



Universiteit  
Leiden  
The Netherlands

## Genome-wide association studies of missing metabolite measures from two population-based studies

Faquih, T.O.; Imtiaz, M.A.; Talevi, V.; Landstra, E.N.; Vlieg, A.V.; Li-Gao, R.; ... ; Dijk, K.W. van

### Citation

Faquih, T. O., Imtiaz, M. A., Talevi, V., Landstra, E. N., Vlieg, A. V., Li-Gao, R., ... Dijk, K. W. van. (2026). Genome-wide association studies of missing metabolite measures from two population-based studies. *Genome Biology*, 27(1). doi:10.1186/s13059-026-04132-9

Version: Publisher's Version

License: [Creative Commons CC BY 4.0 license](#)

Downloaded from: <https://hdl.handle.net/1887/4308218>

**Note:** To cite this publication please use the final published version (if applicable).

RESEARCH

Open Access



# Genome-wide association studies of missing metabolite measures from two population-based studies

Tariq O. Faquihi<sup>1,2,3†</sup>, Mohammed Aslam Imtiaz<sup>4†</sup>, Valentina Talevi<sup>4</sup>, Elvire N. Landstra<sup>4</sup>, Astrid van Hylckama Vlieg<sup>1</sup>, Ruifang Li-Gao<sup>1</sup>, Frits R. Rosendaal<sup>1</sup>, Raymond Noordam<sup>5</sup>, Diana van Heemst<sup>5</sup>, Dennis O. Mook-Kanamori<sup>1,6</sup>, Monique M. B. Breteler<sup>4,7†</sup>, N. Ahmad Aziz<sup>4,7,8\*†</sup> and Ko Willems van Dijk<sup>9,10,11\*†</sup>

<sup>†</sup>Tariq O. Faquihi and Mohammed Aslam Imtiaz contributed equally as first authors.

<sup>†</sup>Monique M. B. Breteler, N. Ahmad Aziz and Ko Willems van Dijk contributed equally to this work.

\*Correspondence: Ahmad.Aziz@dzne.de; k.willems\_van\_dijk@lumc.nl

<sup>4</sup> German Centre for Neurodegenerative Diseases (DZNE), Population Health Sciences, Bonn, Germany

<sup>9</sup> Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands

Full list of author information is available at the end of the article

## Abstract

**Background:** Metabolomic analyses are increasingly applied in both etiological and predictive research, but frequently report missing values, which are then either imputed or removed from the analyses, and not examined as true missingness due to altered metabolism. We hypothesized that interindividual genetic variation may account for part of this missingness.

**Results:** We perform a logistic GWAS of metabolite missingness from an untargeted mass spectrometry-based platform in the Netherlands Epidemiology of Obesity Study ( $N=594$ ) and the Rhineland Study ( $N=4,165$ ). We consider 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 ( $p < 1.58 \times 10^{-10}$ ), involving 28 metabolites and 41 lead SNPs.

**Conclusions:** Despite considerable pleiotropy, the majority of identified SNP- 'missing metabolite' associations are biologically plausible, relating to beta-oxidation, bile acids, steroids, and xenobiotics metabolism. These findings suggest that missing values in metabolomics are partially non-random and reflect potential genetic variation.

**Keywords:** Genome-wide association study, Missing measurements, Inborn errors of metabolism, Metabolomics, Poor metabolizers

## Background

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].



© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

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 that are designed to maximize the number of measured metabolites at the cost of lower specificity [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 ruled out that missing values of metabolites are caused by genetic variations. 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 2,500 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, are 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 the 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, technical errors, or truly absent) of metabolite measures.

## Results

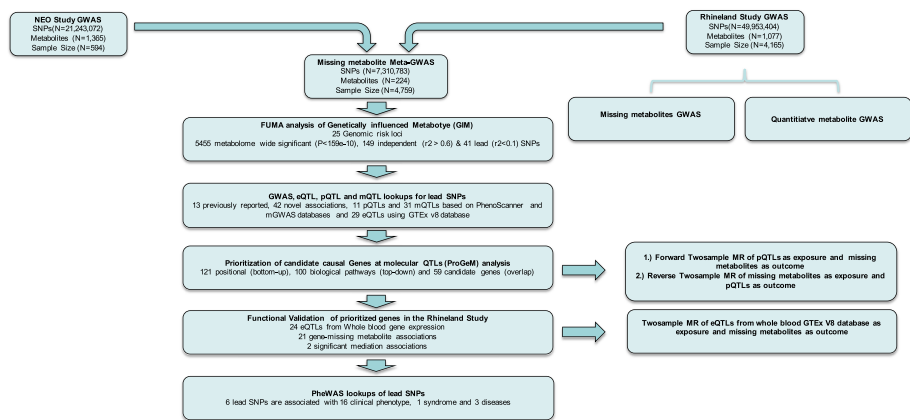
### Discovery genome-wide association studies of missing metabolites

The GWAS of missing metabolite measures was performed separately in 594 individuals from the 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% women). Individual study characteristics and general genotype assay information are summarized in Table 1. GWAS results in NEO identified 712

**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 (5.9)                          | 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 |

SD Standard deviation



**Fig. 1** Workflow of the GWAS and post-GWAS analyses of missing metabolite measures

metabolome-wide significant ( $p < 1.58 \times 10^{-10}$ ) associations between 537 SNPs and 6 out of the 341 included metabolites. In the Rhineland Study, we identified 4,370 metabolome-wide significant ( $p < 1.59 \times 10^{-10}$ ) 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 Additional file 2: Table S1 and Additional file 3: Table S2 for NEO and the Rhineland Study, respectively. In the sensitivity analysis adjusting for prevalent diabetes in the Rhineland Study, the main findings remained largely unchanged. However, genomic risk loci 2 (*ALMS1*) and 12 (*PYROXD2*) that were associated with N2-acetyl, N6, N6-dimethyllysine, and genomic risk locus 20 (*CRX*) that was associated with tetrahydrocortisol sulfate were no longer statistically significant. Among these, only N2-acetyl, N6, N6-dimethyllysine showed directional inconsistency in SNP effect size estimates between the primary model and the sensitivity analysis (Additional file 3: Table S2).

### Quantitative GWAS sensitivity analysis

For the metabolome-wide significant hits identified in missing metabolites, we assessed whether the associated loci had been replicated in quantitative GWAS. Seventeen out of twenty-six metabolites exhibited unique loci in the missing metabolite GWAS that were not observed in their corresponding quantitative GWAS. These metabolites included 3-decenoylcarnitine, 5 $\alpha$ -androstan-3 $\alpha$ ,17 $\alpha$ -diol monosulfate, 5 $\alpha$ -pregnan-3 $\alpha$ -diol disulfate, X-13658, X-12753, X-12410, alliin, dopamine 4-sulfate, estrone 3-sulfate, ferulic acid 4-sulfate, glycocholate glucuronide (1), N2-acetyl,N6,N6-dimethyllysine, N-acetylkynurenine (2), pregnanolone/allopregnanolone sulfate, and tetrahydrocortisol sulfate (Additional file 1: Fig. S1, Additional file 3: Table S2).

### Study-specific SNP $\times$ SNP interaction

For all 224 metabolites overlapping in both NEO and the Rhineland study, we did not observe any metabolome-wide significant SNP  $\times$  SNP interactions across the studies, except for 22 SNPs associated with 10 metabolites that reached genome-wide significance (Additional file 4: Table S3). Our results indicated that for the majority of SNPs among overlapping metabolites, no SNP  $\times$  SNP interactions were detected between the NEO and the Rhineland Study.

### Genome-wide meta-analysis of missing metabolites

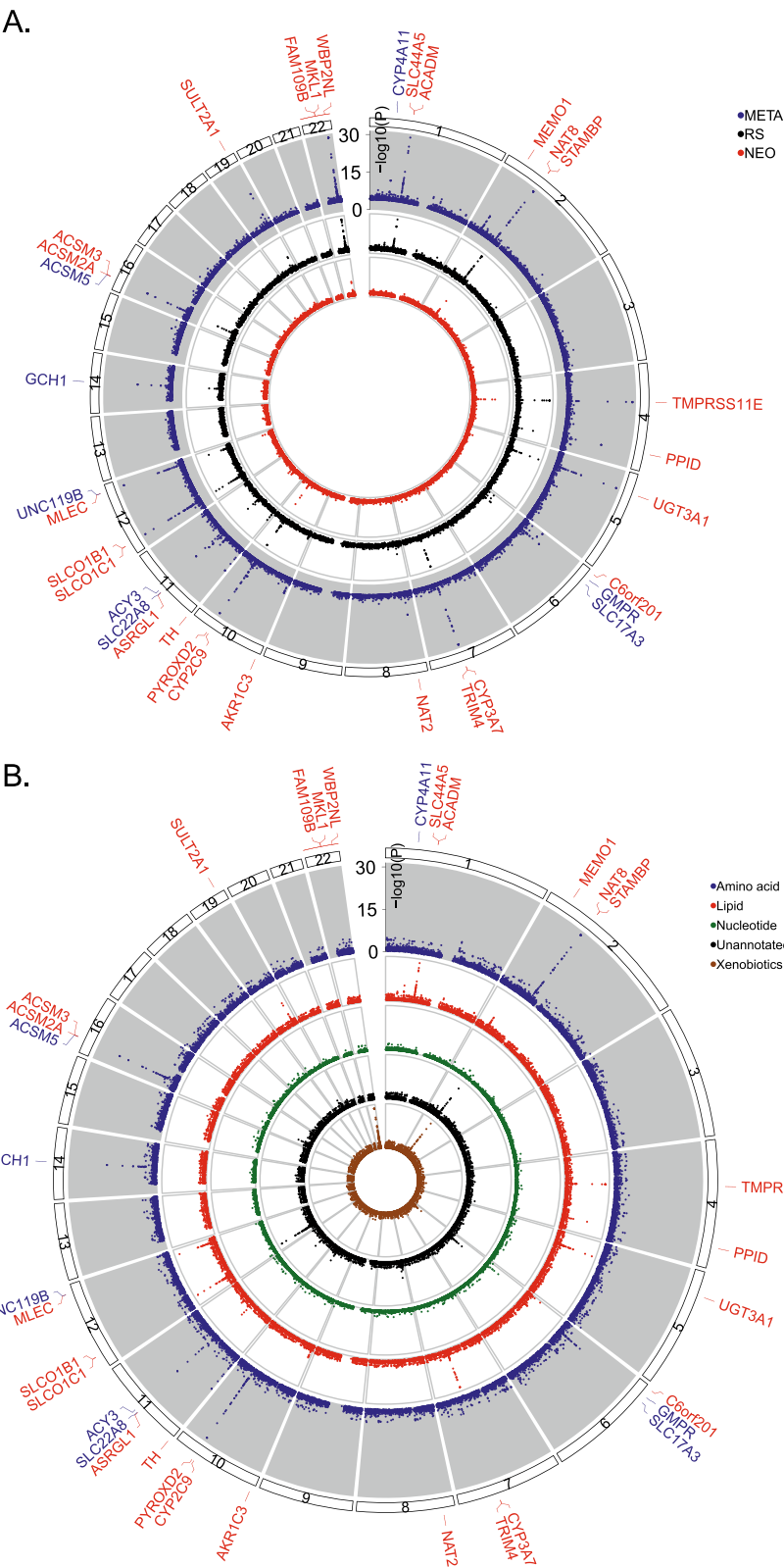
The meta-analysis of the overlapping 224 metabolites in the two studies identified 5,455 significant associations ( $p < 1.59 \times 10^{-10}$ ), including 3,260 SNPs across 33 different metabolites (Fig. 2 and Additional file 6: Table S5). The frequencies of missingness across the 224 metabolites in the NEO and Rhineland Study cohorts are provided in Additional file 5: Table S4. The direction of the associations was similar across both cohorts (Pearson correlation  $R^2 = 0.92$ ) using independent SNP-metabolite ( $r^2 < 0.6$ ) associations (Additional file 1: Fig. S2 and Additional file 7: Table S6). 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 metabolotypes

Genetically Influenced Metabotypes (GIMs), defined as variant-metabolite clusters [9], 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

(See figure on next page.)

