medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314800; this version posted October 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Genome-wide Association Studies of Missing Metabolite Measures: Results From Two Population-based Studies

Authors: Tariq O. Faquih[1,2]*, Mohammed Aslam Imtiaz[3]*, Valentina Talevi[3], Elvire N. Landstra[3], Astrid van Hylckama Vlieg[1], Ruifang Li-Gao[1], Frits R. Rosendaal[1], Raymond Noordam[4], Diana van Heemst[4], Dennis O. Mook-Kanamori[1,5], Monique M. B. Breteler[3,6]**, N. Ahmad Aziz[3,7]**, Ko Willems van Dijk[8,9,10]**

Affiliations:

1. Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, The Netherlands
2. Division of Sleep and Circadian Disorders, Brigham and Women's Hospital, Boston, United States
3. Population Health Sciences, German Centre for Neurodegenerative Diseases (DZNE), Bonn, Germany
4. Department of Internal Medicine, Section of Gerontology and Geriatrics, Leiden University Medical Center, Leiden, The Netherlands
5. Department of Public Health and Primary Care, Leiden University Medical Center, Leiden, The Netherlands
6. Institute for Medical Biometry, Informatics and Epidemiology (IMBIE), Faculty of Medicine, University of Bonn, Bonn, Germany
7. Department of Neurology, Faculty of Medicine, University of Bonn, Bonn, Germany
8. Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands
9. Department of Internal Medicine, Division of Endocrinology, Leiden University Medical Center, Leiden, The Netherlands
10. Einthoven Laboratory for Experimental Vascular Medicine, Leiden University Medical Center, Leiden, The Netherlands

*These authors contributed equally as first authors.

**These authors contributed equally.

Abstract word count: 149

Main manuscript word count: 4006

Number of tables: 4

Number of figures: 4

Number of Extended Figures: 3

Number of Supplementary Tables: 12

Keywords: genome-wide association study, missing measurements, inborn errors of metabolism, metabolomics, poor metabolizers

Corresponding author: Ko Willems van Dijk; E-mail: k.willems_van_dijk@lumc.nl

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314800; this version posted October 7, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Abstract

Metabolomic studies are increasingly used for both etiological and predictive research, but frequently report missing values. We hypothesized that interindividual genetic variation may account for part of this missingness. Therefore, we performed a GWAS of missingness in measured metabolite levels using an untargeted mass spectrometry-based platform in the Netherlands Epidemiology of Obesity Study (N=594) and the Rhineland Study (N=4,165). We considered metabolites missing in 10%-90% of individuals in both cohorts (N=224). GWAS meta-analyses of these metabolites' probability of missingness revealed 55 metabolome-wide significant associations, including 42 novel ones ($p<1.58\times10^{-10}$), involving 28 metabolites and 41 lead SNPs. Despite considerable pleiotropy, the majority of identified SNP-'missing metabolite' associations were biologically plausible, relating to beta-oxidation, bile acids, steroids, and xenobiotics metabolism. These findings suggest that missing values in metabolomics are nonrandom and partly reflect genetic variation, accounting for which is important for both clinical and epidemiological studies, especially nutritional and pharmacogenetics studies.

medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314800; this version posted October 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Introduction

Metabolites are small molecules that are produced or consumed during anabolic or catabolic reactions and constitute the basic building blocks of all biological processes. Circulating metabolite levels are thought to reflect the integrated metabolic response to changes in genetic and non-genetic (including dietary and other environmental) factors.[1] This hypothesis has made metabolomics an attractive field of study for elucidating the biological mechanisms underlying complex multifactorial diseases.[1,2] Recent advances in metabolomics have enabled high-throughput analysis of thousands of metabolites from a single biological sample, and have been applied to study a wide range of cardiovascular,[3,4] metabolic,[5,6] and neurodegenerative outcomes,[7,8] as well as other traits.[9-11]

The field of metabolomics remains relatively new and still faces several challenges. One important challenge is the biological meaning of missing measurements of metabolites, particularly with untargeted approaches.[12,13] Conceptually, missing data could be due to either random or systematic (i.e., technical) measurement errors, or reflect the actual absence of specific metabolites. In addition, when the metabolite concentration in the sample is below the limit of detection of the measurement method, it will be reported as a missing value.[12,13] Indeed, in most studies, missing data are assumed to reflect values below the limit of detection, and consequently are either removed from the analysis or imputed.[12,13] However, *a priori*, it cannot be excluded that missing values of metabolites are caused by genetic variants. In this case, the metabolites with missing values could be truly absent from the sample due to functional alterations of specific biological pathways driven by genetic variation.[9,14,15] Therefore, imputation or removal of those metabolites from the analysis could bias biological interpretation.

Long before large-scale metabolomics data became available, rare genetic mutations affecting metabolism were identified and investigated.[16] Disorders caused by genetic mutations that disrupt metabolism are referred to as inborn errors of metabolism (IEM). Usually, the causal genetic mutations are located in protein coding genes and affect the structure of the encoded proteins to such an extent that their biological function is disrupted.[17] For example, IEM disorders can disrupt carbohydrate metabolism, protein metabolism, fatty acid oxidation, and glycogen storage.[17,18] Collectively, IEM disorders have an overall incidence of 1 in 2500 births.[18] IEM illustrate that certain genetic variants have the potential to prevent the synthesis or breakdown of specific metabolites by disrupting metabolic pathways.[19] We set out to test the hypothesis that at least some of the common missing values in metabolomics data, either due to levels below the limit of detection or otherwise, is caused by common genetic variation. We also hypothesized that the nature and context of the potential associations could provide insights into the potential causes of the missingness (i.e., technical, below limit of detection, or truly absent). To address these hypotheses, we performed genome-wide association studies (GWAS) to discover SNPs associated with the probability of absence (i.e., 'missingness' due to concentrations below the limit of detection or truly absent) of metabolite measures.

Results

Discovery Genome-wide Association Studies of Missing Metabolites

The GWAS of missing metabolite measures was performed separately in 594 individuals from the Netherlands Epidemiology of Obesity (NEO) study (mean (standard deviation (SD)) age: 55.8 (5.9), range: 45-66 years, 53% women), and 4,165 individuals from the Rhineland Study (mean (SD) age: 55.5 (14), range: 30-96 years, 56%

medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314800; this version posted October 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

women). Individual study characteristics and general genotype assay information are summarized in **Table 1**. GWAS results in NEO identified 712 metabolome-wide significant ($p<1.58\times10^{-10}$) associations between 537 SNPs and 6 out of the 341 included metabolites. In the Rhineland Study, we identified 4,370 metabolome-wide significant ($p<1.59\times10^{-10}$) associations between 2,615 SNPs and 32 out of the 425 included metabolites. Restricted to the metabolites that were available in both studies (N=224), the study-specific GWAS identified 523 and 2,613 metabolome-wide significant SNPs for 5 and 26 metabolites in the NEO and the Rhineland study, respectively. The overall workflow of the GWAS analysis and following downstream analyses is illustrated in **Fig. 1**. The summary statistics of GWAS analysis are available in **Supplementary Tables 1** and **2** for NEO and the Rhineland Study, respectively.

Genome-wide Meta-analysis of Missing Metabolites

A meta-analysis of the overlapping 224 metabolites in the two studies identified 5,455 significant associations ($p<1.59\times10^{-10}$), including 3,260 SNPs across 33 different metabolites (**Fig. 2** and **Supplementary Table 3)**. The direction of the associations was similar across both cohorts (Pearson correlation R2=0.92) using independent SNP-metabolite ($r^2<0.6$) associations (**Extended Data Fig. 1** and **Supplementary Table 4**). The majority of these metabolites belonged to the steroid metabolism pathway (N=7), followed by amino-acid metabolism (N=4), fatty acid metabolism (N=4), bile acid metabolism (N=5), and unannotated metabolites (N=8). Other hits belonged to food and plant-derived xenobiotics (i.e., alliin, solanidine, ferulic acid 4-sulfate and caffeic acid sulfate) and nucleotide metabolites (xanthosine).

Genetically Influenced Metabotypes

Genetically Influenced Metabotypes (GIMs), defined as variant-metabolite clusters,[7] were identified by merging the summary statistics of all SNP-metabolite associations and selecting the SNPs with the lowest association P-value, resulting in a set of 7,310,783 unique SNPs. Functional Mapping and Annotation (FUMA) identified 3,260 metabolome-wide significant SNPs indexed by 41 lead SNPs (linkage disequilibrium (LD) $r^2 < 0.1$) located across 25 genomic risk loci (**Supplementary Table 4** and **5**). Those 41 lead SNPs had a total of 55 associations with the odds of missingness of 28 metabolites. Those 55 lead SNP-metabolite associations were further cross-referenced with previous metabolomic GWAS results and metabolome-based GWAS databases to assess the novelty of the associations (**Table 2** and **Supplementary Table 6**). We defined SNP-metabolite associations as "novel" if the SNP-metabolite association was not reported previously, and labelled it as "reported" otherwise. Accordingly, we found 42 novel and 13 previously reported associations. Several of the associated metabolites were connected by shared pathways and genes, as shown in **Fig. 3**, and formed two large clusters. The first cluster was enriched for steroids and bile acid metabolites and contained two pleiotropic SNPs (rs4149056 and rs45446698) associated with 5 and 4 metabolites, respectively. The second cluster comprised of acetylated tryptophan and lysine related metabolites, along with the xenobiotic metabolite alliin. An interactive version of this network is available online at https://tofaquih.github.io/GWASMissingMetabolites/ for further exploration.

Identification of eQTLs, pQTLs and mQTLs

Using GTEx (version 7), we identified 23 expression Quantitative Trait Loci (eQTLs) (**Supplementary Table 6**). Interestingly, among the novel associations, the SNP rs2413667, which was associated with the odds of missingness of solanidine, was an eQTL of *CYP2D6*, an important enzyme for xenobiotic metabolism[20], in adipose

medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314800; this version posted October 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

tissue. The PhenoScanner results showed that five lead SNPs were previously reported as protein Quantitative Trait Loci (pQTLs) for seven different proteins (**Table 3** and **Supplementary Table 7**). These included peptidyl-prolyl cis-trans isomerase D, major histocompatibility class I polypeptide-related sequence B, and DNA repair protein RAD51 homolog. Finally, using the metabolome-based GWAS study databases, we identified 27 lead SNPs as metabolome Quantitative Trait Loci (mQTLs) (**Supplementary Table 8**). For example, we identified rs211710, which was associated with the odds of missingness of 3-decenoylcarnitine in our study, as an mQTL for decenoylcarnitine.

Prioritized Candidate Genes for Missing Metabolites

To map the 41 identified lead SNPs associated with the odds of missingness of the metabolites to candidate casual genes, we used the 'Prioritization of candidate causal Genes at Molecular QTLs' (ProGeM) framework. This framework maps SNPs to genes using two complementary methods. The first is based on positional proximity, which mapped the identified SNPs to 121 genes. The second is based on biological relevance, which mapped the SNPs to 100 relevant genes (**Fig. 4**, **Supplementary Table 9** and **10**). Subsequently, we focused the analysis on genes that were mapped through both methods, resulting in 59 candidate causal genes.

Mediation Analysis Between SNP, Gene Expression Levels, And Odds of Missingness

Out of the 162 identified genes, we conducted a functional analysis for 78 genes whose expression levels were available in 2,575 participants of the Rhineland Study. We found that 18 lead SNPs were significantly associated with 20 genes at a nominal p-value threshold—with the relation between rs10201159 and *ALMS1* emerging as the strongest association (β = -1.55, p-value < 0.001) (**Supplementary Table 11)**. The expression levels of 18 genes were significantly associated with the odds of missingness for 11 metabolites—with the *HPS1* and N2-acetyl, N6,N6-dimethyllysine association emerging as the strongest (β = -0.4532, p-value < 0.001). Finally, mediation analysis (**Table 4**) conducted involving 9 genes whose expression levels were significantly associated with both lead SNPs and the corresponding metabolites, indicated that *SMDT1* and *HPS1* partially mediated the relation between the SNPs and the odds of missingness of the corresponding metabolites (p<0.05). *SMDT1* mediated the effect between rs2413667 and solanidine by 3.2% (β (SE) = -0.02 (0.01)), while *HPS1* mediated the effect between rs2147896 and N2-acetyl, N6,N6-dimethyllysine by 2.8% (β (SE) = -0.02 (0.01)).

