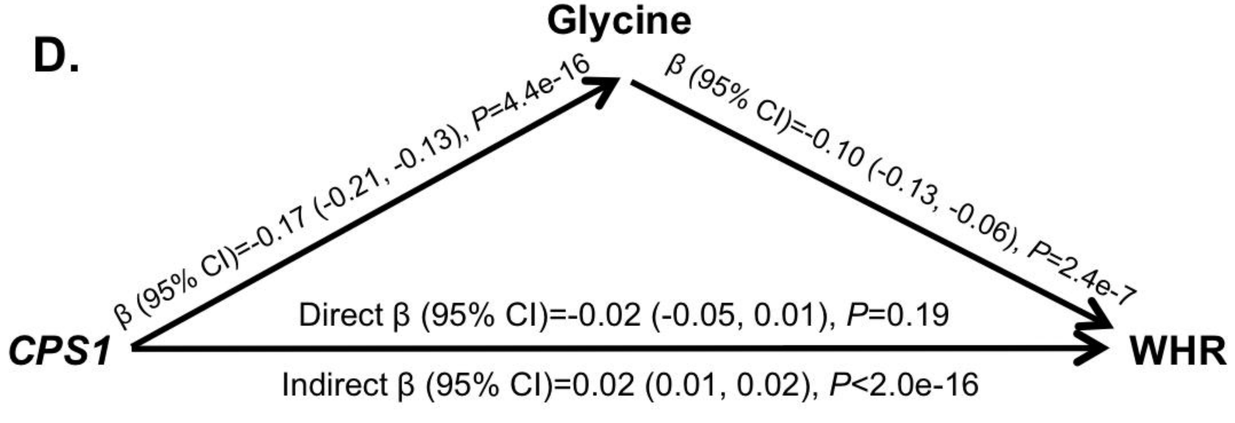
B.



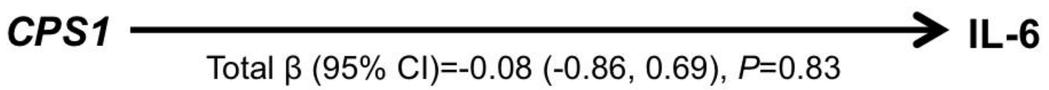
C.



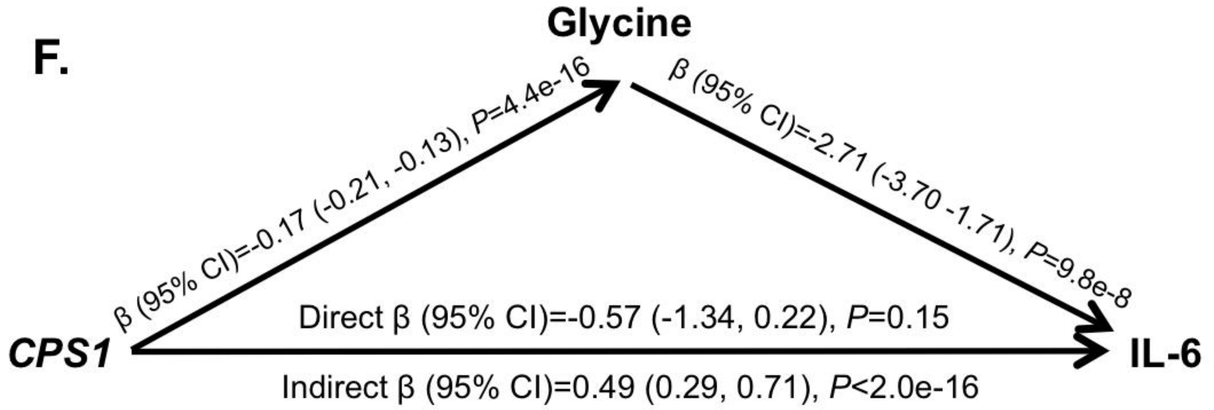
D.



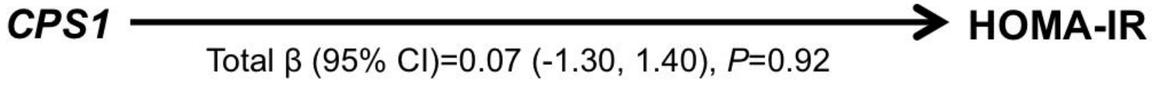
E.



F.



G.



H.