**Fig. 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 ( $r^2 < 0.1$ ) that were previously reported, and red lead SNP loci that were identified to be novel with metabolites. The  $p$ -value axis is truncated at  $p < 1 \times 10^{-30}$  for visualization purposes. All GWAS models were adjusted for age, sex, batch, fasting status and population substructure principal components



**Fig. 2** (See legend on previous page.)

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 (Additional file 7: Table S6 and Additional file 8: Table S7). 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 Additional file 9: Table S8). 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 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 8), we identified 29 expression Quantitative Trait Loci (eQTLs) (Additional file 9: Table S8). 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 [22], in adipose 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 Additional file 10: Table S9). 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) (Additional file 11: Table S10). 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 causal 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, Additional file 12: Table S11 and Additional file 13: Table S12). 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 of which expression levels were available in 2,575 participants of the Rhineland Study. We

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

| Sub pathway                                             | Metabolite name                                    | Missing frequency NEO/RS                  | SNP         | nearestGene | Beta (SE)        | P value                 | Novelty  | HMDB        |
|---------------------------------------------------------|----------------------------------------------------|-------------------------------------------|-------------|-------------|------------------|-------------------------|----------|-------------|
| Androgenic Steroids                                     | 5alpha-androstan-3alpha,17alpha-diol monosulfate   | 23.4/27.4                                 | rs2398186   | AKR1C3      | 0.4495 (0.0684)  | $5.004 \times 10^{-11}$ | Novel    | HMD80000458 |
|                                                         |                                                    |                                           | rs45446698  | CYP3A7      | −2.3785 (0.1532) | $2.441 \times 10^{-54}$ | Reported |             |
|                                                         | 5alpha-androstan-3alpha,17beta-diol 17-glucuronide | 55.72/55.75                               | rs76265464  | TRIM4       | −2.014 (0.249)   | $6.07 \times 10^{-16}$  | Novel    |             |
|                                                         |                                                    |                                           | rs1454247   | TMPPRSS11E  | −0.5275 (0.0717) | $1.931 \times 10^{-13}$ | Novel    |             |
| Corticosteroids                                         | tetrahydrocortisol sulfate (1)                     | 57.74/48.49                               | rs212100    | SULT2A1     | 0.7805 (0.0898)  | $3.556 \times 10^{-18}$ | Novel    | HMD80000949 |
| Estrogenic Steroids                                     | estrone 3-sulfate                                  | 79.46/61.56                               | rs62142080  | MEMO1       | 0.4637 (0.0657)  | $1.636 \times 10^{-12}$ | Novel    | HMD801425   |
|                                                         |                                                    |                                           | rs11045856  | SLC01B1     | −0.5496 (0.0584) | $4.551 \times 10^{-21}$ | Novel    |             |
|                                                         |                                                    |                                           | rs4149056   | SLC01B1     | 0.8336 (0.0634)  | $1.853 \times 10^{-39}$ | Novel    |             |
|                                                         |                                                    |                                           | rs45446698  | CYP3A7      | −1.9913 (0.2056) | $3.486 \times 10^{-22}$ | Novel    |             |
| Fatty Acid Metabolism (Acyl Carnitine, Monounsaturated) | 3-decenoyl carnitine                               | 19.53/29.2                                | rs211710    | SLC44A5     | −0.7772 (0.0612) | $5.372 \times 10^{-37}$ | Novel    | HMD80241067 |
|                                                         |                                                    |                                           | rs629362    | C6orf201    | 0.4453 (0.0597)  | $8.92 \times 10^{-14}$  | Novel    |             |
|                                                         |                                                    |                                           | rs75405265  | ACADM       | −1.3468 (0.1613) | $6.778 \times 10^{-17}$ | Novel    |             |
|                                                         |                                                    |                                           | rs814863    | SLC44A5     | −0.617 (0.0883)  | $2.75 \times 10^{-12}$  | Novel    |             |
| Fatty Acid Metabolism (also BC-AA Metabolism)           | butyrylglycine                                     | 45.96/61.56                               | rs9410      | PPID        | 0.5561 (0.0604)  | $3.23 \times 10^{-20}$  | Novel    | HMD800808   |
|                                                         |                                                    |                                           | rs111409007 | MLEC        | 0.4744 (0.0722)  | $5.008 \times 10^{-11}$ | Novel    |             |
|                                                         |                                                    |                                           | rs12829722  | UNC1198     | 0.6893 (0.0605)  | $4.881 \times 10^{-30}$ | Reported |             |
|                                                         |                                                    |                                           | rs1126742   | CYP4A11     | −0.5556 (0.078)  | $1.085 \times 10^{-12}$ | Reported |             |
| Fatty Acid, Monohydroxy Food Component/Plant            | 3-hydroxysebacate alliin solanidine                | 10.77/17.14<br>67.68/60.48<br>80.13/73.88 | rs10201159  | NAT8        | −0.7113 (0.0588) | $1.075 \times 10^{-33}$ | Novel    | HMD80340579 |
|                                                         |                                                    |                                           | rs116878828 | MKL1        | 0.582 (0.0893)   | $7.267 \times 10^{-11}$ | Novel    | HMD833592   |
|                                                         |                                                    |                                           | rs133338    | WBP2NL      | 0.7751 (0.0648)  | $5.661 \times 10^{-33}$ | Novel    | HMD803236   |
|                                                         |                                                    |                                           | rs2413667   | FAM109B     | 1.1182 (0.0608)  | $2.024 \times 10^{-75}$ | Novel    | HMD803236   |
| Lysine Metabolism                                       | N2-acetyl-N6,N6-dimethyllysine                     | 16.67/13.28                               | rs10201159  | NAT8        | 0.8927 (0.0986)  | $1.334 \times 10^{-19}$ | Novel    |             |
|                                                         |                                                    |                                           | rs11189559  | PYROXD2     | 0.4777 (0.0743)  | $1.317 \times 10^{-10}$ | Novel    |             |
|                                                         |                                                    |                                           | rs2147896   | PYROXD2     | −1.2699 (0.0649) | $3.051 \times 10^{-85}$ | Novel    |             |
|                                                         |                                                    |                                           | rs4919209   | PYROXD2     | 0.4708 (0.0724)  | $8.003 \times 10^{-11}$ | Novel    |             |
| Pregnenolone Steroids                                   | 17alpha-hydroxypregnanolone glucuronide            | 44.78/64.64                               | rs17713514  | SLC22A8     | 1.5434 (0.2094)  | $1.717 \times 10^{-13}$ | Reported | HMD80000363 |

Table 2 (continued)

| Sub pathway                    | Metabolite name                       | Missing frequency NEO/RS | SNP         | nearestGene | Beta (SE)        | P value                 | Novelty  | HMDB        |
|--------------------------------|---------------------------------------|--------------------------|-------------|-------------|------------------|-------------------------|----------|-------------|
| Primary Bile Acid Metabolism   | cholic acid glucuronide               | 39.56/3988               | rs1454247   | TMPRSS11E   | −1.038 (0.0502)  | $8.29 \times 10^{-95}$  | Novel    | HMDB0002577 |
|                                |                                       |                          | rs34976817  | TMPRSS11E   | −0.9448 (0.0802) | $4.947 \times 10^{-32}$ | Reported |             |
|                                |                                       |                          | rs62317501  | TMPRSS11E   | −0.4206 (0.0638) | $4.458 \times 10^{-11}$ | Novel    |             |
| Progestin Steroids             | glyco-beta-muricholate                | 29.12/1657               | rs45446698  | CYP3A7      | 1.8743 (0.2818)  | $2.888 \times 10^{-11}$ | Novel    | HMDB0341323 |
|                                | glycocholate glucuronide              | 79.63/6675               | rs4149056   | SLCO1B1     | 0.4621 (0.0607)  | $2.718 \times 10^{-14}$ | Novel    | HMDB0341324 |
|                                | tauro-beta-muricholate                | 70.03/5181               | rs45446698  | CYP3A7      | 2.0306 (0.1632)  | $1.526 \times 10^{-35}$ | Novel    | HMDB0000932 |
|                                | Salpha-pregnan-diol disulfate         | 56.06/4089               | rs12656482  | UGT3A1      | 1.0108 (0.0788)  | $1.052 \times 10^{-37}$ | Reported |             |
|                                |                                       |                          | rs212100    | SULT2A1     | 0.6352 (0.0697)  | $8.317 \times 10^{-20}$ | Novel    |             |
| Purine Metabolism              | pregnanolone/allopregnanolone sulfate | 53.2/47.54               | rs12656482  | UGT3A1      | 0.8004 (0.0678)  | $3.371 \times 10^{-32}$ | Novel    | HMDB0062782 |
|                                | xanthosine                            | 23.4/59.98               | rs4149056   | SLCO1B1     | 0.4155 (0.0618)  | $1.834 \times 10^{-11}$ | Novel    |             |
|                                | taurodeoxycholic acid 3-sulfate       | 30.81/2137               | rs1042391   | GMPR        | 0.3338 (0.0467)  | $8.819 \times 10^{-13}$ | Reported | HMDB0000299 |
| Secondary Bile Acid Metabolism |                                       |                          | rs4149056   | SLCO1B1     | 0.6427 (0.0784)  | $2.525 \times 10^{-16}$ | Novel    | HMDB0240734 |
| Tryptophan Metabolism          | indoleacetylglutamine                 | 25.42/1407               | rs6497506   | ACSM3       | 0.5559 (0.0713)  | $6.353 \times 10^{-15}$ | Novel    | HMDB0013240 |
|                                |                                       |                          | rs72778603  | ACSM2A      | −1.2624 (0.1369) | $2.882 \times 10^{-20}$ | Novel    |             |
|                                |                                       |                          | rs7498421   | ACSM5       | 0.6891 (0.0686)  | $9.52 \times 10^{-24}$  | Reported |             |
| Tyrosine Metabolism            | Nacetylkynurenine                     | 43.43/33.76              | rs10188058  | STAMBP      | 0.6201 (0.0793)  | $5.423 \times 10^{-15}$ | Novel    | HMDB0240342 |
|                                |                                       |                          | rs10201159  | NAT8        | 1.2782 (0.0738)  | $3.755 \times 10^{-67}$ | Reported |             |
|                                |                                       |                          | rs948445    | ACY3        | 0.6161 (0.0677)  | $9.101 \times 10^{-20}$ | Reported |             |
|                                | dopamine 4-sulfate                    | 15.32/5998               | rs17128050  | GCH1        | −0.6634 (0.0754) | $1.447 \times 10^{-18}$ | Reported | HMDB0004148 |
|                                |                                       |                          | rs67110785  | TH          | 0.4338 (0.0537)  | $6.913 \times 10^{-16}$ | Novel    |             |
|                                | X-12410                               | 15.66/1198               | rs4921913   | NAT2        | −0.6524 (0.087)  | $6.57 \times 10^{-14}$  | Novel    |             |
|                                | X-12456                               | 56.9/5198                | rs11045856  | SLCO1B1     | −0.5904 (0.0554) | $1.635 \times 10^{-26}$ | Novel    |             |
|                                |                                       |                          | rs4149056   | SLCO1B1     | 0.8525 (0.0647)  | $1.098 \times 10^{-39}$ | Reported |             |
|                                |                                       |                          | rs58712885  | SLCO1B1     | −0.6786 (0.1035) | $5.548 \times 10^{-11}$ | Novel    |             |
|                                |                                       |                          | rs974452    | SLCO1C1     | −0.5143 (0.0723) | $1.119 \times 10^{-12}$ | Novel    |             |
|                                | X-12753                               | 43.94/6999               | rs10201159  | NAT8        | 0.5979 (0.055)   | $1.729 \times 10^{-27}$ | Novel    |             |
|                                | X-13658                               | 84.34/5083               | rs61886768  | CYP2C9      | 0.4206 (0.0598)  | $2.066 \times 10^{-12}$ | Novel    |             |
|                                | X-18345                               | 59.43/3169               | rs117699706 | ASRGL1      | −0.5411 (0.0802) | $1.537 \times 10^{-11}$ | Novel    |             |
|                                | X-21312                               | 50.84/2972               | rs1165189   | SLC17A3     | 0.477 (0.0604)   | $2.751 \times 10^{-15}$ | Reported |             |