Missing Metabolites Variants Implicated in Diseases

Our search in the DisGeNET database, using curated data, revealed that six lead SNPs were related to sixteen phenotypes, three diseases and one syndrome (**Supplementary Table 12**). The phenotypes were either related to liver, kidney, or reproductive function. Notably, rs4149056 was associated with 7 phenotypes including levels of thyroxine, sex hormone binding globulin (SHBG), and estradiol. Another notable SNP was rs45446698, which was related to three phenotypes, including birth body weight, body height, and blood protein measurement. Lastly, in the disease category, we found rs4149056 SNP to be associated with squamous cell carcinoma.

Discussion

We conducted a genome-wide meta-analysis to identify genetic variants associated with the odds of missingness of metabolites and identified 55 lead SNP- 'missing metabolite' associations, of which 42 associations were novel (i.e., not found in previous GWAS of metabolite levels). Based on comprehensive *in silico* functional analyses,

medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314800; this version posted October 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

we identified associations between specific groups of metabolites and common metabolic pathways. First, we identified several SNP-metabolite associations involving metabolites that play a biological role in fatty acid beta oxidation pathways—generally containing an acetyl group—and that are also expressed or related to kidney function. Second, we found a group of metabolites and SNPs related to bile acid and steroid metabolism. And third, we identified SNPs associated with two xenobiotics primarily derived from plant sources.

Regarding the first group of SNP-metabolites associated with fatty acid beta oxidation, notable findings were related to the following metabolites: 1) 3-decenoylcarnitine, a medium-chain acylcarnitine that is involved in the production of energy via beta-oxidation by transporting acyl-groups into mitochondria[21]; we found five novel associations between intronic regions of *ACADM* and *ECI2,* as well as exonic loci in *PPID (*also known as *CypD* [22]*)* and 3-decenoylcarnitine's odds of missingness, 2) indoleacetylglutamine, a gut microbiome–derived metabolite[23] involved in tryptophan metabolism and has been reported in various studies regarding gut microbiome and chronic kidney disease.[24,25] Two novel associations near the *ACSM1* and *ACSM2A* genes were associated with its odds of missingness, 3) N-acetylkynurenine, another metabolite involved in tryptophan metabolism. A novel association in the *STAMPB* gene and a novel association (rs10201159) in the intergenic region of *NAT8* were associated with its odds of missingness, 4) N2-acetyl,N6,N6-dimethyllysine, an amino acid and is the precursor of N6,N6,N6-trimethyl-L-lysine (also known as trimethyllysine (TML)), which in turn has been reported as a potential precursor for trimethylamine and trimethylamine N-oxide.[26,27] The rs10201159 near *NAT8* and three novel associations near *PYROXD2* gene were associated with its odds of missingness.

All five novel loci associated with 3-decenoylcarnitine were located in the vicinity of genes that were strongly linked to the regulation of fatty acid beta-oxidation.[28-30] In addition, disruptions of the metabolism of acylcarnitine and the process of beta-oxidation are well known causes for specific IEM. *ACADM* in particular is linked to medium chain acyl-CoA dehydrogenase (MCAD) deficiency,[29,31] supporting our hypothesis of genetic variations affecting the missingness of 3-decenoylcarnitine. The pQTL analysis showed that rs9410 had been associated with PPID protein expression. Since the protein coded by *PPID* (CypD) functions as a transport pore in the mitochondrial membrane,[32] this mutation could affect the transport of 3-decenoylcarnitine in the mitochondria and lead to reduction of its levels below the limit of detection. In addition, our functional validation analyses using gene expression data showed that *ACADM* and *ECI2* expression was associated with their respective genetic variants, as well as with those of 3-decenoylcarnitine. However, the upstream metabolite carnitine had no missingness in our data (none in NEO and only missing in 9 individuals in the Rhineland study). Additionally, we observed that 3-decenoylcarnitine missingness occurred with higher carnitine levels, not when carnitine was low (**Extended Data Fig. 2A** and **Extended Data Fig. 3A**). This is an indication that these genetic variations could be affecting the metabolism or transport of carnitine to 3-decenoylcarnitine leading to an accumulation of carnitine and delayed production of 3-decenoylcarnitine. Similar to 3-decenoylcarnitine, the biological pathways implicated in indoleacetylglutamine metabolism and its respective SNP associations were related to beta-oxidation of fatty acids. Indoleacetylglutamine is also reportedly elevated in urine of patients with Hartnup disease—an IEM.[25,33]

N-acetylkynurenine is a metabolite belonging to the kynurenine pathway, which in turn are crucial for the breakdown of tryptophan. [34] We found the probability of N-acetylkynurenine missingness is linked to the SNP rs10188058 in the *STAMPB* gene. It is worth noting that both N-acetylkynurenine (as well as the kynurenine

medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314800; this version posted October 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

pathway) and *STAMPB* were previously reported to be associated with the IEM microcephaly-capillary malformation syndrome.[34,35] N-acetylkynurenine, along with N2-acetyl,N6,N6-dimethyllysine (lysine metabolism), alliin, and X-12753, are also associated with the SNP rs10201159 in the intergenic region of *NAT8*. In line with the function of *NAT8*, this locus has been reported to be associated with several acetyl forms of amino acids and fatty acids metabolism and with the progression of chronic kidney disease.[36] Overall, these metabolites are associated with beta-oxidation, mitochondrial function, IEMs, and potentially related to kidney function.

Although little is known about N2-acetyl,N6,N6-dimethyllysine in the literature, its precursors—namely TML—has been studied extensively.[26,27,37] TML has also been associated with cardiovascular diseases and reportedly predicted all-cause mortality and cardiovascular disease.[27] The loci associated with the missingness of N2-acetyl,N6,N6-dimethyllysine are near the regions of *PYROXD2* and *NAT8*. These two genes are related as *PYROXD2* has been reported to interact with *NAT8* in several studies. [9,38,39] *PYROXD2* is localized in the mitochondrial inner membrane and has an important role in regulating mitochondrial function.[40] In addition, *PYROXD2* is associated with the IEM disorder trimethylaminuria.[41] *PYROXD2* is normally associated with low levels of trimethylamine in the urine of healthy individuals.[41] Although the commonly reported primary mutations causing this disorder are in the *FMO3* gene, growing evidence suggests that mutations in *PYROXD2* play a role as well. Indeed, genetic variations in *PYROXD2* have been reported to be implicated with the increased levels of trimethylamine in individuals with trimethylaminuria, specifically in sweat, breath, and urine.[41,42] The eQTL analysis further supported that *PYROXD2* expression was associated with the SNPs (rs11189559 and rs2147896) and with the odds of N2-acetyl,N6,N6-dimethyllysine missingness. Our GWAS and eQTL findings could indicate the involvement of our novel SNPs associations in *PYROXD2* in relation to the missingness of N2-acetyl,N6,N6-dimethyllysine. Subsequently, these are possibly associations with poor metabolism of TML and trimethylamine that could additionally relate to the development of trimethylaminuria.

The second group of metabolites we have reported from our analysis were steroids and bile acids with a shared association with two pleiotropic SNPs: rs4149056 and rs45446698. Rs4149056 (*SLCO1B1*) was associated with two bile acids, two estrone metabolites, and an unknown metabolite X-12456—predicted to be analogous to steroid metabolites—which is in line with findings from previous genomic research literature.[10] First, *SLCO1B1* is associated with statin-induced myopathy via the interaction with bile acids and cholesterol.[43] Second, rs4149056 and *SLCO1B1* were reported to be associated with serum estrone levels.[44,45] Similarly, rs45446698 (*CYP3A7*) has been reported in studies relating breast cancer to oestrone and progesterone levels,[46,47] as well as studies related to atorvastatin metabolism.[48] In addition, PheWAS analysis indicated previous associations between rs45446698 and birth weight, a trait with lifetime implications for metabolism.[49] Overall, these two SNPs participate in similar metabolic processes affecting steroid and bile acids. Although it remains unclear how these two SNPs are involved in reducing the metabolite levels to missingness levels, our pQTL suggests that rs45446698 is associated with post-translational modification of DNA repair protein RAD51 homolog 4 (RA51D) in a trans manner.[50,51] These modifications could indicate consequences of the rs45446698 SNP on the RAD51 functionality that could contribute to the missingness of estrone 3-sulfate, tauro-beta-muricholate, "5alpha-androstan-3alpha,17alpha-diol monosulfate", and glyco-beta-muricholate.

medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314800; this version posted October 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

The final group we have identified was comprised of alliin and solanidine, which are both xenobiotic metabolites derived from the consumption of garlic and potatoes, respectively. Alliin is generally known for its health benefits, such as improved glucose tolerance in mice and anti-inflammatory effects in rats and *in vitro* studies.[52,53] Interestingly, the *NAT8* locus was found to be associated with the odds of missingness of alliin. A previous study reported another SNP in *NAT8* to be associated with the acylated form of alliin—N-acetylalliin.[10] These findings indicate that *NAT8* engages in the metabolism and acylation of alliin in some capacity. It is possible that rs10201159 leads to a faster metabolism of alliin into N-acetylalliin. Consequently, the levels of alliin could fall below the detection limit and are therefore reported as missing in metabolomic analyses.

Solanidine is a steroidal alkaloid, slightly toxic metabolite in low quantities derived from potatoes and other plants of the Solanaceae family.[54,55] We identified three novel SNPs associations from three genes strongly associated with solanidine. The rs2413667 SNP in the eQTL region of *CYP2D6* was particularly noteworthy. The coded protein from this gene is responsible for the metabolism of approximately 25% of drugs used in clinical settings and its association with solanidine has been studied in relation to metabolism efficiency.[56,57] Solanidine has also been reported as a potential dietary marker to assess the efficiency of *CYP2D6* functionality.[58] This was further supported by a clinical trial reporting *CYP2D6* inhibition to be associated with up to 4.56 fold increase of solanidine levels, indicating compromised xenobiotic metabolism. Therefore, additional studies used solanidine as a biomarker to identify "poor metabolizers". [56,58,59] Based on our findings regarding the association of rs2413667 and the odds of missingness of solanidine in tandem with previous studies examining solanidine and *CYP2D6* metabolism, rs2413667 may be utilized as a new pharmacogenomic marker to identify poor metabolizers (or conversely rapid metabolizers) of drugs [60] and could be utilized in identifying and developing personalized nutritional interventions. [61] This may also be true for all the reported SNPs and metabolites in this study. The rs2413667 SNP is also in proximity of the eQTL region of *SMDT1* and based on our mediation analysis, this eQTL has a significant mediation effect of 3.2%. The *SMDT1* encoded protein from this gene partakes in forming a calcium uniporter complex in the mitochondria. In line with the protein function, solanidine and solanine toxicity are characterized by the disruption of calcium transport in mitochondria.[55] Therefore, rs2413667 may affect solanidine metabolism through its influence on *SMDT1* expression.

Four pleiotropy patterns characterized the total 55 associations (**Fig. 3**). First, the odds of missingness of 13 metabolites was associated with multiple loci in different genes, as was the case, e.g., for 3-decenoylcarnitine. Second, five loci illustrated pleiotropy and were associated with the odds of missingness of multiple metabolites, usually belonging to similar metabolic pathways. A noteworthy case of this was rs4149056, which was associated with the odds of missingness of five metabolite measures—three of which were steroid metabolites. Third, we observed three instances where multiple SNPs within the same gene were associated with the odds of missingness of corresponding metabolites. For example, three SNPs in *PYROXD2* were associated with the odds of missingness of N2-acetyl,N6,N6-dimethyllysine. Fourth, the remaining associations were exclusively single SNP-metabolite associations. Taken together, our findings indicate considerable genetic pleiotropy regarding the odds of missing metabolite measures, which, however, converge on common metabolic pathways.

An important consideration for this study is the interpretation of results originating from a non-traditional phenotype—missingness of metabolites. Missingness of a metabolite measurement can be caused by either a

medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314800; this version posted October 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

technical issue, i.e. failure of metabolite identification in the spectral data due to a deconvolution issue, or the metabolite concentration being below the limit of detection, or real missingness, i.e. the metabolite concentration is null. It can be expected that failure of metabolite detection in the spectral data due to a deconvolution issue is to an extent random and unlikely to be caused by genetics. However, it is extremely difficult to distinguish between a metabolite measure being below the limit of detection and a metabolite truly being absent, such as the case of dopamine 4-sulphate. The SNP rs67110785 is found to be associated with missingness of dopamine 4-sulphate measures and is located in close proximity to the tyrosine hydroxylase (TH) encoding gene—a rate limiting enzyme in the synthesis of dopamine.[62] The same SNP is also an eQTL for *TH* in the Genotype-Tissue Expression (GTEx) database. At face value, TH deficiency would lead to missingness of dopamine 4-sulphate; however, TH deficiency also leads to severe neurological problems that were not reported by any of the participants. Dopamine 4-sulphate is produced from dopamine by the enzyme sulfotransferase family 1A member 3 (SULT1A3), which also produces dopamine 3-sulphate.[63] When plotting the levels of dopamine 4-sulphate against dopamine 3-sulphate in the NEO study, it was clear that the individuals with missing values of dopamine 4-sulphate had low levels of dopamine 3-sulphate and could thus not be TH or SULT1A3 deficiency (**Extended Data Fig. 2B** and **Extended Data Fig. 3B**). Missingness of dopamine 4-sulphate is therefore likely due to the measures being below the limit of detection, probably caused by lower levels of dopamine, rather than its absence.