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 ( $\beta = -1.55$ ,  $p$ -value  $< 0.001$ ) (Additional file 14: Table S13). 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 ( $\beta = -0.45$ ,  $p$ -value  $< 0.001$ ). Finally, the mediation analysis (Table 4) conducted involving 9 genes whose expression levels were significantly associated with both lead SNPs and the corresponding metabolites, indicated that *HPS1* (unadjusted  $p$ -value = 0.004, FDR-adjusted  $p$ -value = 0.04) significantly partially mediated the effect between the rs2147896 SNP and the odds of missingness of the N2-acetyl, N6,N6-dimethyllysine metabolite by 2.8% ( $\beta$  (SE) =  $-0.02$  (0.007)). This was followed by *SMDT1* (unadjusted  $p$ -value = 0.06, FDR  $p$ -value = 0.25) that showed suggestive but non-significant partial mediation with the effect between rs2413667 and solanidine by 3.2% ( $\beta$  (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 (Additional file 15: Table S14). 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 the rs4149056 SNP to be associated with squamous cell carcinoma.

#### Mendelian randomization

For the eQTL based MR analysis, we identified two associations based on the IVW: *ACADM* with 3-decenoylcarnitine (IVW:  $\beta = 3.12$ , SE = 0.41, FDR  $p$ -value =  $1.6 \times 10^{-13}$ ), and *CYP2D6* with solanidine (IVW:  $\beta = -0.36$ , SE = 0.158, FDR  $p$ -value = 0.03). For the pQTL MR analysis, we identified 5 associations using IVW MR (forward). Notable associations were between *RAD51D* with 5alpha-androstan-3alpha,17alpha-diol monosulfate (IVW:  $\beta = 1.98$ , SE = 0.23, FDR  $p$ -value =  $5.51 \times 10^{-16}$ , MR-radial IVW:  $\beta = 2.11$ , SE = 0.12,  $p$ -value =  $3.88 \times 10^{-66}$ )—which had the largest effect and significance—and between *ACY3* with N-acetylkynurenine (IVW:  $\beta = -0.93$ , SE = 0.266, FDR  $p$ -value = 0.001, MR-radial IVW:  $\beta = -0.83$ , SE = 0.24,  $p$ -value = 0.005). In the reverse pQTL MR analysis, we identified 3 associations. Notable associations were 5alpha-androstan-3alpha,17alpha-diol monosulfate with *RAD51D* (IVW:  $\beta = 0.43$ , SE = 0.07, FDR  $p$ -value =  $1.49 \times 10^{-9}$ ; MR-radial IVW:  $\beta = 2.1$ , SE = 0.12,  $p$ -value =  $3.88 \times 10^{-66}$ ), the only association found in both the forward and reverse MR analyses, and X-12410 with transferrin (IVW:  $\beta = -0.11$ , SE = 0.05, FDR  $p$ -value = 0.03). All significant associations are presented in Table 5, and the full results of the MR analyses are provided in Additional file 16: Table S15 and Additional file 17: Table S16.

We further explored the results of the sensitivity analysis by performing a two-sample Mendelian randomization analyses using N2-acetyl-N6,N6-dimethyllysine as the

exposure and diabetes as the outcome, deriving genetic instruments for diabetes from a recent large GWAS on type 2 diabetes [21]. These analyses indicated that genetically predicted non-missingness of N2-acetyl-N6,N6-dimethyllysine was associated with a reduced risk of type 2 diabetes (IVW:  $\beta = -0.03$ ,  $p = 0.04$ ; Weighted median:  $\beta = -0.03$ ,  $p = 4.70 \times 10^{-5}$ ). Evidence of heterogeneity across genetic instruments was observed (Cochran's  $Q = 7.42$ ,  $Q-p = 0.02$ ); however, there was no indication of directional horizontal pleiotropy based on the MR-Egger intercept test (intercept = 0.10, SE = 0.04,  $p = 0.22$ ). The estimates remained consistent using the MR-Radial method and no instrumental outliers were detected.

## Discussion

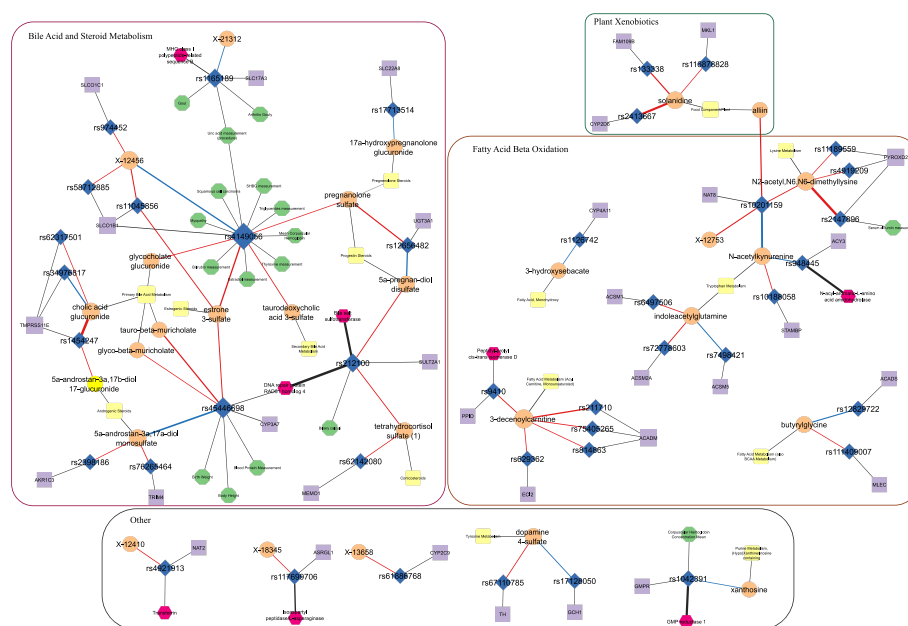
We conducted a genome-wide meta-analysis to identify genetic variants associated with the odds of missingness of metabolites and identified 55 lead SNP- 'missing metabolite' associations, of which 42 associations were novel (i.e., not found in previous GWAS of metabolite levels). Based on comprehensive in silico functional analyses, 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. 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: 3-decenoylcarnitine, N-acetylkynurenine, N2-acetyl-N6,N6-dimethyllysine, and indoleacetylglutamine. This group of metabolites were overall associated with beta-oxidation, mitochondrial function, IEMs, and potentially related to kidney function.

3-decenoylcarnitine is a medium-chain acylcarnitine that is involved in the production of energy via beta-oxidation by transporting acyl-groups into mitochondria [23]. We identified five novel associations with 3-decenoylcarnitine's odds of missingness in the intronic regions of *ACADM* and *ECI2*, in addition to an exonic loci in *PPID* (also known as *CypD* [24]). 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 [25–27]. 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 [26, 28], supporting our hypothesis of genetic variations affecting the missingness of 3-decenoylcarnitine. Further evidence to support this hypothesis was also found in our study. First, the loci was only found in our missingness GWAS and not the quantitative GWAS or any previous GWAS studies for this metabolite. Second, our functional validation analyses using gene expression data showed that *ACADM* and *ECI2* expression were associated with their respective genetic variants, as well as with those of 3-decenoylcarnitine. This was further reinforced by the eQTL MR analysis that showed a strong potentially causal association between *ACADM* and the missingness of 3-decenoylcarnitine. Finally, the pQTL analysis using PhenoScanner showed that rs9410 (*PPID* gene) had been previously associated with PPID protein

expression. The protein encoded by *PPID* (CypD) functions as a transport pore in the mitochondrial membrane [29], and plays an important role in the acetylation of proteins in the fatty acid oxidation and branched-chain amino acid metabolism, as reported in mice models [30]. Therefore, it is possible that mutations in this gene might affect the transport of 3-decenoylcarnitine in the mitochondria and lead to a reduction of its levels below the limit of detection, but further studies are required to substantiate this hypothesis. Additionally, the upstream metabolite carnitine had nearly no missingness in our data (none in NEO and only missing in 9 individuals in the Rhineland study). However, we observed that 3-decenoylcarnitine missingness occurred with higher carnitine levels, not when carnitine was low (Additional file 1: Fig. S3A and Additional file 1: Fig. S4A). This is an indication that these genetic variations could be affecting the expression or protein functions responsible for the metabolism or transport of carnitine to 3-decenoylcarnitine leading to an accumulation of carnitine and delayed production of 3-decenoylcarnitine.