Although the exact mechanism through which these SNPs would induce missingness of metabolites remains to be fully elucidated, findings from previous studies related to IEM and poor metabolism, as well as our eQTL and pQTL analyses supports the hypothesis that genetic factors influence the probability of metabolites absence. Future studies are needed for deeper investigation of the underlying biological pathway of the missingness of the reported metabolites. A limitation of our study is the relatively small samples sizes used in the GWAS. However, by using two studies and a meta-analysis approach, we found strong associations between the genetic variants and missingness, despite the sample size limitation, and were able to replicate our findings. A second potential limitation was using different blood sample collection methods, with serum used in NEO and plasma in the Rhineland study. The differing blood sampling methodology could have influenced the levels and detection of some metabolites and may explain some of the disparity in the total measured metabolites between the two studies. However, we did find a very strong correlation between the loci-metabolites effect estimates from the NEO and Rhineland study. Additionally, high correlations were previously reported for metabolite measurements from serum and plasma samples collected from the same individuals.[64] Therefore, the choice of blood sampling type may have a limited impact on the overall metabolite profiles and our findings. A third limitation was the nature of the untargeted platforms. These platforms can be prone to missing data due to systematic errors.[12,13] We accounted for this limitation by excluding metabolites outside the missingness limits (<10% or >90% missingness) to avoid the inclusion of metabolites that were simply missing due to systematic errors as much as possible. Future targeted metabolomics studies that measure the absolute concentrations of our reported metabolites can aid in replicating our findings. Finally, our study included populations from European ancestry only and, therefore, further studies are required to investigate metabolite missingness in different populations and ethnicities.

In summary, we identified 55 associations between genetic variants and the odds of missingness of numerous metabolites, 42 of which were completely novel associations. These associations involved 24 SNP-metabolite pairs related to fatty acid beta oxidation and kidney function. In addition, two pleiotropic SNPs were notable for

medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314800; this version posted October 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

their associations with metabolites partaking in steroid and bile acid metabolism, as well as metabolism of dietary and xenobiotic metabolites. Our results provide novel insights into the role of genetics in determining the absence of certain metabolites, with potential implications for the identification of both "poor metabolizers" and IEM. Indeed, the novel genetic variants reported here could have potential value in future etiological and prediction studies, especially in the fields of metabolomics, nutritional epidemiology, pharmacogenomics, and IEM disorders.

medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314800; this version posted October 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Online Methods

Study Populations

We included 594 and 4,165 individuals of European ancestry from the Netherlands Epidemiology of Obesity (NEO) study and the Rhineland Study, respectively, who had both genetic and metabolomics data. NEO is a population-based, prospective cohort study, initiated in 2008. All participants recruited in this study gave written informed consent and the Medical Ethical Committee of the Leiden University Medical Center (LUMC) approved the study design. A detailed description of the study design and data collection can be found elsewhere.[65] Briefly, men and women aged between 45 and 65 years with a self-reported body mass index (BMI) of 27 kg/m^2 or higher living in the greater area of Leiden (in the west of the Netherlands) were eligible to participate in NEO. Participants were invited for a baseline visit at the NEO center in the LUMC after an overnight fast. At the baseline visit, fasting blood samples were drawn. The Rhineland Study is an ongoing prospective population-based cohort study in Bonn, Germany. People aged 30 years or above who lived in two geographically defined areas in Bonn were invited to participate with the only exclusion criterium being insufficient command of the German language to provide informed consent. These participants underwent deep phenotyping to obtain whole-blood, genetic, imaging, socio-demographic, and clinical data.

Metabolomics Measurements and Missingness Inclusion Criteria

Metabolites were measured on the Metabolon HD4 platform in the fasting state serum samples (N= 594) from NEO and fasting state plasma samples (N= 4,165) from the Rhineland Study. Details on the metabolomics pipeline have been described elsewhere.[11] In brief, the Metabolon HD4 platform employs an untargeted measurement approach that utilizes ultra-performance liquid chromatography (UPLC) tandem mass spectrometry (MS/MS) combined with a positive ion mode electrospray ionization, RP/UPLC/-MS/MS combined with a negative ion mode electrospray, and HILIC/UPLC-MS/MS combined with a negative ion mode electrospray ionization. In total, 1,365 metabolites were measured in NEO and 1,077 were measured in the Rhineland Study. Of these, 847 and 467 were endogenous in NEO and the Rhineland Study, respectively. Based on the pathway annotations by Metabolon, these endogenous metabolites spanned 10 pathway groups: amino acids, cofactors and vitamins, lipids, energy, nucleotides, peptides, carbohydrate, and partially characterized molecules. In addition, it included measurements of 222 and 321 xenobiotic metabolites as well as 296 and 289 unannotated metabolites from NEO and the Rhineland Study, respectively.

Most of missingness for endogenous metabolites measured by the metabolon platform occurs in less than 10% of the measured population. These cases are commonly due to systematic or random errors in measurements. On the other hand, for xenobiotic metabolites most missing values are found in 90% of the population as these metabolites depend on specific external exposures (i.e., medication use, nicotine exposure etc.). Therefore, we included metabolites with a moderate number of missing values by excluding metabolites that had a missingness percentage that was either below 10% or above 90% within each study. The rationale for this approach was to exclude metabolites with a high probability of having missing values due to systematic and random errors (<10%), are truly missing (>90%), or other unknown causes. Accordingly, we selected 341 and 425 metabolites in NEO and the Rhineland Study, respectively.

medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314800; this version posted October 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Genotyping and imputation

In NEO, DNA was extracted from 6,671 venous blood samples obtained from the antecubital vein. Genotyping was performed in the Centre National de Génotypage (Evry Cedex, France), using Illumina HumanCoreExome-24 BeadChip (Illumina Inc., San Diego, California, United States of America). The detailed quality control process has been described previously.[66] Genotypes were imputed to the Haplotype Reference Consortium (HRC) release 1.1.[67] In the Rhineland Study, 4,165 DNA samples isolated from buffy coats extracted from blood samples were genotyped using the Illumina Omni-2.5 exome array and processed with GenomeStudio (version 2.0.5). Quality control was performed using PLINK (version 1.9). SNPs were excluded based on poor genotyping rate (< 99%) or Hardy-Weinberg Equilibrium ($p < 1\times10^{-6}$). Additionally, participants with poor quality DNA samples were excluded, on account of a poor call rate (<95%) (N=41), abnormal heterozygosity (N=69), cryptic relatedness (N=261), or a sex mismatch (N=28). To account for variation in the population structure, which may otherwise cause systematic differences in allele frequencies.[68] We used EIGENSTRAT (version 16000). EIGENSTRAT uses principal component analysis to detect and correct for population structure, which resulted in the exclusion of an additional 164 participants from non-European descent. Finally, imputation was performed with IMPUTE (version 2),[69] using the 1000 Genomes version 3 phase 5 as the reference panel.[70]

Genome-wide Association Analyses

In NEO, we performed the GWAS of missing metabolites using the SNPTEST v2 software, employing logistic regression analysis under an additive model. In the Rhineland Study, the GWAS was performed using the REGENIE software (v2.2),[71] fitting a firth logistic regression model to the data. REGENIE computation is composed of two steps. Step 1 uses a subset of genetic markers to fit a whole genome regression model that captures the phenotype variance attributable to genetic effects. In step 2, a larger set of imputed SNPs are used in order to test for their association with the different phenotypes conditional upon the prediction from step 1 and using a leave one chromosome out scheme. Genotyped SNPs were pruned using a linkage disequilibrium (LD) r^2-threshold of 0.9 with a window size of 1,000 and a step size of 100 markers.

Overall, we included 21,243,072 and 49,953,404 imputed SNPs in the GWAS analysis in NEO and the Rhineland Study, respectively. These analyses were restricted to SNPs with an imputation quality > 0.3 and minor allele frequency (MAF) > 0.01. The missing metabolites were adjusted for age, sex, fasting status and five genetic principal components in the NEO study, and age, sex, fasting status and the first ten genetic principal components in the Rhineland Study. Genome-wide significance level was set at $p < 5\times10^{-8}$. However, because of the large number of outcome variables (i.e., missing metabolites), we corrected for multiple testing using the method of Li & Ji,[72] which estimates the effective number of independent tests. Accordingly, we estimated the effective number of independent tests to be 315 and 313 in NEO and the Rhineland Study, resulting in a metabolome-wide significance level of $p < 1.58\times10^{-10}$ ($\approx 5\times10^{-8}/315$) and $p < 1.59\times10^{-10}$ ($\approx 5\times10^{-8}/313$), respectively.

medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314800; this version posted October 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Meta-analysis of GWAS

For the GWAS meta-analysis, we selected and used 7,310,783 SNPs which had MAF > 0.01 in the Rhineland Study, as it had the larger sample size between the two studies. We then harmonized the SNPs with the overlapping 6,610,552/7,310,783 SNPs in the NEO study that had a MAF > 0.01. Meta-analysis was performed employing an inverse variance-weighted fixed-effects model using METAL.[73] Identification of allele flips and applying genomic control was performed using METAL as well for each cohort prior to performing the meta-analysis. The metabolome-wide significance level for the meta-analysis was set at $p < 1.59\times10^{-10}$, which was the more stringent cut-off used in the Rhineland Study.

Definition of Genomic Risk Loci

To identify genomic regions associated with missing metabolites, a single dataset was created by identifying the minimum p-value for each SNP across all meta-GWAS summary statistics of missing metabolites (6). This dataset was LD-clumped (r2 < 0.6) using the Functional Mapping and Annotation (FUMA) platform[74] with the 1000 Genomes Phase 3 European reference panel to account for the LD structure. We further represented those clumped SNPs by lead SNPs, which are a subset of the independent significant SNPs that are in approximate LD with each other at $r^2 > 0.1$. Finally, we identified associated genomic risk loci by merging any physically overlapping lead SNPs (LD blocks < 250 kb apart).

Novel Associations

To identify novel GWAS hits, we used curated metabolome GWAS databases such as mGWAS-Explorer[75] and PhenoScanner.[76] We further manually verified whether our lead SNPs were previously identified in metabolomic-based GWAS studies[10] as metabolome quantitative trait locus (mQTLs).

Identifying Candidate Genes

We aimed to identify candidate genes tagged by the lead SNPs that may influence the probability of missingness of certain metabolites. To achieve this, we first used the "prioritization of candidate causal genes at molecular QTLs" (ProGeM) framework.[77] This framework implements a two-step approach (bottom-up and top-down) to identify putative causal genes. In the bottom-up approach, the three closest protein-coding genes within 500 kb of the lead SNP are selected, while the top-down approach uses curated gene function databases (e.g., Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), Mouse Genome Informatics (MGI) and Orphanet) to identify biologically relevant genes that are present within 500 kb of the lead SNPs. We used Variant Effect Predictor (VEP)[78] to search for the closest protein coding genes with the lead SNP and also calculated the impact factor score of the lead SNPs based on its function as either missense, start loss or stop gain. ProGEM also assesses whether the lead SNPs are eQTL using the GTEX v7 database.[79] The genes that were identified through top-down and bottom-up approach, were prioritized as candidate genes.

PheWAS

We performed a phenome-wide association study (PheWAS) using the disgenet2r package to check which lead SNPs had previously been reported to be associated with any other clinical or disease outcomes, as contained in the disease-gene association database (DisGeNET). The package ranks the associations using Variant Disease Association (VDA) scores ranging from 0 to 1, where a higher score represents stronger evidence of a SNP association with a disease outcome. Only lead SNPs with a score > 0.7 were reported.

medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314800; this version posted October 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

RNA sequencing data in the Rhineland Study

Total RNA was isolated from 3,384 whole blood samples, stored, and stabilized in PAXgene Blood RNA tubes (PreAnalytix/Qiagen) using PAXgene Blood miRNA Kit in accordance with the manufacturer's instructions and following the semi-automatic purification protocol (PreAnalytix/Qiagen). RNA integrity and quantity were assessed using the Tapestation 4200 system (Agilent). After using 750 ng of total RNA to generate next generation-sequencing libraries for total RNA sequencing (TruSeq stranded total RNA kit,Illumina), a Ribo-Zero Globin reduction was performed. Libraries were quantified using Qubit HS dsDNA assay (Invitrogen) and clustered at 250 pM concentrations on a NovaSeq6000 instrument using NovaSeq S2 v1 chemistry (Illumina) in XP mode for the first batch of 3,000 samples and NovaSeq S4 v1.5 chemistry for the last batch of 384 samples and sequenced paired-end 2*50 cycles. Sequencing data was demultiplexed and converted into fastq format using bcl2fastq2 v2.20. We performed the quality control check on raw sequencing reads using FastQC v0.11.9 and we filtered low-quality score reads using Trimmomatic v.0.39. Next, we used STAR v2.7.1 aligner to align the sequencing reads to the human reference genome GRCh38.p13 and to generate the gene count matrix through "STAR –quantMode GeneCounts" using the human gene annotation version GRCh38.101. Genes with overall mean expression greater than 15 reads and expressed in at least 5% of the participants were selected for further analysis. Finally, we used the "varianceStabilizingTransformation" function from DESeq2 v1.30.1 to normalize and transform the raw counts.

Gene Expression quantitative trait loci

To functionally validate the GWAS results using gene expression data, we used a three stepped approach analysis conducted on the first 3,384 consecutive participants of the Rhineland Study on which genetics, gene expression and metabolite data were available and quality controlled (N = 2,575). First, we assessed the associations between the lead SNPs and the corresponding genes, selected through bottom-up and top-down approaches. We adjusted gene expression levels for age, sex, the first 10 principal components, red and white blood cell counts, the relative fractions of basophils, eosinophils, lymphocytes, monocytes and neutrophils, and batch effect and we extracted the residuals. Next, linear regression analysis was performed to assess the associations between the lead SNPs (independent variable) and the residuals of candidate genes (dependent variable). Second, we evaluated the relations between the residuals of gene expression data, obtained after adjustment for identical covariates as before (independent variable) and the significant metabolites with missing values, adjusting for metabolomics' batch effect using logistic regression analysis. Third, we employed a mediation analysis using the R package lavaan v.06-11 to investigate which candidate genes mediated the associations between lead SNPs and missing metabolite with 1000 bootstrapping iterations.

Protein quantitative trait analysis

We performed a protein quantitative trait analysis by using the Phenoscanner. [76] Briefly, Phenoscanner holds publicly available protein quantitative trait loci (pQTLs) results from large-scale genome-wide association studies. We verified whether our lead SNPs associated with missing metabolites were previously identified as pQTLs. Accordingly, we filtered results for protein wide association studies with significant SNP-protein associations ($p<5\times10^{-8}$) with our lead SNP or a proxy SNP ($r2>0.9$) with our lead SNPs.

medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314800; this version posted October 7, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Network Representation of Gene-SNP-metabolite Associations

We used the associations identified through our meta-analysis, ProGEM mapping, and PheWAS to construct a comprehensive interactive network. Each SNP, metabolite, gene, metabolite sub-pathway (as annotated in the metabolon dataset), disease or phenotype associations from DisGeNET, and pQTL associations from Phenoscanner were presented as "nodes" with distinct colours. The width of the "edges" connecting the SNP-metabolites was determined according to the -log10 of the p-value of the effect estimate. Additionally, the colour of these edges was chosen to reflect the novelty of the association. Visualization and layout of the networks were created using Cytoscape version 3.10.2 and Gephi v0.10 and then exported as an interactive HTML5 using the sigmaExporter plugin [80] at the following URL https://tofaquih.github.io/GWASMissingMetabolites/.

medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314800; this version posted October 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Author Contributions

T.O. Faquih: Conceptualization, Methodology, Formal Analysis, Writing – Original Draft Preparation, Visualization; M.A. Imtiaz: Conceptualization, Methodology, Formal Analysis, Writing – Original Draft Preparation, Visualization; V. Talevi: Methodology, Formal Analysis, Writing – Reviewing and Editing; E.N. Landstra.: Conceptualization, Methodology, Writing – Reviewing and Editing; A. van Hylckama-Vlieg: Conceptualization, Supervision, Writing – Reviewing and Editing; N. A. Aziz: Conceptualization, Methodology, Writing – Review and Editing, Supervision; R. Li-Gao: Writing – Review and Editing; F.R. Rosendaal.: Study design, Funding acquisition, Conceptualization; R. Noordam and D. van Heemst: Funding acquisition, Writing – Review and Editing; D.O. Mook-Kanamori: Conceptualization, Methodology, Resources, Writing – Reviewing and Editing, Funding Acquisition, Supervision; M.M.B. Breteler: Conceptualization, Methodology, Resources, Writing – Reviewing and Editing, Data Curation, Funding Acquisition, Supervision; K.Willems van Dijk: Conceptualization, Supervision, Writing – Reviewing and Editing. All authors read and approved the final manuscript.

Acknowledgements

We would like to thank all participants and the study personnel of the Rhineland Study. The authors of the NEO study thank all participants, all participating general practitioners for inviting eligible participants, all research nurses for data collection, and the NEO study group: Pat van Beelen, Petra Noordijk, and Ingeborg de Jonge for study coordination, laboratory, and data management.

Funding

The NEO study is supported by the participating Departments, Division, and Board of Directors of the Leiden University Medical Center, and by the Leiden University, Research Profile Area Vascular and Regenerative Medicine. D.O. Mook-Kanamori. is supported by Dutch Science Organization (ZonMW-VENI Grant No. 916.14.023). D. van Heemst. and R. Noordam were supported by a grant of the VELUX Stiftung [grant number 1156]. T.O. Faquih was supported by the King Abdullah Scholarship Program and King Faisal Specialist Hospital & Research Center [No. 1012879283]. The Rhineland Study is funded by the German Center for Neurodegenerative Diseases (DZNE). The work was further partly supported by the German Research Foundation (DFG) under Germany's Excellence Strategy (EXC2151-390873048) and SFB1454—project number 432325352; the Federal Ministry of Education and Research under the Diet-Body-Brain Competence Cluster in Nutrition Research (grant numbers 01EA1410C and FKZ:01EA1809C) and in the framework "PreBeDem—Mit Prävention und Behandlung gegen Demenz" (FKZ: 01KX2230); and the Helmholtz Association under the Initiative and Networking Fund (No. RA-285/19) and the 2023 Innovation Pool. N.A. Aziz is partly supported by a European Research Council Starting Grant (#101041677).

Competing Interests

All authors have no relevant financial or non-financial interests to declare.

medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314800; this version posted October 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Ethical Approval

The NEO study was approved by the medical ethical committee of the Leiden University Medical Centre (LUMC) and all participants gave their written informed consent. The ethics committee of the medical faculty of the University of Bonn approved the undertaking of the Rhineland Study and it was carried out according to the recommendations of the "International Council for Harmonisation Good Clinical Practice" standards.

Consent to Participate

Written informed consent was acquired from all participants per the Declaration of Helsinki in both the NEO and Rhineland Study.

Data Availability

To protect participant privacy and comply with legal regulations, the NEO study data is not publicly accessible. Qualified researchers can request access by contacting the NEO Executive Board at https://www.lumc.nl/org/neo-studie/contact/. Similarly, the Rhineland Study data used in this manuscript is restricted to public access due to data protection laws. Researchers seeking access to these datasets can submit requests to RS-DUAC@dzne.de, providing evidence of their qualifications and adherence to the respective study's data use policies. All authors had full access to the data from their respective studies and are responsible for the accuracy and integrity of the data and analysis.

References:

1. Clish, C.B. Metabolomics: an emerging but powerful tool for precision medicine. *Cold Spring Harb Mol Case Stud* **1**, a000588 (2015).
2. Tzoulaki, I., Ebbels, T.M., Valdes, A., Elliott, P. & Ioannidis, J.P. Design and analysis of metabolomics studies in epidemiologic research: a primer on -omic technologies. *Am J Epidemiol* **180**, 129-39 (2014).
3. Shah, S.H., Kraus, W.E. & Newgard, C.B. Metabolomic profiling for the identification of novel biomarkers and mechanisms related to common cardiovascular diseases: form and function. *Circulation* **126**, 1110-20 (2012).
4. Zheng, Y. *et al.* Metabolomics and incident hypertension among blacks: the atherosclerosis risk in communities study. *Hypertension* **62**, 398-403 (2013).
5. Ahola-Olli, A.V. *et al.* Circulating metabolites and the risk of type 2 diabetes: a prospective study of 11,896 young adults from four Finnish cohorts. *Diabetologia* **62**, 2298-2309 (2019).
6. Menni, C. *et al.* Biomarkers for type 2 diabetes and impaired fasting glucose using a nontargeted metabolomics approach. *Diabetes* **62**, 4270-6 (2013).
7. Faquih, T.O. *et al.* Normal range CAG repeat size variations in the HTT gene are associated with an adverse lipoprotein profile partially mediated by body mass index. *Human Molecular Genetics* **32**, 1741-1752 (2023).
8. Hatano, T., Saiki, S., Okuzumi, A., Mohney, R.P. & Hattori, N. Identification of novel biomarkers for Parkinson's disease by metabolomic technologies. *Journal of Neurology, Neurosurgery & Psychiatry* **87**, 295-301 (2016).
9. Shin, S.Y. *et al.* An atlas of genetic influences on human blood metabolites. *Nat Genet* **46**, 543-550 (2014).
10. Surendran, P. *et al.* Rare and common genetic determinants of metabolic individuality and their effects on human health. *Nat Med* **28**, 2321-2332 (2022).
11. Faquih, T.O. *et al.* Hepatic triglyceride content is intricately associated with numerous metabolites and biochemical pathways. *Liver Int* **43**, 1458-1472 (2023).
12. Faquih, T. *et al.* A Workflow for Missing Values Imputation of Untargeted Metabolomics Data. *Metabolites* **10**(2020).
13. Do, K.T. *et al.* Characterization of missing values in untargeted MS-based metabolomics data and evaluation of missing data handling strategies. *Metabolomics* **14**, 128 (2018).
14. Yousri, N.A. *et al.* Whole-exome sequencing identifies common and rare variant metabolic QTLs in a Middle Eastern population. *Nat Commun* **9**, 333 (2018).

medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314800; this version posted October 7, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