N-acetylkynurenine is a metabolite belonging to the kynurenine pathway, which is crucial for the breakdown of tryptophan [31]. In our study we found that 3 loci were associated with its missingness. Among those, the locus in *ACY3* (rs948445), a gene encoding a deacetylation protein, was only associated with the missingness of N-acetylkynurenine and not the quantitative levels of the metabolite in our study (but was reported in a previous GWAS [10]). In addition, we identified a potentially causal association between this gene and the missingness of N-acetylkynurenine in the pQTL MR analysis, implying a role between the loci affecting the protein function, which in



**Fig. 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

**Table 3** Top pQTL associations with the lead SNPs

| Metabolite name                                                   | Sub pathway                                             | Lead SNP    | Nearest gene | Proxy pQTL SNPs*      | pQTL associated trait (Phenoscaner)                             | 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                            |             |              |                       |                                                                 | -                                                                 |
| 5 $\alpha$ -Androstane-3 $\alpha$ ,17 $\alpha$ -diol mono-sulfate | Androgenic Steroids                                     |             |              |                       |                                                                 | -                                                                 |
| glyco-beta-muricholate                                            | Primary Bile Acid Metabolism                            |             |              |                       |                                                                 | -                                                                 |
| X-12410                                                           | -                                                       | rs4921913   | NAT2         | rs4921915             | Transferrin                                                     | -                                                                 |
| 5 $\alpha$ pha-pregnan-diol disulfate                             | Progestin Steroids                                      | rs212100    | SULT2A1      | rs212100              | Bile salt sulfotransferase                                      | SULT2A1                                                           |
| tetrahydrocortisol sulfate                                        | Corticosteroids                                         | rs212100    | SULT2A1      | rs212100              | Bile salt sulfotransferase , DNA repair protein RAD51 homolog 4 | SULT2A1                                                           |
| 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 Table S7

| SNP        | Genes      | Positional | Functional | eQTL | pQTL |
|------------|------------|------------|------------|------|------|
| rs1126742  | CYP4A11    |            |            |      |      |
|            | CYP4A22    |            |            |      |      |
|            | CYP4B1     |            |            |      |      |
| rs814863   | ACADM      |            |            |      |      |
|            | ST6GALNAC3 |            |            |      |      |
| rs211710   | ACADM      |            |            |      |      |
|            | SLC44A5    |            |            |      |      |
| rs75405265 | ACADM      |            |            |      |      |
| rs62142080 | ST6GALNAC3 |            |            |      |      |
|            | SPAST      |            |            |      |      |
| rs10201159 | NAT8       |            |            |      |      |
|            | DGUOK      |            |            |      |      |
|            | ALMS1      |            |            |      |      |
| rs62317501 | TPRKB      |            |            |      |      |
|            | YTHDC1     |            |            |      |      |
| rs1454247  | TMPRSS11E  |            |            |      |      |
|            | UGT2B17    |            |            |      |      |
|            | 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      |            |            |      |      |
| rs2398186   | NAT2      |            |            |      |      |
|             | AKR1C4    |            |            |      |      |
| rs61886768  | AKR1C3    |            |            |      |      |
|             | AKR1C2    |            |            |      |      |
| rs11189559  | CYP2C8    |            |            |      |      |
|             | C10orf129 |            |            |      |      |
|             | CYP2C19   |            |            |      |      |
| rs2147896   | CYP2C9    |            |            |      |      |
|             | PYROXD2   |            |            |      |      |
| rs67110785  | PYROXD2   |            |            |      |      |
|             | HPS1      |            |            |      |      |
| X-18345     | HPSE2     |            |            |      |      |
|             | INS       |            |            |      |      |
| rs948445    | TH        |            |            |      |      |
|             | ASRGL1    |            |            |      |      |
| rs974452    | ALDH3B2   |            |            |      |      |
|             | ACY3      |            |            |      |      |
| rs4149056   | SLCO1B3   |            |            |      |      |
|             | PDE3A     |            |            |      |      |
| rs58712885  | SLCO1C1   |            |            |      |      |
|             | SLCO1A2   |            |            |      |      |
| rs11045856  | SLCO1B1   |            |            |      |      |
|             | SLCO1A2   |            |            |      |      |
| rs116878828 | SLCO1B1   |            |            |      |      |
|             | SLCO1A2   |            |            |      |      |
| rs2413667   | SLCO1B1   |            |            |      |      |
|             | LDHB      |            |            |      |      |

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

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

turn affects the deacetylation and the missingness of N-acetylkynurenine. Unlike it is paralog *ACY1*, *ACY3* was not reported to be associated with any IEM disorders [32]. We also found that 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 pathway) and *STAMPB* were previously reported to be associated with the IEM microcephaly-capillary malformation syndrome [31, 33]. The third locus (tagged by rs10201159) was located near *NAT8* and has been associated with the levels of acetyl forms on amino acids such as N-acetylglutamine [34]. These associations are expected as *NAT8* encodes an N-acetyltransferase [35]. In addition, *NAT8* is associated with kidney function and chronic kidney disease (CKD) progression. Specifically, genetic studies found an association between *NAT8*, higher levels of 15 acetylated amino acids (no association reported for N-acetylglutamine), and the progression of CKD [36]. Furthermore, *ACY3* [37, 38] and *STAMPB* [38, 39] have also been reported to be associated with CKD or eGFR. Further characterization of these loci in future studies may provide new insights into the role of N-acetylkynurenine and related pathways in CKD progression.

N2-acetyl,N6,N6-dimethyllysine is 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 [40, 41]. Although little is known about N2-acetyl,N6,N6-dimethyllysine in the literature,

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

| Exposure     | Mediator       | Metabolite                     | Indirect effect |         | FDR pval | Direct effect   |                        | Total effect    |                        | Percentage of mediation <sup>±</sup> |
|--------------|----------------|--------------------------------|-----------------|---------|----------|-----------------|------------------------|-----------------|------------------------|--------------------------------------|
|              |                |                                | β estimate (SE) | P value |          | β estimate (SE) | P value                | β estimate (SE) | P value                |                                      |
| rs117699706  | <i>ASRGL1</i>  | X-18345                        | −0.030 (0.023)  | 0.212   | 0.382    | −0.30 (0.067)   | $6.6 \times 10^{-06}$  | −0.333 (0.062)  | $9.6 \times 10^{-08}$  | 8.9%                                 |
| rs629362     | <i>ECI2</i>    | 3-Decenoylcarnitine            | −0.015 (0.013)  | 0.259   | 0.388    | −0.236 (0.047)  | $7.6 \times 10^{-07}$  | −0.252 (0.045)  | $2.9 \times 10^{-08}$  | 6.2%                                 |
| rs10201159   | <i>ALMS1</i>   | X-12753                        | −0.025 (0.025)  | 0.320   | 0.411    | −0.407 (0.053)  | $1.8 \times 10^{-14}$  | −0.433 (0.046)  | $< 10 \times 10^{-16}$ | 5.9%                                 |
| rs814863     | <i>ACADM</i>   | 3-Decenoylcarnitine            | −0.011 (0.007)  | 0.172   | 0.382    | −0.319 (0.063)  | $4.9 \times 10^{-07}$  | −0.330 (0.063)  | $2.0 \times 10^{-07}$  | 3.2%                                 |
| rs2413667    | <i>SMDT1</i>   | Solanidine                     | −0.024 (0.012)  | 0.055   | 0.247    | −0.722 (0.050)  | $< 10 \times 10^{-16}$ | −0.746 (0.047)  | $< 10 \times 10^{-16}$ | 3.2%                                 |
| rs10201159_T | <i>ALMS1</i>   | N2-acetyl,N6,N6-dimethyllysine | −0.015 (0.035)  | 0.684   | 0.731    | −0.495 (0.074)  | $2.9 \times 10^{-11}$  | −0.509 (0.066)  | $1.3 \times 10^{-14}$  | 2.9%                                 |
| rs2147896    | <i>HPS1</i>    | N2-acetyl,N6,N6-dimethyllysine | −0.022 (0.007)  | 0.004   | 0.036*   | −0.745 (0.046)  | $< 10 \times 10^{-16}$ | −0.766 (0.046)  | $< 10 \times 10^{-16}$ | 2.8%                                 |
| rs11189559_G | <i>PYROXD2</i> | N2-acetyl,N6,N6-dimethyllysine | −0.005 (0.004)  | 0.176   | 0.382    | −0.236 (0.049)  | $1.9 \times 10^{-06}$  | −0.241 (0.049)  | $1.1 \times 10^{-06}$  | 2.2%                                 |
| rs211710     | <i>ACADM</i>   | 3-Decenoylcarnitine            | −0.002 (0.006)  | 0.731   | 0.731    | −0.556 (0.048)  | $< 10 \times 10^{-16}$ | −0.557 (0.048)  | $< 10 \times 10^{-16}$ | 0.4%                                 |

\* Indirect effect's FDR pval = <0.05  
± (Indirect effect/total effect) \* 100

**Table 5** Inverse variance weighted Two-sample MR associations for pQTLs and eQTLs with missing metabolites

| Exposure                                         | Outcome                                          | N SNPs | Beta (CI)               | pvalue                    | FDR pval                  | F      | Q     | Q_pval                     | Beta (CI)               | Radial-MR pvalue <sup>±</sup> |
|--------------------------------------------------|--------------------------------------------------|--------|-------------------------|---------------------------|---------------------------|--------|-------|----------------------------|-------------------------|-------------------------------|
| ACY3                                             | N-acetylkynurenine                               | 20     | −0.93 (−1.45, −0.41)    | 0.0004                    | 0.00102                   | 118.49 | 65.6  | 4.807 × 10 <sup>−7</sup> * | −0.745 (−1.064, −0.426) | 0.005                         |
| GMPIR1                                           | xanthosine                                       | 5      | 1.15 (0.59, 1.71)       | 5.36 × 10 <sup>−5</sup>   | 0.000196                  | 209.9  | 10.3  | 0.035*                     | -                       | -                             |
| RAD51D                                           | 5alpha-androstan-3alpha,17alpha-diol monosulfate | 3      | 1.99 (1.53, 2.45)       | 5.011 × 10 <sup>−17</sup> | 5.512 × 10 <sup>−16</sup> | 144.81 | 8.4   | 0.0146*                    | 2.106 (1.865, 2.347)    | 3.88 × 10 <sup>−66</sup>      |
| SULT2A1                                          | 5alpha-pregnan-diol disulfate                    | 10     | 1.16 (0.54, 1.78)       | 0.00025                   | 0.00069                   | 162.17 | 33.5  | 0.0001*                    | 1.245 (0.753, 1.737)    | 0.007                         |
| SULT2A1                                          | tauro-beta-tetrahydrocortisol sulfate            | 10     | 1.76 (1.12, 2.4)        | 9.3763 × 10 <sup>−8</sup> | 5.156 × 10 <sup>−7</sup>  | 162.17 | 22.3  | 0.00793*                   | 1.681 (1.109, 2.253)    | 0.0009                        |
| X-12410                                          | Transferrin                                      | 3      | −0.11 (−0.206, −0.014)  | 0.017                     | 0.03                      | 39.14  | 0.98  | 0.613                      | -                       | -                             |
| 5alpha-androstan-3alpha,17alpha-diol monosulfate | RA51D                                            | 3      | 0.43 (0.299, 0.561)     | 1.66 × 10 <sup>−10</sup>  | 1.49 × 10 <sup>−9</sup>   | 115.45 | 18.22 | 0.0001*                    | 0.463 (−0.016, 0.94)    | 4.06 × 10 <sup>−86</sup>      |
| tauro-beta-tetrahydrocortisol sulfate            | RA51D                                            | 2      | 0.211 (0.107, 0.315)    | 8.4 × 10 <sup>−105</sup>  | 0.0002                    | 62.67  | 2.6   | 0.1                        | -                       | -                             |
| ACADM                                            | 3-decenoylcarnitine                              | 2      | 3.12 (2.309, 3.931)     | 5.3 × 10 <sup>−14</sup>   | 1.6 × 10 <sup>−13</sup>   | 41.2   | 3.4   | 0.06                       | -                       | -                             |
| CYP2D6                                           | solanidine                                       | 3      | −0.361 (−0.671, −0.051) | 0.022                     | 0.03                      | 95.03  | 4.4   | 0.11                       | -                       | -                             |

FDR: False discovery rate  
Q: Cochran's Q statistic  
Q\_pval: Corresponding p-value assessing heterogeneity  
\* Significant heterogeneity (P value < 0.05)  
± Radial MR was only performed if outlier variants were present

its precursor—namely TML—have been studied extensively [40–42]. TML has been associated with cardiovascular diseases and reportedly predicted all-cause mortality and cardiovascular disease [41]. 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, 43, 44]. *PYROXD2* is localized in the mitochondrial inner membrane and has an important role in regulating mitochondrial function [45]. In addition, *PYROXD2* is associated with the IEM disorder trimethylaminuria [46]. *PYROXD2* is normally associated with low levels of trimethylamine in the urine of healthy individuals [46]. Our PheWAS analysis indicated the lead SNP rs2147896 of N2-acetyl,N6,N6-dimethyllysine mapped to *PYROXD2* gene is associated with serum albumin measurement, which is a key indicator of kidney function. 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 [46, 47]. 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. In addition, our two-sample Mendelian randomization analysis using genetically predicted non-missingness of N2-acetyl-N6,N6-dimethyllysine (*ALMS1* and *PYROXD2*) as the exposure and T2D as the outcome indicated that the non-missingness of N2-acetyl-N6,N6-dimethyllysine is associated with a reduced risk of T2D (or, equivalently, that N2-acetyl-N6,N6-dimethyllysine missingness is associated with a higher risk of T2D). This could indicate that either (part) of the association of the respective genetic loci with the odds of missingness of this particular metabolite may be due to their association with diabetes, or vice versa. Interestingly, in a previous study, higher levels of N6, N6-dimethyllysine were found to be potentially causally associated with a lower risk of diabetic retinopathy [20], providing indirect support for the latter possibility. While this finding should be interpreted with caution, it suggests a potential link between this metabolite and diabetes-related metabolic pathways. Consequently, these associations may reflect poor metabolism of TML and trimethylamine, which in turn might be related to the development of trimethylaminuria. Furthermore, from our GWAS, eQTL, Mendelian randomization analyses, and the evidence from the literature, the missingness of N2-acetyl,N6,N6-dimethyllysine and the loci associated with its missingness could be associated with other comorbidities such as cardiovascular disease, CKD, and particularly T2D.

Finally, indoleacetylglutamine, a gut microbiome-derived metabolite [48] involved in tryptophan metabolism and has been reported in various studies regarding gut microbiome and chronic kidney disease [49, 50]. In our study, we identified two novel loci near the *ACSM1* and *ACSM2A* genes that were associated with the odds of missingness of this metabolite. Similar to *ACADM*, both *ACSM1* and *ACSM2A* are involved in the metabolism of acyl-CoA in medium-chain fatty acids. In addition, significant associations have been reported between genetic variants in these genes and eGFR and serum creatinine levels [51, 52], indicating a role in kidney function. This is further supported by their association with indoleacetylglutamine, which is part of the tryptophan

metabolism pathway (shown in Fig. 3) that also involves N-acetylkynurenine and *NAT8*, which in turn are related to CKD.

The second group of metabolites we report was composed of steroids and bile acid metabolites with a shared association with two pleiotropic SNPs: rs4149056 and rs45446698. The rs4149056 (*SLCO1B1*) SNP 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 [53]. Second, rs4149056 and *SLCO1B1* were reported to be associated with serum estrone levels [54, 55]. Similarly, rs45446698 (*CYP3A7*) has been reported in studies relating breast cancer to estrone and progesterone levels [56–58]. In addition, PheWAS analysis indicated previous associations between rs45446698 and birth weight, a trait with lifetime implications for metabolism [59]. 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 (RAD51D) in a trans manner [60, 61]. This was further supported by our forward and reverse pQTL MR analysis. These modifications could indicate consequences of the 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.

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 [62, 63]. Interestingly, the *NAT8* locus was found to be associated with the odds of missingness of alliin but not with the quantitative levels of this metabolite. In addition, a previous study reported a different SNP in *NAT8* that was associated with the acylated form of alliin—N-acetylalliin [10]. This novel finding indicates that the locus in *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, a slightly toxic metabolite in low quantities derived from potatoes and other plants of the Solanaceae family [64, 65]. We identified three novel SNP 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 [66, 67]. Solanidine has also been reported as a potential dietary marker to assess the efficiency of *CYP2D6* functionality [68]. This was further supported by a clinical trial reporting *CYP2D6* inhibition to be associated with up to a 4.56-fold increase in solanidine levels, indicating compromised xenobiotic metabolism, and additional studies used solanidine as a biomarker to identify “poor metabolizers”

[66, 68, 69]. Based on our findings regarding the association of rs2413667 and the odds of missingness of solanidine—from the GWAS and eQTL MR analyses—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 [70] and could be utilized in identifying and developing personalized nutritional interventions [71]. This may also be true for all the reported SNPs and metabolites we have identified, but would require further studies. The rs2413667 SNP is also in proximity to the eQTL region of *SMDT1*, and based on our mediation analysis, this eQTL has a suggestive 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 [65]. 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 were 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—the missingness of metabolites. Missingness of a metabolite measurement can be caused by either a technical issue, i.e., failure of metabolite identification in the spectral data due to a deconvolution issue, 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 the 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 [72]. 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 [73]. 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 deficient (Additional file 1: Fig. S3B and Additional file 1: Fig. S4B). The 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 the 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, support the hypothesis that genetic factors influence the probability of metabolites' absence. Future studies are needed for a deeper investigation of the underlying biological pathway of the missingness of the reported metabolites. A limitation of our study is the relatively small sample sizes used in the GWAS. However, despite this limitation we were able to identify strong associations in both studies. Furthermore, by using two studies and a meta-analysis approach, we replicated the associations between the genetic variants and missingness. We expect that with a larger sample size even more associations will be identified. 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 [74]. 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.

## Conclusions

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 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.

## 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 complete 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 [75]. Briefly, men and women aged between 45 and 65 years with a self-reported body mass index (BMI) of  $27 \text{ kg/m}^2$  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 criterion being insufficient command of the German language to provide informed consent. Each participant undergoes a comprehensive 7-h examination protocol, which includes detailed questionnaires, brain imaging, neuropsychological examinations, and blood sample analyses. Prevalent diabetes was defined based on the current use of antidiabetic medication (ATC code A10), glycated hemoglobin ( $\text{HbA1c} \geq 6.5\%$ ), fasting plasma glucose ( $\geq 126 \text{ mg/dL}$ ), or self-reported physician-diagnosed diabetes obtained during the medical interview.  $\text{HbA1c} < 6.5\%$  and fasting plasma glucose  $< 126 \text{ mg/dL}$  were considered indicative of no diabetes. Data collection in the Rhineland Study was approved by relevant ethics committees and/or institutions and was conducted according to the Declaration of Helsinki. All participants provided written informed consent.

### 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. Only very few samples were collected in a non-fasting state ( $N=51$ ). Details on the metabolomics pipeline have been described elsewhere [11]. In brief, the Metabolon HD4 platform employs an untargeted measurement approach that utilizes reverse phase 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 hydrophilic interaction liquid chromatography (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, carbohydrates, 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 the 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, the majority of xenobiotic metabolites are missing in at least 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 ( $< 10\%$ ), that are truly missing ( $> 90\%$ ), or other unknown causes. Accordingly, we selected 341 in the NEO study and 425 metabolites in the Rhineland Study.

### 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 G  notypage (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 [76]. Genotypes were imputed to the Haplotype Reference Consortium (HRC) release 1.1 [77]. 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 ( $p < 1 \times 10^{-6}$ ). 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, we used EIGENSTRAT (version 16,000), a principal component-based analysis to assess population structure [78]. In addition, prior to the GWAS, individuals of non-European ancestry were identified based on the clustering of their genetic principal components relative to the main European ancestry cluster and were excluded from further analyses. To ensure a genetically homogeneous study population of European ancestry, 164 individuals who fell outside the PCA-defined European ancestry cluster were removed. Finally, imputation was performed with IMPUTE (version 2) [79], using the 1000 Genomes version 3 phase 5 as the reference panel [80]. We sought a high imputation quality by only including SNPs with an R-squared ( $R^2$ ) info metric  $> 0.3$ , which indicates reliable imputation [81]. After quality control, all autosomal SNPs passing these criteria—representing over 95% of common variants ( $MAF > 0.01$ )—were retained for the subsequent analyses.

### 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) [82], fitting a first 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. Genetic principal components were included to control for residual confounding due to population structure, as they capture ancestry-related variation through linear combinations of genetic markers and are therefore standard in GWAS. Because analyses were restricted to individuals of European ancestry and additionally adjusted for genetic principal components, residual confounding due to population stratification is expected to be minimal. Genome-wide significance level was set at  $p < 5 \times 10^{-8}$  [83]. However, because of the large number of outcome variables (i.e., missing metabolites), we corrected for multiple testing using the method of Li & Ji [84], which estimates the effective number of independent tests. Accordingly, we estimated the effective number of independent tests to be 315 and 313 in the NEO and the Rhineland Study, respectively, resulting in a metabolome-wide significance level of  $p < 1.58 \times 10^{-10}$  ( $\approx 5 \times 10^{-8}/315$ ) and  $p < 1.59 \times 10^{-10}$  ( $\approx 5 \times 10^{-8}/313$ ), respectively. As a sensitivity analysis, we additionally adjusted for prevalent diabetes ( $N=205$ ) in the Rhineland Study to assess the robustness of associations after accounting for diabetes status.

#### Quantitative GWAS sensitivity analysis

To assess the uniqueness of the associations identified through GWAS on missing metabolites, we additionally performed a GWAS for the quantitative measures of the metabolites in the Rhineland study that were significant in the primary binary GWAS (26 metabolites). We used the same  $p < 1.59 \times 10^{-10}$  cutoff as the binary GWAS as well. Missing values in these metabolites were left as is. We used REGENIE software (v2.2) [82], and used linear regression for association testing. Prior to association testing, metabolite levels were adjusted for age, sex, fasting status, batch effects, and the first ten genetic principal components using linear regression. The residuals from these models were then transformed using a rank-based inverse normal transformation to approximate a normal distribution. These transformed residuals were used as input phenotypes in the GWAS performed with REGENIE (v2.2).