15. Tin, A. *et al.* Target genes, variants, tissues and transcriptional pathways influencing human serum urate levels. *Nat Genet* **51**, 1459-1474 (2019).
16. Hsia, D.Y.-y. *Inborn errors of metabolism / [by] David Yi-yung Hsia*, (Chicago : Year Book Publishers, Chicago, 1959).
17. Dharuri, H. *et al.* Genetics of the human metabolome, what is next? *Biochim Biophys Acta* **1842**, 1923-1931 (2014).
18. Jeanmonod, R., Asuka, E. & Jeanmonod, D. Inborn Errors of Metabolism. in *StatPearls* (StatPearls Publishing
Copyright © 2024, StatPearls Publishing LLC., Treasure Island (FL) ineligible companies. Disclosure: Edinen Asuka declares no relevant financial relationships with ineligible companies. Disclosure: Donald Jeanmonod declares no relevant financial relationships with ineligible companies., 2024).
19. Balasubramaniam, S., Duley, J.A. & Christodoulou, J. Inborn errors of purine metabolism: clinical update and therapies. *J Inherit Metab Dis* **37**, 669-86 (2014).
20. Taylor, C. *et al.* A Review of the Important Role of CYP2D6 in Pharmacogenomics. *Genes* **11**, 1295 (2020).
21. Dambrova, M. *et al.* Acylcarnitines: Nomenclature, Biomarkers, Therapeutic Potential, Drug Targets, and Clinical Trials. *Pharmacol Rev* **74**, 506-551 (2022).
22. Kieffer, L.J. *et al.* Cyclophilin-40, a protein with homology to the P59 component of the steroid receptor complex. Cloning of the cDNA and further characterization. *J Biol Chem* **268**, 12303-10 (1993).
23. Lee, A.M. *et al.* Circulating Metabolomic Associations with Neurocognitive Outcomes in Pediatric CKD. *Clinical Journal of the American Society of Nephrology* **19**(2024).
24. Hu, J.R. *et al.* A metabolomics approach identified toxins associated with uremic symptoms in advanced chronic kidney disease. *Kidney Int* **101**, 369-378 (2022).
25. Ludwig, G.D. & Epstein, I.S. A Genetic Study of Two Families Having the Acute Intermittent Type of Porphyria. *Annals of Internal Medicine* **55**, 81-93 (1961).
26. Koeth, R.A. *et al.* γ-Butyrobetaine is a proatherogenic intermediate in gut microbial metabolism of L-carnitine to TMAO. *Cell Metab* **20**, 799-812 (2014).
27. Bjørnestad, E. *et al.* Trimethyllysine predicts all-cause and cardiovascular mortality in community-dwelling adults and patients with coronary heart disease. *Eur Heart J Open* **1**, oeab007 (2021).
28. Tavecchio, M., Lisanti, S., Bennett, M.J., Languino, L.R. & Altieri, D.C. Deletion of Cyclophilin D Impairs β-Oxidation and Promotes Glucose Metabolism. *Sci Rep* **5**, 15981 (2015).
29. Vishwanath, V.A. Fatty Acid Beta-Oxidation Disorders: A Brief Review. *Ann Neurosci* **23**, 51-5 (2016).
30. Houten, S.M., Violante, S., Ventura, F.V. & Wanders, R.J. The Biochemistry and Physiology of Mitochondrial Fatty Acid β-Oxidation and Its Genetic Disorders. *Annu Rev Physiol* **78**, 23-44 (2016).
31. Merritt, J.L., 2nd & Chang, I.J. Medium-Chain Acyl-Coenzyme A Dehydrogenase Deficiency. in *GeneReviews(®)* (eds. Adam, M.P. *et al.*) (University of Washington, Seattle
Copyright © 1993-2024, University of Washington, Seattle. GeneReviews is a registered trademark of the University of Washington, Seattle. All rights reserved., Seattle (WA), 1993).
32. Elrod, J.W. & Molkentin, J.D. Physiologic functions of cyclophilin D and the mitochondrial permeability transition pore. *Circ J* **77**, 1111-22 (2013).
33. Proceedings of the Biochemical Society. *Biochem J* **72**, 27p-37p (1959).
34. Shi, H. *et al.* NAD Deficiency, Congenital Malformations, and Niacin Supplementation. *New England Journal of Medicine* **377**, 544-552 (2017).
35. Carter, M.T., Mirzaa, G., McDonell, L.M. & Boycott, K.M. Microcephaly-Capillary Malformation Syndrome. in *GeneReviews(®)* (eds. Adam, M.P. *et al.*) (University of Washington, Seattle
Copyright © 1993-2024, University of Washington, Seattle. GeneReviews is a registered trademark of the University of Washington, Seattle. All rights reserved., Seattle (WA), 1993).
36. Luo, S. *et al.* NAT8 Variants, N-Acetylated Amino Acids, and Progression of CKD. *Clin J Am Soc Nephrol* **16**, 37-47 (2020).
37. Maas, M.N., Hintzen, J.C.J., Porzberg, M.R.B. & Mecinović, J. Trimethyllysine: From Carnitine Biosynthesis to Epigenetics. *Int J Mol Sci* **21**(2020).
38. Nicholson, G. *et al.* A genome-wide metabolic QTL analysis in Europeans implicates two loci shaped by recent positive selection. *PLoS Genet* **7**, e1002270 (2011).
39. Montoliu, I. *et al.* Current status on genome-metabolome-wide associations: an opportunity in nutrition research. *Genes Nutr* **8**, 19-27 (2013).
40. Wang, T. *et al.* Pyridine nucleotide-disulphide oxidoreductase domain 2 (PYROXD2): Role in mitochondrial function. *Mitochondrion* **47**, 114-124 (2019).
41. Guo, Y. *et al.* Genetic analysis of impaired trimethylamine metabolism using whole exome sequencing. *BMC Med Genet* **18**, 11 (2017).

medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314800; this version posted October 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

42. Cheng, Y. *et al.* Rare genetic variants affecting urine metabolite levels link population variation to inborn errors of metabolism. *Nat Commun* **12**, 964 (2021).
43. Turongkaravee, S. *et al.* A systematic review and meta-analysis of genotype-based and individualized data analysis of SLCO1B1 gene and statin-induced myopathy. *Pharmacogenomics J* **21**, 296-307 (2021).
44. Moyer, A.M. *et al.* SLCO1B1 genetic variation and hormone therapy in menopausal women. *Menopause* **25**, 877-882 (2018).
45. Dudenkov, T.M. *et al.* SLCO1B1 polymorphisms and plasma estrone conjugates in postmenopausal women with ER+ breast cancer: genome-wide association studies of the estrone pathway. *Breast Cancer Res Treat* **164**, 189-199 (2017).
46. Johnson, N. *et al.* CYP3A7*1C allele: linking premenopausal oestrone and progesterone levels with risk of hormone receptor-positive breast cancers. *Br J Cancer* **124**, 842-854 (2021).
47. Johnson, N. *et al.* Cytochrome P450 Allele CYP3A7*1C Associates with Adverse Outcomes in Chronic Lymphocytic Leukemia, Breast, and Lung Cancer. *Cancer Res* **76**, 1485-1493 (2016).
48. Turner, R.M. *et al.* A Genome-wide Association Study of Circulating Levels of Atorvastatin and Its Major Metabolites. *Clin Pharmacol Ther* **108**, 287-297 (2020).
49. Casirati, A. *et al.* Preterm birth and metabolic implications on later life: A narrative review focused on body composition. *Front Nutr* **9**, 978271 (2022).
50. Sun, B.B. *et al.* Genomic atlas of the human plasma proteome. *Nature* **558**, 73-79 (2018).
51. Zhu, J. *et al.* Associations between Genetically Predicted Circulating Protein Concentrations and Endometrial Cancer Risk. *Cancers (Basel)* **13**(2021).
52. Salehi, B. *et al.* Allicin and health: A comprehensive review. *Trends in Food Science & Technology* **86**, 502-516 (2019).
53. Deng, Y., Ho, C.-T., Lan, Y., Xiao, J. & Lu, M. Bioavailability, Health Benefits, and Delivery Systems of Allicin: A Review. *Journal of Agricultural and Food Chemistry* **71**, 19207-19220 (2023).
54. Harvey, M.H., McMillan, M., Morgan, M.R. & Chan, H.W. Solanidine is present in sera of healthy individuals and in amounts dependent on their dietary potato consumption. *Hum Toxicol* **4**, 187-94 (1985).
55. Friedman, M. Potato glycoalkaloids and metabolites: roles in the plant and in the diet. *J Agric Food Chem* **54**, 8655-81 (2006).
56. Magliocco, G. *et al.* Metabolomics reveals biomarkers in human urine and plasma to predict cytochrome P450 2D6 (CYP2D6) activity. *Br J Pharmacol* **178**, 4708-4725 (2021).
57. Gaedigk, A. Complexities of CYP2D6 gene analysis and interpretation. *Int Rev Psychiatry* **25**, 534-53 (2013).
58. Wollmann, B.M. *et al.* Evidence for solanidine as a dietary CYP2D6 biomarker: Significant correlation with risperidone metabolism. *Br J Clin Pharmacol* **90**, 740-747 (2024).
59. Wollmann, B.M., Størset, E., Kringen, M.K., Molden, E. & Smith, R.L. Prediction of CYP2D6 poor metabolizers by measurements of solanidine and metabolites-a study in 839 patients with known CYP2D6 genotype. *Eur J Clin Pharmacol* **79**, 523-531 (2023).
60. Kiiski, J.I. *et al.* Solanidine is a sensitive and specific dietary biomarker for CYP2D6 activity. *Human Genomics* **18**, 11 (2024).
61. Haugabrooks, E. Chapter 1 - The interrelationships between food, nutrition, and toxicology. in *History of Food and Nutrition Toxicology* (eds. Haugabrooks, E. & Hayes, A.W.) 1-31 (Academic Press, 2023).
62. Daubner, S.C., Le, T. & Wang, S. Tyrosine hydroxylase and regulation of dopamine synthesis. *Arch Biochem Biophys* **508**, 1-12 (2011).
63. Sidharthan, N.P., Minchin, R.F. & Butcher, N.J. Cytosolic sulfotransferase 1A3 is induced by dopamine and protects neuronal cells from dopamine toxicity: role of D1 receptor-N-methyl-D-aspartate receptor coupling. *J Biol Chem* **288**, 34364-74 (2013).
64. Yu, Z. *et al.* Differences between human plasma and serum metabolite profiles. *PLoS One* **6**, e21230 (2011).
65. de Mutsert, R. *et al.* The Netherlands Epidemiology of Obesity (NEO) study: study design and data collection. *Eur J Epidemiol* **28**, 513-23 (2013).
66. Bos, M.M. *et al.* Investigating the relationships between unfavourable habitual sleep and metabolomic traits: evidence from multi-cohort multivariable regression and Mendelian randomization analyses. *BMC Med* **19**, 69 (2021).
67. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* **48**, 1279-83 (2016).
68. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904-9 (2006).
69. Howie, B.N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**, e1000529 (2009).
70. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).

medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314800; this version posted October 7, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

71. Mbatchou, J. *et al.* Computationally efficient whole-genome regression for quantitative and binary traits. *Nat Genet* **53**, 1097-1103 (2021).
72. Li, J. & Ji, L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity (Edinb)* **95**, 221-7 (2005).
73. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190-1 (2010).
74. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun* **8**, 1826 (2017).
75. Chang, L., Zhou, G., Ou, H. & Xia, J. mGWAS-Explorer: Linking SNPs, Genes, Metabolites, and Diseases for Functional Insights. *Metabolites* **12**(2022).
76. Kamat, M.A. *et al.* PhenoScanner V2: an expanded tool for searching human genotype-phenotype associations. *Bioinformatics* **35**, 4851-4853 (2019).
77. Stacey, D. *et al.* ProGeM: a framework for the prioritization of candidate causal genes at molecular quantitative trait loci. *Nucleic Acids Res* **47**, e3 (2019).
78. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122 (2016).
79. McCall, M.N., Illei, P.B. & Halushka, M.K. Complex Sources of Variation in Tissue Expression Data: Analysis of the GTEx Lung Transcriptome. *Am J Hum Genet* **99**, 624-635 (2016).
80. Bastian, M., Heymann, S. & Jacomy, M. Gephi: An Open Source Software for Exploring and Manipulating Networks. *Proceedings of the International AAAI Conference on Web and Social Media* **3**, 361-362 (2009).

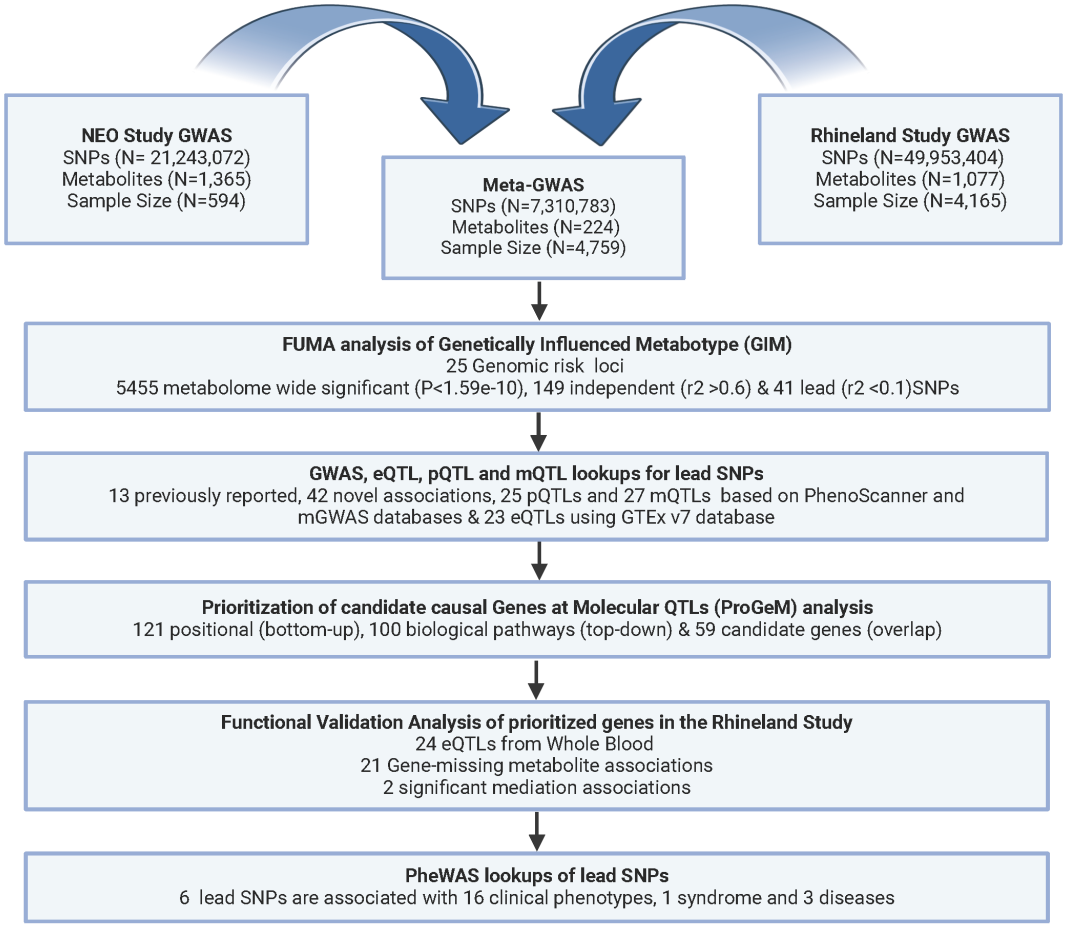
medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314800; this version posted October 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Tables and Figures

Table 1: Overview of the sample characteristics and general genotype assay information.

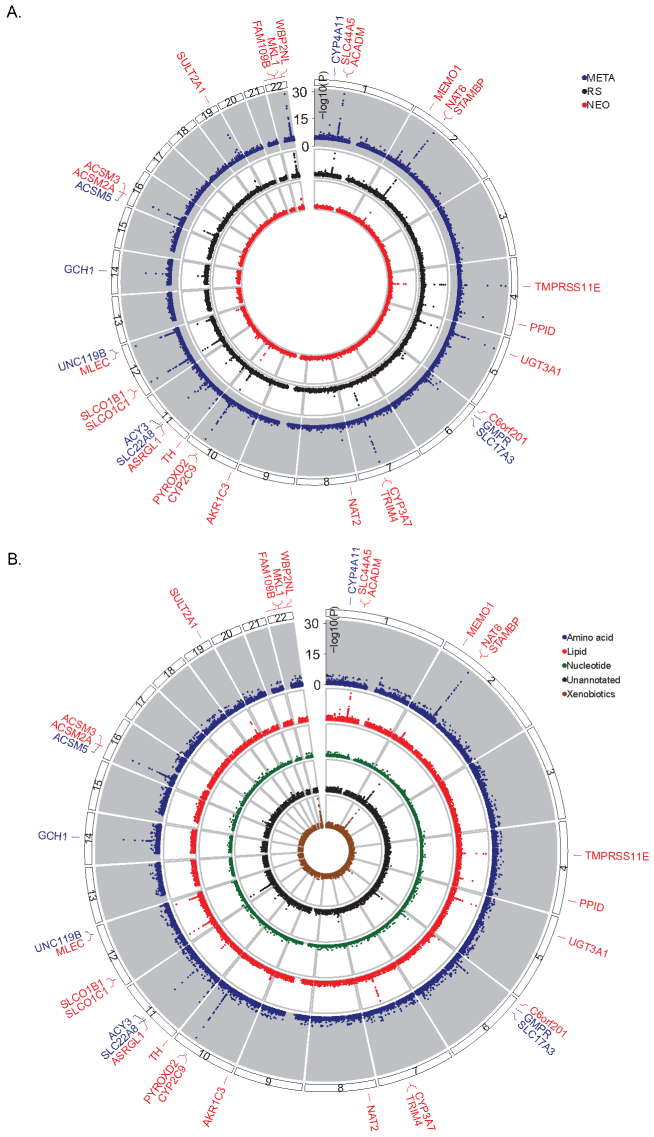
Characteristic	NEO Study (n= 594)	Rhineland Study (n= 4165)
Age: mean (sd)	55.8 (4.5)	55.5 (13.96)
Sex: % female	53%	56%
Metabolomics Platform	Metabolon-HD4	Metabolon-HD4
N of metabolites with missing values	433	340
Blood sampling	Serum	Plasma
Fasting status %	100%	99.4%
Genotyping Array	Illumina HumanCoreExome-24 BeadChip	Omni 2.5 Exome Array
Genotype Imputation Panel	HRC (r1.1)	1000 genome phase 3 version 5

Figure 1: Workflow of the GWAS and post-GWAS analyses of missing metabolite measures



medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314800; this version posted October 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Figure 2: Genomic loci associated with missingness of metabolites.



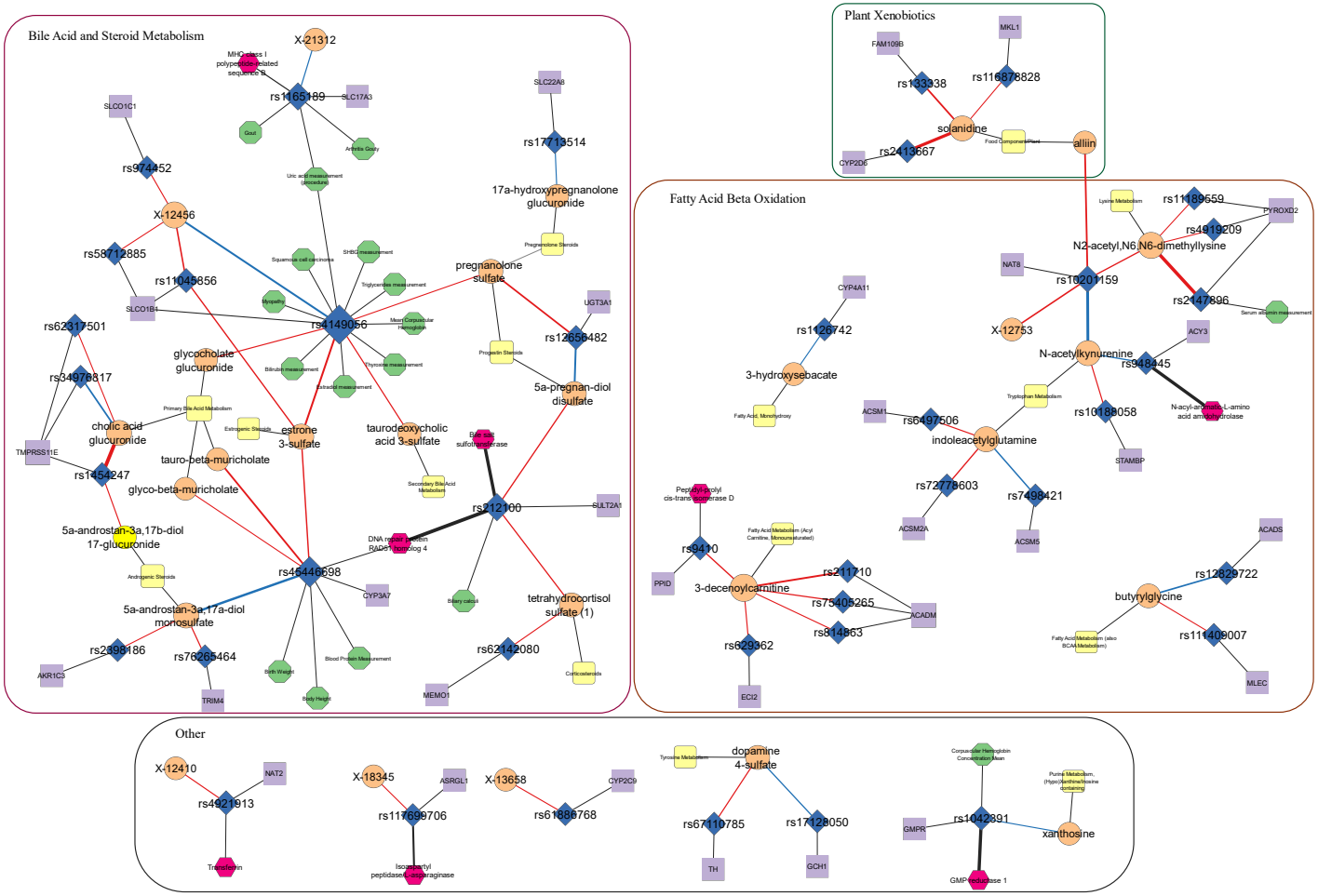
A, Circular Manhattan plot showing both cohort-specific and meta-analysis GWAS results. **B,** Circular Manhattan plot showing regional associations of genomic locations per metabolite class. The color of the genes indicates whether the identified lead SNP locus is novel. Dark blue indicates lead SNP loci (r2<0.1) that were previously reported, and red lead SNP loci that were identified to be novel with metabolites. The p-value axis is truncated at $p<1\times10^{-30}$ for visualization purposes. All GWAS models were adjusted for age, sex, batch, fasting status and population substructure principal components.

Table 2: Novel and previously reported lead SNP-Metabolite associations using PhenoScanner

Sub Pathway	Metabolite Name	SNP	nearestGene	Beta (SE)	P value	Novelty	HMDB*
Androgenic Steroids	5alpha-androstan-3alpha,17alpha-diol monosulfate	rs2398186	*AKR1C3*	0.4495 (0.0684)	5.004×10^{-11}	Novel	HMDB0000458
		rs45446698	*CYP3A7*	-2.3785 (0.1532)	2.441×10^{-54}	Reported	
		rs76265464	*TRIM4*	-2.014 (0.249)	6.07×10^{-16}	Novel	
	5alpha-androstan-3alpha,17beta-diol 17-glucuronide	rs1454247	*TMPRSS11E*	-0.5275 (0.0717)	1.931×10^{-13}	Novel	
Corticosteroids	tetrahydrocortisol sulfate (1)	rs212100	*SULT2A1*	0.7805 (0.0898)	3.556×10^{-18}	Novel	HMDB0000949
		rs62142080	*MEMO1*	0.4637 (0.0657)	1.636×10^{-12}	Novel	
Estrogenic Steroids	estrone 3-sulfate	rs11045856	*SLCO1B1*	-0.5496 (0.0584)	4.551×10^{-21}	Novel	HMDB01425
		rs4149056	*SLCO1B1*	0.8336 (0.0634)	1.853×10^{-39}	Novel	
		rs45446698	*CYP3A7*	-1.9913 (0.2056)	3.486×10^{-22}	Novel	
Fatty Acid Metabolism (Acyl Carnitine, Monounsaturated)	3-decenoylcarnitine	rs211710	*SLC44A5*	-0.7772 (0.0612)	5.372×10^{-37}	Novel	HMDB0241067
		rs629362	*C6orf201*	0.4453 (0.0597)	8.92×10^{-14}	Novel	
		rs75405265	*ACADM*	-1.3468 (0.1613)	6.778×10^{-17}	Novel	
		rs814863	*SLC44A5*	-0.617 (0.0883)	2.75×10^{-12}	Novel	
		rs9410	*PPID*	0.5561 (0.0604)	3.23×10^{-20}	Novel	
Fatty Acid Metabolism (also BCAA Metabolism)	butyrylglycine	rs111409007	*MLEC*	0.4744 (0.0722)	5.008×10^{-11}	Novel	HMDB00808
		rs12829722	*UNC119B*	0.6893 (0.0605)	4.881×10^{-30}	Reported	
Fatty Acid, Monohydroxy	3-hydroxysebacate	rs1126742	*CYP4A11*	-0.5556 (0.078)	1.085×10^{-12}	Reported	HMDB0340579
Food Component/Plant	alliin	rs10201159	*NAT8*	-0.7113 (0.0588)	1.075×10^{-33}	Novel	HMDB33592
	solanidine	rs116878828	*MKL1*	0.582 (0.0893)	7.267×10^{-11}	Novel	HMDB03236
		rs133338	*WBP2NL*	0.7751 (0.0648)	5.661×10^{-33}	Novel	
		rs2413667	*FAM109B*	1.1182 (0.0608)	2.024×10^{-75}	Novel	
Lysine Metabolism	N2-acetyl,N6,N6-dimethyllysine	rs10201159	*NAT8*	0.8927 (0.0986)	1.334×10^{-19}	Novel	
		rs11189559	*PYROXD2*	0.4777 (0.0743)	1.317×10^{-10}	Novel	
		rs2147896	*PYROXD2*	-1.2699 (0.0649)	3.051×10^{-85}	Novel	
		rs4919209	*PYROXD2*	0.4708 (0.0724)	8.003×10^{-11}	Novel	
Pregnenolone Steroids	17alpha-hydroxypregnanolone glucuronide	rs17713514	*SLC22A8*	1.5434 (0.2094)	1.717×10^{-13}	Reported	HMDB0000363
Primary Bile Acid Metabolism	cholic acid glucuronide	rs1454247	*TMPRSS11E*	-1.038 (0.0502)	8.29×10^{-95}	Novel	HMDB0002577
		rs34976817	*TMPRSS11E*	-0.9448 (0.0802)	4.947×10^{-32}	Reported	
		rs62317501	*TMPRSS11E*	-0.4206 (0.0638)	4.458×10^{-11}	Novel	
	glyco-beta-muricholate	rs45446698	*CYP3A7*	1.8743 (0.2818)	2.888×10^{-11}	Novel	HMDB0341323
	glycocholate glucuronide	rs4149056	*SLCO1B1*	0.4621 (0.0607)	2.718×10^{-14}	Novel	HMDB0341324
	tauro-beta-muricholate	rs45446698	*CYP3A7*	2.0306 (0.1632)	1.526×10^{-35}	Novel	HMDB0000932
Progestin Steroids	5alpha-pregnan-diol disulfate	rs12656482	*UGT3A1*	1.0108 (0.0788)	1.052×10^{-37}	Reported	
		rs212100	*SULT2A1*	0.6352 (0.0697)	8.317×10^{-20}	Novel	
	pregnanolone/allopregnanolone sulfate	rs12656482	*UGT3A1*	0.8004 (0.0678)	3.371×10^{-32}	Novel	HMDB0062782
		rs4149056	*SLCO1B1*	0.4155 (0.0618)	1.834×10^{-11}	Novel	
Purine Metabolism	xanthosine	rs1042391	*GMPR*	0.3338 (0.0467)	8.819×10^{-13}	Reported	HMDB0000299