#### Assessment of potential study-specific effects

To assess potential study-specific effects, we performed SNP  $\times$  study interaction analyses using the GWAS summary statistics for each study for each metabolite (224 metabolites). Heterogeneity in SNP effect estimates between the two studies was tested using EasyStrata [85]. We considered an interaction effect to be metabolome-wide significant if it attained a  $1.59 \times 10^{-10}$ .

### Meta-analysis of GWAS

For the GWAS meta-analysis, we selected and used 7,310,783 SNPs that 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 SNPs in the NEO study that had a MAF > 0.01 (6,610,552/7,310,783 SNPs). Meta-analysis was performed employing an inverse variance-weighted fixed-effects model using METAL [86]. 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  $p < 1.59 \times 10^{-10}$ , which was the more stringent cut-off used in the Rhineland Study. To verify if our findings for missing metabolites were indeed novel, we used curated metabolome GWAS databases such as mGWAS-Explorer [87] and Phe-noScanner [88]. We further manually verified whether our lead SNPs were previously identified in metabolomic-based GWAS studies [10] as metabolome quantitative trait locus (mQTLs).

### 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. This dataset was LD-clumped ( $r^2 < 0.6$ ) using the Functional Mapping and Annotation (FUMA) platform [89] 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).

### 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 [90]. 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) [91] 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 their function as either missense, start loss, or stop gain. ProGEM also assesses whether the lead SNPs are eQTL using the GTEX v8 database [92]. The genes that were identified through the 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.

#### **RNA sequencing data in the Rhineland study**

Total RNA was isolated from 3,384 whole blood samples in the Rhineland Study. Those samples were 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  $2 \times 50$  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 v0.39. 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 –quant-Mode 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-step approach analysis leveraging data from the first 3,384 consecutive participants of the Rhineland Study, on whom genotype, gene expression, and metabolite data were available and passed quality control ( $N=2,575$ ). First, we assessed the associations between the lead SNPs and the corresponding genes, selected through bottom-up and top-down approaches. Each lead SNP was analyzed separately in relation to its corresponding candidate gene as the outcome. 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 the metabolomics' batch effect using logistic regression analysis. Third, we employed a mediation analysis using the R package lavaan v0.6–11 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 [88]. 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 ( $p < 5 \times 10^{-8}$ ) with our lead SNP or a proxy SNP ( $r^2 > 0.9$ ) with our lead SNPs.

### 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  $-\log_{10}$  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 [93].

### Mendelian randomization of eQTLs and pQTLs

We evaluated the potential causal effects for the genes we identified on the missingness of the metabolites reported from the GWAS analyses using Mendelian randomization analysis (MR). For each gene, we used SNPs for their eQTL and pQTL data. We first used whole blood eQTLs from GTEX v8 as genetic instruments, treating gene expression as the exposure and missing metabolite phenotypes as the outcomes. Instruments were selected using LD clumping with a 10 Mb window, an LD threshold of  $r^2 < 0.01$ , and a significance threshold of  $P < 1 \times 10^{-4}$ . We used the Two Sample Mendelian Randomization (MR) R package (v 0.6.4) [94]. Second, we conducted two-sample MR for the pQTL data per gene, which were used as instruments to assess the causal effect of protein levels on metabolite missingness. We applied the same clumping parameters with a stricter significant threshold of  $P < 5 \times 10^{-8}$ . Additionally, as a sensitivity analysis, we performed reverse MR analyses to evaluate whether the missingness of specific metabolites could causally influence protein expression. For all MR analyses, we performed MR Egger, simple mode, weighted median, weighted mode, Wald ratio, and inverse variance weighted (IVW) analyses—focusing on the IVW for interpretation. We also assessed the heterogeneity using Cochran’s Q and horizontal pleiotropy using the MR-egger intercept. Bias due to variant outliers was accounted for using Radial MR. Briefly, this method applies a simulation-based approach to detect outlying variants and remove them to re-estimate the associations. This analysis was performed using the Radial MR R package (v 1.1) [95].

## Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13059-026-04132-9>.

Additional file 1: Figs. S1–S4. Supplementary figures showing quantitative GWAS sensitivity analyses, correlation of SNP effect estimates between NEO and Rhineland study cohorts, and visualization of missingness of 3-decenoyl-carnitine and dopamine 4-sulphate in relation to carnitine and dopamine 3-sulphate in NEO and Rhineland study cohorts

Additional file 2: Table S1. Genomic risk loci of metabolome wide significant SNPs in the NEO study

Additional file 3: Table S2. Genomic risk loci of metabolome-wide significant SNPs in the Rhineland Study

Additional file 4: Table S3. SNP\*SNP interaction analyses using EasyStrata

Additional file 5: Table S4. Frequency of missing metabolites in the NEO study and Rhineland Study

Additional file 6: Table S5. Overall SNPs in linkage disequilibrium with lead SNPs identified in meta-analysis

Additional file 7: Table S6. Comparison of genomic risk loci identified in the GWAS-meta analysis with cohort specific GWAS results from the NEO and Rhineland Study cohorts

Additional file 8: Table S7. Overview of the lead SNPs in GWAS-meta analysis

Additional file 9: Table S8. Association of lead SNPs with inter and intra class metabolites

Additional file 10: Table S9. Lead SNPs look ups with protein quantitative trait loci using PhenoScanner V2

Additional file 11: Table S10. Lead SNPs look ups with metabolome quantitative trait loci using PhenoScanner V2, mGWAS and Praveen et.al database

Additional file 12: Table S11. ProGeM analysis of identifying positional genes using bottom-up approach for the lead SNPs

Additional file 13: Table S12. ProGeM analysis of identifying candidate genes using top-down approach for the lead SNPs

Additional file 14: Table S13. Functional validation analyses using gene expression data

Additional file 15: Table S14. DisGeNet phenome-wide association study analyses of lead SNPs

Additional file 16: Table S15. Two sample Mendelian Randomization of gene expression and missing metabolites

Additional file 17: Table S16. Two sample MR of pQTLs and missing metabolites

## 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.

## Peer review information

Fengchao Yu and Tim Sands were the primary editors of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. The peer-review history is available in the online version of this article.

## Authors' contributions

Tariq O. Faquih, Mohammed Aslam Imtiaz, Elvire N. Landstra, Astrid van Hylckama Vlieg, Frits R. Rosendaal, Dennis O. Mook-Kanamori, Monique M. B. Breteler, N. Ahmad Aziz, Ko Willems van Dijk, Conceptualization. Tariq O. Faquih, Mohammed Aslam Imtiaz, Valentina Talevi, Elvire N. Landstra, Dennis O. Mook-Kanamori, Monique M. B. Breteler, N. Ahmad Aziz, Ko Willems van Dijk Methodology, Tariq O. Faquih, Mohammed Aslam Imtiaz, Valentina Talevi, Formal Analysis. Tariq O. Faquih, Mohammed Aslam Imtiaz, Writing – Original Draft Preparation, Tariq O. Faquih, Mohammed Aslam Imtiaz, Visualization. Valentina Talevi, Elvire N. Landstra, Astrid van Hylckama Vlieg, Ruifang Li-Gao, Raymond Noordam, Diana van Heemst, Dennis O. Mook-Kanamori, Monique M. B. Breteler, N. Ahmad Aziz, Ko Willems van Dijk. Writing – Reviewing and Editing. Astrid van Hylckama Vlieg, Dennis O. Mook-Kanamori, Monique M. B. Breteler, N. Ahmad Aziz, Ko Willems van Dijk. Supervision. Frits R. Rosendaal, Raymond Noordam, Diana van Heemst, Dennis O. Mook-Kanamori, Monique M. B. Breteler, N. Ahmad Aziz, Ko Willems van Dijk. Funding acquisition. Dennis O. Mook-Kanamori, Monique M. B. Breteler. Resources. Monique M. B. Breteler, Data Curation. All authors read and approved the final manuscript.

## Funding

Open Access funding enabled and organized by Projekt DEAL. 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).

### 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/> and <https://www.lumc.nl/en/afdelingen/Clinical-Epidemiology/obesity-related-diseases/>. Similarly, the Rhineland Study <https://www.rheinland-studie.de/en/> 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. The summary statistics of the meta-analysis GWAS are publicly available via the GWAS catalog under accession number GCP001677.

### Declarations

#### Ethics approval and consent to participate

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

All study participants in the respective cohorts provided their written informed consent. No additional ethical approval was required for the present meta-analysis, as all participating studies had previously obtained local ethical approval and written informed consent from participants.

#### Consent for publication

All authors read and approved the final manuscript.

#### Competing interests

The authors declare no competing interests.

#### Author details

<sup>1</sup>Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, The Netherlands. <sup>2</sup>Division of Sleep and Circadian Disorders, Brigham and Women's Hospital, Boston, USA. <sup>3</sup>King Faisal Specialist Hospital & Research Center, Riyadh, Kingdom of Saudi Arabia. <sup>4</sup>German Centre for Neurodegenerative Diseases (DZNE), Population Health Sciences, Bonn, Germany. <sup>5</sup>Department of Internal Medicine, Section of Gerontology and Geriatrics, Leiden University Medical Center, Leiden, The Netherlands. <sup>6</sup>Department of Public Health and Primary Care, Leiden University Medical Center, Leiden, The Netherlands. <sup>7</sup>University of Bonn, University Hospital Bonn, Institute for Medical Biometry, Informatics and Epidemiology (IMBIE), Faculty of Medicine, Bonn, Germany. <sup>8</sup>Center for Neurology, University of Bonn, University Hospital Bonn, Bonn, Germany. <sup>9</sup>Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands. <sup>10</sup>Department of Internal Medicine, Division of Endocrinology, Leiden University Medical Center, Leiden, The Netherlands. <sup>11</sup>Eindhoven Laboratory for Experimental Vascular Medicine, Leiden University Medical Center, Leiden, The Netherlands.

Received: 4 October 2025 Accepted: 27 May 2026

Published online: 10 July 2026

### References

- Clish CB. Metabolomics: an emerging but powerful tool for precision medicine. *Cold Spring Harb Mol Case Stud.* 2015;1(1):a000588. <https://doi.org/10.1101/mcs.a000588>.
- Tzoulaki I, Ebbels TM, Valdes A, et al. Design and analysis of metabolomics studies in epidemiologic research: a primer on -omic technologies. *Am J Epidemiol.* 2014;180(2):129–39. <https://doi.org/10.1093/aje/kwu143>.
- Shah SH, Kraus WE, Newgard CB. Metabolomic profiling for the identification of novel biomarkers and mechanisms related to common cardiovascular diseases: form and function. *Circulation.* 2012;126(9):1110–20. <https://doi.org/10.1161/circulationaha.111.060368>.
- Zheng Y, Yu B, Alexander D, et al. Metabolomics and incident hypertension among Blacks: the Atherosclerosis Risk in Communities Study. *Hypertension.* 2013;62(2):398–403. <https://doi.org/10.1161/HYPERTENSIONAHA.113.01166>.
- Ahola-Olli AV, Mustelin L, Kalimeri M, et al. Circulating metabolites and the risk of type 2 diabetes: a prospective study of 11,896 young adults from four Finnish cohorts. *Diabetologia.* 2019;62(12):2298–309. <https://doi.org/10.1007/s00125-019-05001-w>.
- Menni C, Fauman E, Erte I, et al. Biomarkers for type 2 diabetes and impaired fasting glucose using a nontargeted metabolomics approach. *Diabetes.* 2013;62(12):4270–6. <https://doi.org/10.2337/db13-0570>.
- Faquih TO, Aziz NA, Gardiner SL, 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. *Hum Mol Genet.* 2023;32(10):1741–52. <https://doi.org/10.1093/hmg/ddad020>.
- Hatano T, Saiki S, Okuzumi A, et al. Identification of novel biomarkers for Parkinson's disease by metabolomic technologies. *J Neurol Neurosurg Psychiatry.* 2016;87(3):295–301. <https://doi.org/10.1136/jnnp-2014-309676>.
- Shin SY, Fauman EB, Petersen AK, et al. An atlas of genetic influences on human blood metabolites. *Nat Genet.* 2014;46(6):543–50. <https://doi.org/10.1038/ng.2982>.
- Surendran P, Stewart ID, Au Yeung VPW, et al. Rare and common genetic determinants of metabolic individuality and their effects on human health. *Nat Med.* 2022;28(11):2321–32. <https://doi.org/10.1038/s41591-022-02046-0>.

11. Faquih TO, van Klinken JB, Li-Gao R, et al. Hepatic triglyceride content is intricately associated with numerous metabolites and biochemical pathways. *Liver Int.* 2023;43(7):1458–72. <https://doi.org/10.1111/liv.15575>.
12. Faquih T, van Smeden M, Luo J, et al. A workflow for missing values imputation of untargeted metabolomics data. *Metabolites.* 2020(12). <https://doi.org/10.3390/metabo10120486>.
13. Do KT, Wahl S, Raffler J, et al. Characterization of missing values in untargeted MS-based metabolomics data and evaluation of missing data handling strategies. *Metabolomics.* 2018;14(10):128. <https://doi.org/10.1007/s11306-018-1420-2>.
14. Yousri NA, Fakhro KA, Robay A, et al. Whole-exome sequencing identifies common and rare variant metabolic QTLs in a Middle Eastern population. *Nat Commun.* 2018;9(1):333. <https://doi.org/10.1038/s41467-017-01972-9>.
15. Tin A, Marten J, Halperin Kuhns VL, et al. Target genes, variants, tissues and transcriptional pathways influencing human serum urate levels. *Nat Genet.* 2019;51(10):1459–74. <https://doi.org/10.1038/s41588-019-0504-x>.
16. Hsia DY-y. Inborn errors of metabolism / [by] David Yi-yung Hsia. Chicago: Chicago : Year Book Publishers 1959.
17. Dharuri H, Demirkan A, van Klinken JB, et al. Genetics of the human metabolome, what is next? *Biochim Biophys Acta.* 2014;1842(10):1923–31. <https://doi.org/10.1016/j.bbadis.2014.05.030>.
18. Jeanmonod R, Asuka E, Jeanmonod D. Inborn Errors of Metabolism. StatPearls. 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.: StatPearls Publishing Copyright © 2024, StatPearls Publishing LLC. 2024.
19. Balasubramaniam S, Duley JA, Christodoulou J. Inborn errors of purine metabolism: clinical update and therapies. *J Inher Metab Dis.* 2014;37(5):669–86. <https://doi.org/10.1007/s10545-014-9731-6>.
20. Li B, Zhao X, Xie W, et al. Causal association of circulating metabolites with diabetic retinopathy: a bidirectional Mendelian randomization analysis. *Front Endocrinol.* 2024. <https://doi.org/10.3389/fendo.2024.1359502>.
21. Verma A, Huffman JE, Rodriguez A, et al. Diversity and scale: genetic architecture of 2068 traits in the VA Million Veteran Program. *Science.* 2024;385(6706):eadj1182. <https://doi.org/10.1126/science.adj1182>.
22. Taylor C, Crosby I, Yip V, et al. A review of the important role of CYP2D6 in pharmacogenomics. *Genes.* 2020;11(11):1295.
23. Dambrova M, Makrecka-Kuka M, Kuka J, et al. Acylcarnitines: nomenclature, biomarkers, therapeutic potential, drug targets, and clinical trials. *Pharmacol Rev.* 2022;74(3):506–51. <https://doi.org/10.1124/pharmrev.121.000408>.
24. Kieffer LJ, Seng TW, Li W, 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.* 1993;268(17):12303–10.
25. Tavecchio M, Lisanti S, Bennett MJ, et al. Deletion of Cyclophilin D impairs  $\beta$ -oxidation and promotes glucose metabolism. *Sci Rep.* 2015;5(1):15981. <https://doi.org/10.1038/srep15981>.
26. Vishwanath VA. Fatty acid beta-oxidation disorders: a brief review. *Ann Neurosci.* 2016;23(1):51–5. <https://doi.org/10.1159/000443556>.
27. Houten SM, Violante S, Ventura FV, et al. The biochemistry and physiology of mitochondrial fatty acid  $\beta$ -oxidation and its genetic disorders. *Annu Rev Physiol.* 2016;78(1):23–44. <https://doi.org/10.1146/annurev-physiol-021115-105045>.
28. Merritt JL, 2nd, Chang JJ. Medium-Chain Acyl-Coenzyme A Dehydrogenase Deficiency. In: Adam MP, Feldman J, Mirzaa GM, et al., eds. *GeneReviews*®. Seattle (WA): 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. 1993.
29. Elrod JW, Molkentin JD. Physiologic functions of cyclophilin D and the mitochondrial permeability transition pore. *Circ J.* 2013;77(5):1111–22. <https://doi.org/10.1253/circj.cj-13-0321>. ([published Online First: 20130329]).
30. Nguyen TTM, Wong R, Menazza S, et al. Cyclophilin D modulates mitochondrial acetylome. *Circ Res.* 2013;113(12):1308–19. <https://doi.org/10.1161/CIRCRESAHA.113.301867>.
31. Shi H, Enriquez A, Rapadas M, et al. NAD deficiency, congenital malformations, and niacin supplementation. *N Engl J Med.* 2017;377(6):544–52. <https://doi.org/10.1056/NEJMoa1616361>.
32. Bortoluzzi VT, Ribeiro RT, Zemniçak AB, et al. Disturbance of mitochondrial functions caused by N-acetylglutamate and N-acetylmethionine in brain of adolescent rats: potential relevance in aminoacylase 1 deficiency. *Neurochem Int.* 2023;171:105631. <https://doi.org/10.1016/j.neuint.2023.105631>.
33. Carter MT, Mirzaa G, McDonell LM, et al. Microcephaly-Capillary Malformation Syndrome. In: Adam MP, Feldman J, Mirzaa GM, et al., eds. *GeneReviews*®. Seattle (WA): 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. 1993.
34. Tahir UA, Katz DH, Avila-Pachecho J, et al. Whole Genome Association Study of the Plasma Metabolome Identifies Metabolites Linked to Cardiometabolic Disease in Black Individuals. *Nat Commun.* 2022;13(1):4923. <https://doi.org/10.1038/s41467-022-32275-3>[publishedOnlineFirst:20220822].
35. Raffler J, Friedrich N, Arnold M, et al. Genome-wide association study with targeted and non-targeted NMR metabolomics identifies 15 novel loci of urinary human metabolic individuality. *PLoS Genet.* 2015;11(9): e1005487. <https://doi.org/10.1371/journal.pgen.1005487>. [published Online First: 20150909].
36. Luo S, Surapaneni A, Zheng Z, et al. NAT8 variants, N-acetylated amino acids, and progression of CKD. *Clin J Am Soc Nephrol.* 2020;16(1):37–47. <https://doi.org/10.2215/CJN.08600520>.
37. Luo S, Feofanova EV, Tin A, et al. Genome-wide association study of serum metabolites in the African American Study of Kidney Disease and Hypertension. *Kidney Int.* 2021;100(2):430–9. <https://doi.org/10.1016/j.kint.2021.03.026>.
38. Zhang R, Sun Y, Huang Z, et al. Abstract 045: Genome-wide association study for metabolites in the Tryptophan Metabolism Pathway among African American patients with Chronic Kidney Disease. *Circ.* 2025. [https://doi.org/10.1161/cir.151.suppl\\_1.045](https://doi.org/10.1161/cir.151.suppl_1.045).
39. Nano J, Schöttker B, Lin J-s, et al. Novel biomarkers of inflammation, kidney function and chronic kidney disease in the general population. *Nephrol Dial Transplant.* 2021;37(10):1916–26. <https://doi.org/10.1093/ndt/gfab294>.
40. Koeth RA, Levison BS, Culley MK, et al.  $\gamma$ -Butyrobetaine is a proatherogenic intermediate in gut microbial metabolism of L-carnitine to TMAO. *Cell Metab.* 2014;20(5):799–812. <https://doi.org/10.1016/j.cmet.2014.10.006>.

41. Bjørnstad E, Dhar I, Svingen GFT, et al. Trimethyllysine predicts all-cause and cardiovascular mortality in community-dwelling adults and patients with coronary heart disease. *Eur Heart J Open*. 2021;1(1):oeab007. <https://doi.org/10.1093/ehjopen/oeab007>.
42. Maas MN, Hintzen JCJ, Porzberg MRB, et al. Trimethyllysine: from Carnitine biosynthesis to epigenetics. *Int J Mol Sci*. 2020. <https://doi.org/10.3390/ijms21249451>.
43. Nicholson G, Rantalainen M, Li JV, et al. A genome-wide metabolic QTL analysis in Europeans implicates two loci shaped by recent positive selection. *PLoS Genet*. 2011;7(9):e1002270. <https://doi.org/10.1371/journal.pgen.1002270>.
44. Montoliu I, Genick U, Ledda M, et al. Current status on genome-metabolome-wide associations: an opportunity in nutrition research. *Genes Nutr*. 2013;8(1):19–27. <https://doi.org/10.1007/s12263-012-0313-7>.
45. Wang T, Xie X, Liu H, et al. Pyridine nucleotide-disulphide oxidoreductase domain 2 (PYROXD2): role in mitochondrial function. *Mitochondrion*. 2019;47:114–24. <https://doi.org/10.1016/j.mito.2019.05.007>.
46. Guo Y, Hwang LD, Li J, et al. Genetic analysis of impaired trimethylamine metabolism using whole exome sequencing. *BMC Med Genet*. 2017;18(1):11. <https://doi.org/10.1186/s12881-017-0369-8>.
47. Cheng Y, Schlosser P, Hertel J, et al. Rare genetic variants affecting urine metabolite levels link population variation to inborn errors of metabolism. *Nat Commun*. 2021;12(1):964. <https://doi.org/10.1038/s41467-020-20877-8>.
48. Lee AM, Xu Y, Hooper SR, et al. Circulating metabolomic associations with neurocognitive outcomes in pediatric CKD. *Clin J Am Soc Nephrol*. 2024. <https://doi.org/10.2215/CJN.0000000000000318>.
49. Hu JR, Myint L, Levey AS, et al. A metabolomics approach identified toxins associated with uremic symptoms in advanced chronic kidney disease. *Kidney Int*. 2022;101(2):369–78. <https://doi.org/10.1016/j.kint.2021.10.035>.
50. Ludwig GD, Epstein IS. A genetic study of two families having the acute intermittent type of porphyria. *Ann Intern Med*. 1961;55(1):81–93. <https://doi.org/10.7326/0003-4819-55-1-81>.
51. Common Metabolic Diseases Knowledge Portal (cmdkp.org). ACSM1 Gene page. [Available from: <https://hugeamp.org/gene.html?gene=ACSM12025> Dec 17.
52. Common Metabolic Diseases Knowledge Portal (cmdkp.org). ACSM2A Gene page. [Available from: <https://hugeamp.org/gene.html?gene=ACSM2A2025> Dec 17.
53. Turongkaravee S, Jittikoon J, Lukkunaprasit T, et al. A systematic review and meta-analysis of genotype-based and individualized data analysis of SLC01B1 gene and statin-induced myopathy. *Pharmacogenomics J*. 2021;21(3):296–307. <https://doi.org/10.1038/s41397-021-00208-w>.
54. Moyer AM, de Andrade M, Faubion SS, et al. SLC01B1 genetic variation and hormone therapy in menopausal women. *Menopause*. 2018;25(8):877–82. <https://doi.org/10.1097/GME.0000000000001109>.
55. Dudenkov TM, Ingle JN, Buzdar AU, et al. SLC01B1 polymorphisms and plasma estrone conjugates in postmenopausal women with ER+ breast cancer: genome-wide association studies of the estrone pathway. *Breast Cancer Res Treat*. 2017;164(1):189–99. <https://doi.org/10.1007/s10549-017-4243-3>.
56. Johnson N, Maguire S, Morra A, et al. CYP3A7\*1C allele: linking premenopausal oestrone and progesterone levels with risk of hormone receptor-positive breast cancers. *Br J Cancer*. 2021;124(4):842–54. <https://doi.org/10.1038/s41416-020-01185-w>.
57. Johnson N, De Ieso P, Migliorini G, et al. Cytochrome P450 Allele CYP3A7\*1C Associates with Adverse Outcomes in Chronic Lymphocytic Leukemia, Breast, and Lung Cancer. *Cancer Res*. 2016;76(6):1485–93. <https://doi.org/10.1158/0008-5472.CAN-15-1410>.
58. Turner RM, Fontana V, Zhang JE, et al. A Genome-wide Association Study of Circulating Levels of Atorvastatin and Its Major Metabolites. *Clin Pharmacol Ther*. 2020;108(2):287–97. <https://doi.org/10.1002/cpt.1820>.
59. Casirati A, Somaschini A, Perrone M, et al. Preterm birth and metabolic implications on later life: A narrative review focused on body composition. *Front Nutr*. 2022;9:978271. <https://doi.org/10.3389/fnut.2022.978271>.
60. Sun BB, Maranville JC, Peters JE, et al. Genomic atlas of the human plasma proteome. *Nature*. 2018;558(7708):73–9. <https://doi.org/10.1038/s41586-018-0175-2>.
61. Zhu J, O'Mara TA, Liu D, et al. Associations between genetically predicted circulating protein concentrations and endometrial cancer risk. *Cancers (Basel)*. 2021. <https://doi.org/10.3390/cancers13092088>.
62. Salehi B, Zucca P, Orhan IE, et al. Allicin and health: A comprehensive review. *Trends Food Sci Technol*. 2019;86:502–16. <https://doi.org/10.1016/j.tifs.2019.03.003>.
63. Deng Y, Ho C-T, Lan Y, et al. Bioavailability, Health Benefits, and Delivery Systems of Allicin: A Review. *J Agric Food Chem*. 2023;71(49):19207–20. <https://doi.org/10.1021/acs.jafc.3c05602>.
64. Harvey MH, McMillan M, Morgan MR, et al. Solanidine is present in sera of healthy individuals and in amounts dependent on their dietary potato consumption. *Hum Exp Toxicol*. 1985;4(2):187–94. <https://doi.org/10.1177/096032718500400209>.
65. Friedman M. Potato glycoalkaloids and metabolites: roles in the plant and in the diet. *J Agric Food Chem*. 2006;54(23):8655–81. <https://doi.org/10.1021/jf061471t>.
66. Magliocco G, Desmeules J, Matthey A, et al. Metabolomics reveals biomarkers in human urine and plasma to predict cytochrome P450 2D6 (CYP2D6) activity. *Br J Pharmacol*. 2021;178(23):4708–25. <https://doi.org/10.1111/bph.15651>.
67. Gaedigk A. Complexities of CYP2D6 gene analysis and interpretation. *Int Rev Psychiatry*. 2013;25(5):534–53. <https://doi.org/10.3109/09540261.2013.825581>.
68. Wollmann BM, Smith RL, Kringen MK, et al. Evidence for solanidine as a dietary CYP2D6 biomarker: significant correlation with risperidone metabolism. *Br J Clin Pharmacol*. 2024;90(3):740–7. <https://doi.org/10.1111/bcp.15721>.
69. Wollmann BM, Størset E, Kringen MK, et al. Prediction of CYP2D6 poor metabolizers by measurements of solanidine and metabolites—a study in 839 patients with known CYP2D6 genotype. *Eur J Clin Pharmacol*. 2023;79(4):523–31. <https://doi.org/10.1007/s00228-023-03462-y>.
70. Kiiski JI, Neuvonen M, Kurkela M, et al. Solanidine is a sensitive and specific dietary biomarker for CYP2D6 activity. *Hum Genomics*. 2024;18(1):11. <https://doi.org/10.1186/s40246-024-00579-8>.
71. Haugabrooks E. Chapter 1 - the interrelationships between food, nutrition, and toxicology. In: Haugabrooks E, Hayes AW, editors. *History of food and nutrition toxicology*. Academic Press; 2023. p. 1–31.
72. Daubner SC, Le T, Wang S. Tyrosine hydroxylase and regulation of dopamine synthesis. *Arch Biochem Biophys*. 2011;508(1):1–12. <https://doi.org/10.1016/j.jabb.2010.12.017>.

73. Sidharthan NP, Minchin RF, Butcher NJ. 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*. 2013;288(48):34364–74. <https://doi.org/10.1074/jbc.M113.493239>.
74. Yu Z, Kastenmuller G, He Y, et al. Differences between human plasma and serum metabolite profiles. *PLoS One*. 2011;6(7):e21230. <https://doi.org/10.1371/journal.pone.0021230>.
75. de Mutsert R, den Heijer M, Rabelink TJ, et al. The Netherlands Epidemiology of Obesity (NEO) study: study design and data collection. *Eur J Epidemiol*. 2013;28(6):513–23. <https://doi.org/10.1007/s10654-013-9801-3>.
76. Bos MM, Goulding NJ, Lee MA, et al. Investigating the relationships between unfavourable habitual sleep and metabolic traits: evidence from multi-cohort multivariable regression and Mendelian randomization analyses. *BMC Med*. 2021;19(1):69. <https://doi.org/10.1186/s12916-021-01939-0>.
77. McCarthy S, Das S, Kretschmar W, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet*. 2016;48(10):1279–83. <https://doi.org/10.1038/ng.3643> [publishedOnlineFirst:2016/08/23].
78. Price AL, Patterson NJ, Plenge RM, et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*. 2006;38(8):904–9. <https://doi.org/10.1038/ng1847>.
79. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet*. 2009;5(6):e1000529. <https://doi.org/10.1371/journal.pgen.1000529>.
80. Auton A, Brooks LD, Durbin RM, et al. A global reference for human genetic variation. *Nature*. 2015;526(7571):68–74. <https://doi.org/10.1038/nature15393>.
81. Stanaway IB, Hall TO, Rosenthal EA, et al. The eMERGE genotype set of 83,717 subjects imputed to ~40 million variants genome wide and association with the herpes zoster medical record phenotype. *Genet Epidemiol*. 2019;43(1):63–81. <https://doi.org/10.1002/gepi.22167> [publishedOnlineFirst:20181008].
82. Mbatchou J, Barnard L, Backman J, et al. Computationally efficient whole-genome regression for quantitative and binary traits. *Nat Genet*. 2021;53(7):1097–103. <https://doi.org/10.1038/s41588-021-00870-7>.
83. Pe'er I, Yelensky R, Altshuler D, et al. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet Epidemiol*. 2008;32(4):381–5. <https://doi.org/10.1002/gepi.20303>.
84. Li J, Ji L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity (Edinb)*. 2005;95(3):221–7. <https://doi.org/10.1038/sj.hdy.6800717>.
85. Winkler TW, Kutalik Z, Gorski M, et al. EasyStrata: evaluation and visualization of stratified genome-wide association meta-analysis data. *Bioinformatics*. 2014;31(2):259–61. <https://doi.org/10.1093/bioinformatics/btu621>.
86. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*. 2010;26(17):2190–1. <https://doi.org/10.1093/bioinformatics/btq340>.
87. Chang L, Zhou G, Ou H, et al. mGWAS-Explorer: linking SNPs, genes, metabolites, and diseases for functional insights. *Metabolites*. 2022(6). <https://doi.org/10.3390/metabo12060526>.
88. Kamat MA, Blackshaw JA, Young R, et al. PhenoScanner V2: an expanded tool for searching human genotype-phenotype associations. *Bioinformatics*. 2019;35(22):4851–3. <https://doi.org/10.1093/bioinformatics/btz469>.
89. Watanabe K, Taskesen E, van Bochoven A, et al. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun*. 2017;8(1):1826. <https://doi.org/10.1038/s41467-017-01261-5>.
90. Stacey D, Fauman EB, Ziemek D, et al. ProGeM: a framework for the prioritization of candidate causal genes at molecular quantitative trait loci. *Nucleic Acids Res*. 2019;47(1):e3. <https://doi.org/10.1093/nar/gky837>.
91. McLaren W, Gil L, Hunt SE, et al. The Ensembl variant effect predictor. *Genome Biol*. 2016;17(1):122. <https://doi.org/10.1186/s13059-016-0974-4>.
92. McCall MN, Illei PB, Halushka MK. Complex sources of variation in tissue expression data: analysis of the GTEx lung transcriptome. *Am J Hum Genet*. 2016;99(3):624–35. <https://doi.org/10.1016/j.ajhg.2016.07.007>.
93. 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*. 2009;3(1):361–2. <https://doi.org/10.1609/icwsm.v3i1.13937>.
94. Hemani G, Zheng J, Elsworth B, et al. The MR-Base platform supports systematic causal inference across the human phenotype. *Elife*. 2018. <https://doi.org/10.7554/eLife.34408>. ([published Online First: 20180530]).
95. Bowden J, Spiller W, Del Greco MF, et al. Improving the visualization, interpretation and analysis of two-sample summary data Mendelian randomization via the radial plot and radial regression. *Int J Epidemiol*. 2018;47(4):1264–78. <https://doi.org/10.1093/ije/dyy101>.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.