Sub Pathway	Metabolite Name	SNP	nearestGene	Beta (SE)	P value	Novelty	HMDB*
Secondary Bile Acid Metabolism	taurodeoxycholic acid 3-sulfate	rs4149056	*SLCO1B1*	0.6427 (0.0784)	2.525×10^{-16}	Novel	HMDB0240734
Tryptophan Metabolism	indoleacetylglutamine	rs6497506	*ACSM3*	0.5559 (0.0713)	6.353×10^{-15}	Novel	HMDB0013240
		rs72778603	*ACSM2A*	-1.2624 (0.1369)	2.882×10^{-20}	Novel	
		rs7498421	*ACSM5*	0.6891 (0.0686)	9.52×10^{-24}	Reported	
	N-acetylkynurenine	rs10188058	*STAMBP*	0.6201 (0.0793)	5.423×10^{-15}	Novel	HMDB0240342
		rs10201159	*NAT8*	1.2782 (0.0738)	3.755×10^{-67}	Reported	
		rs948445	*ACY3*	0.6161 (0.0677)	9.101×10^{-20}	Reported	
Tyrosine Metabolism	dopamine 4-sulfate	rs17128050	*GCH1*	-0.6634 (0.0754)	1.447×10^{-18}	Reported	HMDB0004148
		rs67110785	*TH*	0.4338 (0.0537)	6.913×10^{-16}	Novel	
	X-12410	rs4921913	*NAT2*	-0.6524 (0.087)	6.57×10^{-14}	Novel	
	X-12456	rs11045856	*SLCO1B1*	-0.5904 (0.0554)	1.635×10^{-26}	Novel	
		rs4149056	*SLCO1B1*	0.8525 (0.0647)	1.098×10^{-39}	Reported	
		rs58712885	*SLCO1B1*	-0.6786 (0.1035)	5.548×10^{-11}	Novel	
		rs974452	*SLCO1C1*	-0.5143 (0.0723)	1.119×10^{-12}	Novel	
	X-12753	rs10201159	*NAT8*	0.5979 (0.055)	1.729×10^{-27}	Novel	
	X-13658	rs61886768	*CYP2C9*	0.4206 (0.0598)	2.066×10^{-12}	Novel	
	X-18345	rs117699706	*ASRGL1*	-0.5411 (0.0802)	1.537×10^{-11}	Novel	
	X-21312	rs1165189	*SLC17A3*	0.477 (0.0604)	2.751×10^{-15}	Reported	

Figure 3: Network representation of the 41 lead SNP-metabolite associations.



The network shows the associations between the 41 lead SNPs (blue diamonds) and metabolites (orange circles). The network also includes the mapped genes to each SNP (purple squares), assigned sub-pathway from the measurement platform (yellow rounded squares), traits and diseases associated with the SNPs from DisGeNET (green octagons), and pQTL associations from Phenoscanner (pink hexagons). Novel and reported associations between the SNP-metabolites are represented in red- and orange-colored lines respectively.

Figure 4: Summary of top ranked candidate gene identification, eQTL, and pQTL associations per identified SNP-metabolite.

SNP	Genes	Positional	Functional	eQTL	pQTL
rs1126742	*CYP4A11*				
	CYP4A22				
	CYP4B1				
rs814863	*ACADM*				
	ST6GALNAC3				
rs211710	*ACADM*				
	SLC44A5				
rs75405265	*ACADM*				
	ST6GALNAC3				
rs62142080	*SPAST*				
rs10201159	*NAT8*				
	DGUOK				
	ALMS1				
	TPRKB				
rs62317501	*YTHDC1*				
rs1454247	*UGT2B15*				
rs9410	*ETFDH*				
	PPID				
rs12656482	*UGT3A1*				
	UGT3A2				
rs629362	*ECI2*				
rs1042391	*GMPR*				
	MYLIP				
rs1165189	*SLC17A3*				
	HFE				
	SLC17A4				
	SLC17A1				
	SLC17A2				
rs45446698	*ATP5J2*				
	CYP3A4				
	CYP3A7				
	CYP3A5				

SNP	Genes	Positional	Functional	eQTL	pQTL
rs76265464	*TRIM4*				
	CYP3A43				
rs4921913	*PSD3*				
	NAT1				
	NAT2				
rs2398186	*AKR1C4*				
	AKR1C3				
	AKR1C2				
rs61886768	*CYP2C8*				
	C10orf129				
	CYP2C19				
	CYP2C9				
rs11189559	*PYROXD2*				
rs2147896	*PYROXD2*				
	HPS1				
	HPSE2				
rs67110785	*INS*				
	TH				
rs117699706	*ASRGL1*				
rs948445	*ALDH3B2*				
	ACY3				
rs974452	*SLCO1B3*				
	PDE3A				
	SLCO1C1				
rs4149056	*SLCO1A2*				
	SLCO1B1				
rs58712885	*SLCO1A2*				
	SLCO1B1				
rs11045856	*SLCO1A2*				
	SLCO1B1				
	LDHB				

SNP	Genes	Positional	Functional	eQTL	pQTL
rs111409007	*ACADS*				
	COQ5				
	MLEC				
	UNC119B				
	GATC				
rs12829722	*ACADS*				
	MLEC				
rs17128050	*GCH1*				
rs7498421	*ACSM5*				
	ACSM1				
	ACSM2A				
rs72778603	*ACSM2A*				
	ACSM5				
rs6497506	*ACSM3*				
	ACSM2B				
	ACSM5				
	ACSM1				
rs212100	*SULT2A1*				
rs116878828	*MCHR1*				
rs133338	*CYP2D6*				
	NAGA				
rs2413667	*CYP2D6*				
	NAGA				
	SREBF2				
	SMDT1				

Candidate genes identified by the ProGeM positional approach (bottom-up) are highlighted in orange under the "positional" column and highlighted in yellow under the "functional" column based on metabolic and phenotypic relevance approach (top-down). Genes with a significant eQTL association are highlighted in green under the eQTL column and in blue for SNP-pQTL associations under the pQTL column.

Table 3: Top pQTL associations with the lead SNPs

Metabolite Name	Sub Pathway	Lead SNP	Nearest gene	Proxy pQTL SNPs*	pQTL associated trait (Phenoscanner)	pQTL associated trait (UK biobank)
3-decenoylcarnitine	Fatty Acid Metabolism (Acyl Carnitine. Monounsaturated)	rs9410	*PPID*	rs8396	Peptidyl-prolyl cis-trans isomerase D	-
X-21312	-	rs1165189	*SLC17A3*	rs13200784, rs3757132	MHC class I polypeptide-related sequence B	-
estrone 3-sulfate	Estrogenic Steroids	rs45446698	*CYP3A7*	rs45446698	DNA repair protein RAD51 homolog 4	-
tauro-beta-muricholate	Primary Bile Acid Metabolism					-
5a-Androstane-3a,17a-diol monosulfate	Androgenic Steroids					-
glyco-beta-muricholate	Primary Bile Acid Metabolism					-
X-12410	-	rs4921913	*NAT2*	rs4921915	Transferrin	-
5a-Androstane-3a,17a-diol disulfate	Progestin Steroids	rs212100	*SULT2A1*	rs212100	Bile salt sulfotransferase	SULT2A1
tetrahydrocortisol sulfate	Corticosteroids				DNA repair protein RAD51 homolog 4	
X-18345	-	rs117699706	*ASRGL1*	-	-	Isoaspartyl peptidase/L-asparaginase
N-acetylkynurenine (2)	Amino acid	rs948445	*ACY3*	-	-	N-acyl-aromatic-L-amino acid amidohydrolase (carboxylate-forming)
xanthosine	Nucleotide	rs1042391	*GMPR*	-	-	GMP reductase 1

*Proxy SNPs selected based on their correlation (r^2) with the lead SNP. Full results available in Supplementary Table 7.

Table 4: eQTL mediation of the association between the lead SNPs and missing metabolites by primary candidate genes.

Exposure	Mediator	Metabolite	Indirect effect		Direct effect		Total effect		Percentage of mediation[±]
			β estimate (SE)	P value	β estimate (SE)	P value	β estimate (SE)	P value	
rs117699706	*ASRGL1*	X-18345	-0.030 (0.023)	0.212	-0.30 (0.067)	6.6×10^{-06}	-0.333 (0.062)	9.6×10^{-08}	8.9%
rs629362	*ECI2*	3-Decenoylcarnitine	-0.015 (0.013)	0.259	-0.236 (0.047)	7.6×10^{-07}	-0.252 (0.045)	2.9×10^{-08}	6.2%
rs10201159	*ALMS1*	X-12753	-0.025 (0.025)	0.320	-0.407 (0.053)	1.8×10^{-14}	-0.433 (0.046)	$<10\times10^{-16}$	5.9%
rs814863	*ACADM*	3-Decenoylcarnitine	-0.011 (0.007)	0.172	-0.319 (0.063)	4.9×10^{-07}	-0.330 (0.063)	2.0×10^{-07}	3.2%
rs2413667	*SMDT1*	Solanidine	-0.024 (0.012)	0.055*	-0.722 (0.050)	$<10\times10^{-16}$	-0.746 (0.047)	$<10\times10^{-16}$	3.2%
rs10201159_T	*ALMS1*	N2-acetyl.N6.N6-dimethyllysine	-0.015 (0.035)	0.684	-0.495 (0.074)	2.9×10^{-11}	-0.509 (0.066)	1.3×10^{-14}	2.9%
rs2147896	*HPS1*	N2-acetyl.N6.N6-dimethyllysine	-0.022 (0.007)	0.004*	-0.745 (0.046)	$<10\times10^{-16}$	-0.766 (0.046)	$<10\times10^{-16}$	2.8%
rs11189559_G	*PYROXD2*	N2-acetyl.N6.N6-dimethyllysine	-0.005 (0.004)	0.176	-0.236 (0.049)	1.9×10^{-06}	-0.241 (0.049)	1.1×10^{-06}	2.2%
rs211710	*ACADM*	3-Decenoylcarnitine	-0.002 (0.006)	0.731	-0.556 (0.048)	$<10\times10^{-16}$	-0.557 (0.048)	$<10\times10^{-16}$	0.4%

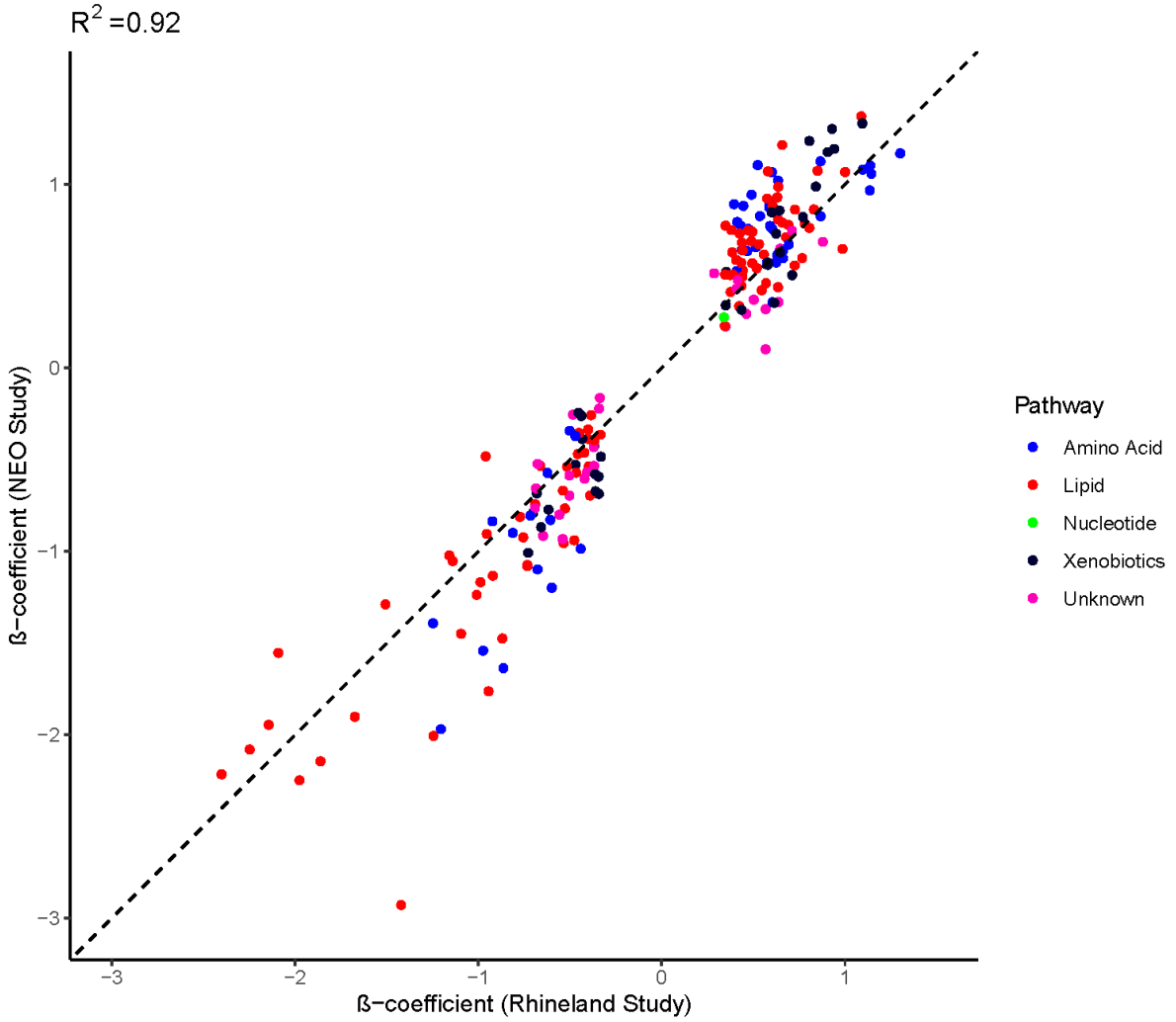
*Indirect effect's P value =< 0.05

[±] (Indirect effect / total effect) * 100

medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314800; this version posted October 7, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Extended Data Figures

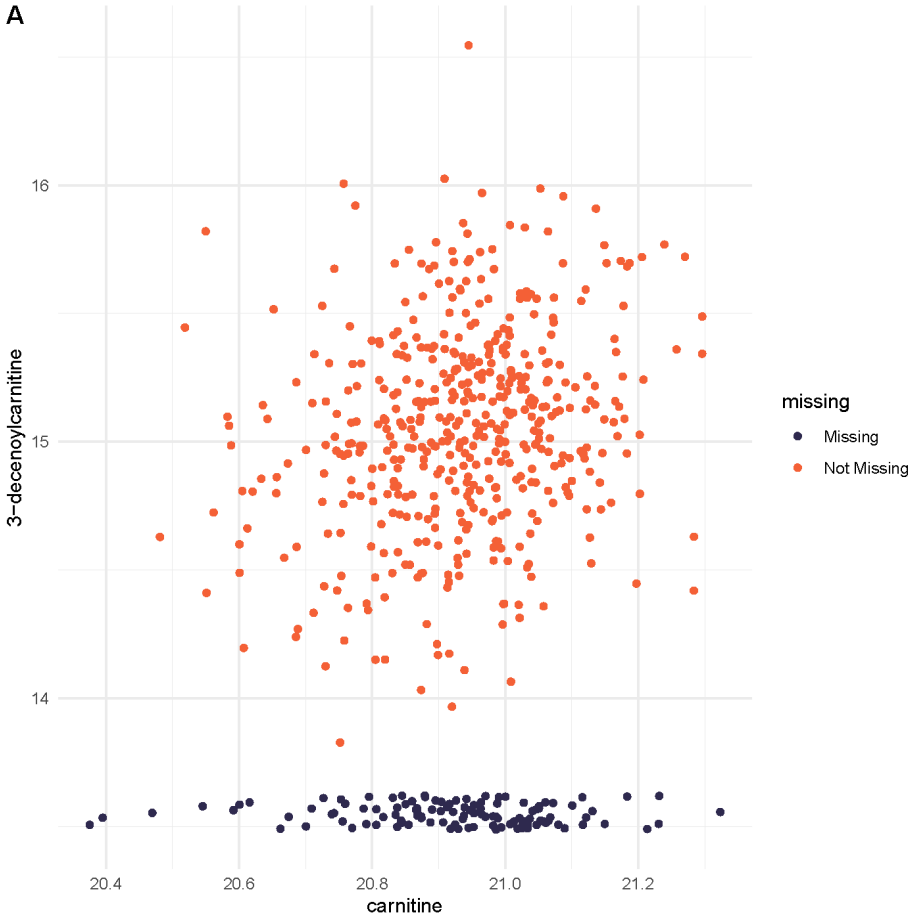
Extended Data Fig. 1 Correlation between the GWAS meta-analysis independent SNP (r^2<0.6)-metabolite effect sizes from the Rhineland Study and the NEO Study

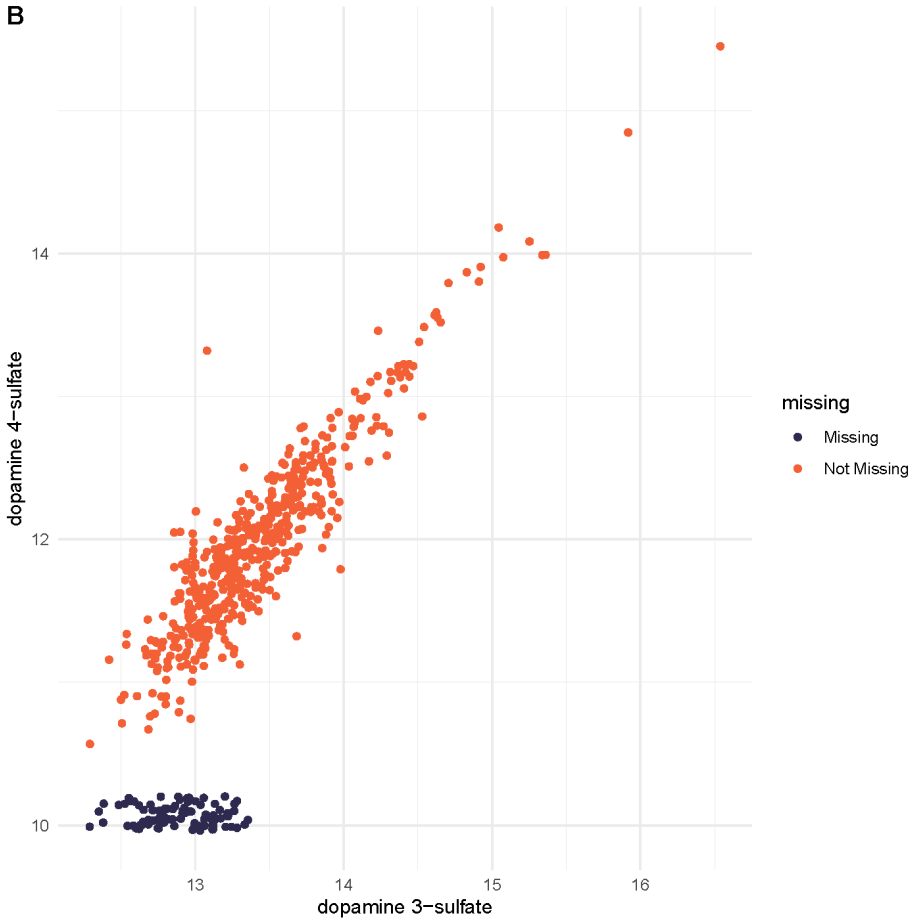


medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314800; this version posted October 7, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

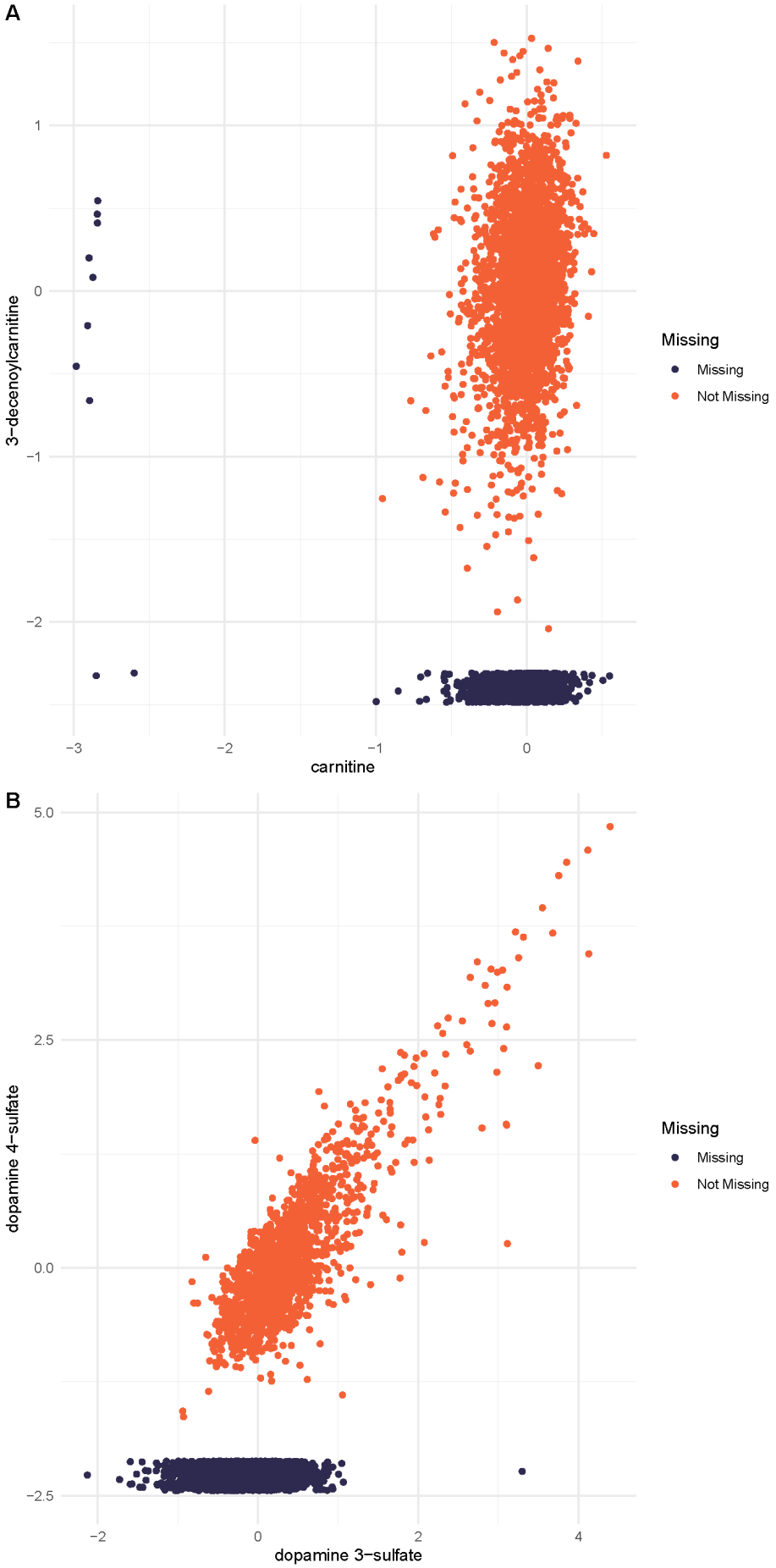
Extended Data Fig. 2 Correlation and visualization of missingness of 3-decenoylcarnitine and dopamine 4-sulphate in relation to carnitine and dopamine 3-sulphate in the NEO study

medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314800; this version posted October 7, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .





medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314800; this version posted October 7, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Extended Data Fig. 3 Correlation and visualization of missingness of 3-decenoylcarnitine and dopamine 4-sulphate in relation to carnitine and dopamine 3-sulphate in the Rhineland Study

medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314800; this version posted October 7, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .