The genomic architecture of blood metabolites based on a decade of genome-wide analyses

Fiona A. Hagenbeek[1,2*], René Pool[1,2], Jenny van Dongen[1,2], Harmen H.M. Draisma[1], Jouke Jan Hottenga[1], Gonneke Willemsen[1], Abdel Abdellaoui[1], Iryna O. Fedko[1], Anouk den Braber[1,3,4], Pieter Jelle Visser[3,5], Eco J.C.N. de Geus[1,2,4], Ko Willems van Dijk[6], Aswin Verhoeven[7], H. Eka Suchiman[8], Marian Beekman[8], P. Eline Slagboom[8], Cornelia M. van Duijn[9], BBMRI Metabolomics Consortium[10], Amy C. Harms[11], Thomas Hankemeier[11], Meike Bartels[1,2,4], Michel G. Nivard[1,2,4*¥] and Dorret I. Boomsma[1,2,4*¥]

[1]Department of Biological Psychology, Vrije Universiteit Amsterdam, Amsterdam, the Netherlands.

[2]Amsterdam Public Health research institute, Amsterdam, the Netherlands.

[3]Alzheimer Center Amsterdam, Department of Neurology, VU Amsterdam, Amsterdam UMC, Amsterdam, The Netherlands.

[4]Amsterdam Neuroscience, Amsterdam, the Netherlands.

[5]Department of Psychiatry and Neuropsychology, School of Mental Health and Neuroscience, Alzheimer Center Limburg, Maastricht University, Maastricht, The Netherlands.

[6]Einthoven Laboratory for Experimental Vascular Medicine, Leiden University Medical Center, Leiden, The Netherlands; Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands; Department of Internal Medicine division Endocrinology, Leiden University Medical Center, Leiden, The Netherlands

[7]Center for Proteomics and Metabolomics, Leiden University Medical Center, Leiden, The Netherlands

[8]Department of Biomedical Data Sciences, section of Molecular Epidemiology, Leiden University Medical Center, Leiden, The Netherlands

[9]Department of Epidemiology, Erasmus Medical Center, Rotterdam, The Netherlands.

[10] Members of the BBMRI Metabolomics Consortium are listed before the references.

[11]Division of Analytical Biosciences, Leiden Academic Center for Drug Research, Leiden University, Leiden, the Netherlands; The Netherlands Metabolomics Centre, Leiden, The Netherlands.

[¥] These authors contributed equally.

[*]Correspondence to: Fiona A. Hagenbeek, Dorret I. Boomsma or Michel G. Nivard, Department of Biological Psychology, Vrije Universiteit Amsterdam, Van der Boechorststraat 7-10, 1081 BT Amsterdam, The Netherlands. E-mail: f.a.hagenbeek@vu.nl; di.boomsma@vu.nl; m.g.nivard@vu.nl

Word count:

Abstract: 147; Main text: 3,520; Methods: 3,001;

References incl. methods: 79; Tables: 4; Figures: 3

Supplementary Material:

Supplementary Notes: 4; Supplementary Figures: 6; Supplementary Tables: 19; Supplementary Data: 1

Abstract

Metabolomics examines the small molecules involved in cellular metabolism. Approximately 50% of total phenotypic differences in metabolite levels is due to genetic variance, but heritability estimates differ across metabolite classes and lipid species. We performed a review of all genetic association studies, and identified > 800 class-specific metabolite loci that influence metabolite levels. In a twin-family cohort (*N* = 5,117), these metabolite loci were leveraged to simultaneously estimate total heritability (h^2_{total}), and the proportion of heritability captured by known metabolite loci ($h^2_{Metabolite\text{-}hits}$) for 309 lipids and 52 organic acids. Our study revealed significant differences in $h^2_{Metabolite\text{-}hits}$ among different classes of lipids and organic acids. Furthermore, phosphatidylcholines with a high degree of unsaturation had higher $h^2_{Metabolite\text{-}hits}$ estimates than phosphatidylcholines with a low degree of unsaturation. This study highlights the importance of common genetic variants for metabolite levels, and elucidates the genetic architecture of metabolite classes and lipid species.

The metabolome is defined as the collection of metabolites, i.e., small molecules involved in cellular metabolism, which are produced in cells[1] and consist of many classes[2–5]. The overall aim of metabolomics is to provide a holistic overview of the metabolome[1], and its role in biological mechanisms and metabolic disturbances in diseases. Elucidating this role may offer new therapeutic targets or new biomarkers for disease diagnosis[6]. Variation in metabolite levels can arise due to gender[7], and age[8], as well as physiologic effects, behavior, and lifestyle factors, such as diet[9]. Genetic differences may be a source of direct variation in metabolomics profiles, or an indirect source of variation through genetic influences on physiology, behavior, and (or) lifestyle.

Genome- and metabolome-wide analysis of common genetic variants in human metabolism have successfully identified genetically influenced metabolites[10]. In 2008, the first genome-wide association study (GWAS; N = 284 participants) identified four genetic variants associated with metabolite levels[11]. Thereafter, GWAS with increasing sample sizes, and in diverse populations, identified hundreds of Single Nucleotide Polymorphism (SNP) associations with metabolites from a wide range of metabolite classes[10]. Additional metabolite loci have been identified by leveraging low-frequency and rare-variant analyses using (exome-) sequencing. We conducted a comprehensive review of all quantitative trait locus (QTL) discovery for metabolites and supply the complete reference list in **Supplementary Note 1**.

Twin and family studies have established that the heritability (h^2; proportion of phenotypic variance due to genetic factors) of metabolite levels is 50% on average, with a range from h^2 = 0% to h^2 = 80%[9,12–19]. Several studies reported differences in heritability estimates among different classes of lipid species[16,18] or lipoprotein subclasses[17]. For example, Rhee et al. (2013) reported higher heritability estimates for amino acids than for lipids[15]. Essential amino acids, which cannot be synthesized by an organism *de novo*[20], had lower heritability than non-essential amino acids[15], that are synthesized within the body[20]. Several techniques are available to estimate the contribution of measured SNPs to trait heritability[21], and, given SNP data in family members, to simultaneously estimate SNP-associated (h^2_{SNP}) and pedigree-associated genetic variance (h^2_{ped})[22]. Together the SNP- and pedigree-associated genetic effects account for the narrow-sense heritability. However, when including data of family members, the variance explained by genetic effects (h^2_{total}) may be biased upwards by shared environmental factors and/or non-additive genetic effects [22,23].

An improved understanding of the genetic architecture of metabolites will benefit our understanding of the aetiology of diseases and traits, such as cardiometabolic diseases[24], migraine[25], psychiatric disorders[26], and cognition[27]. Here we aim to further our understanding of the contribution of genetic factors to variation in fasting blood metabolic measures (henceforth referred to as *metabolites* for brevity) by the analysis of data from multiple metabolomics platforms in a large cohort of twins and family members (N = 5,117). Specifically, we aim to estimate the total genetic variance of metabolite levels (h^2_{total}), and to elucidate the contribution to metabolite levels of known metabolite class-specific and metabolite class-unspecific loci ($h^2_{Metabolite\text{-}hits}$), on the basis of the results of a decade of GWA and (exome-) sequencing studies (**Supplementary Data 1**). To this end, we characterized all published metabolite-SNP associations by metabolite classification, and used linear-mixed models to estimate the h^2_{total}, h^2_{SNP} and $h^2_{Metabolite\text{-}hits}$ simultaneously for 369 metabolites (**Figure 1**). In these models, the $h^2_{Metabolite\text{-}hits}$ consists of two variance components, a component attributable to metabolite loci associated with metabolites of a specific superclass ($h^2_{Class\text{-}hits}$) and a component attributable other metabolite loci ($h^2_{Notclass\text{-}hits}$; **Figure 1**). We further expand on the current knowledge of the genetic aetiology of metabolite classes by employing mixed-effect meta-regression models to test differences in heritability estimates among metabolite classes and among lipid species.

Intriguingly, phosphatidylcholines[14] and triglycerides (TGs)[19] show increasing heritability with increasing number of carbon atoms and/or double bonds in their fatty acyl side chains. Draisma et. al speculated this might be attributable to differences in the number of metabolic conversion rounds for phosphatidylcholines or TGs with a variable number of carbon atoms[14]. To distinguish between the effects of the number of carbon atoms or number of double bonds in the fatty acyl side chains of phosphatidylcholines and TGs, we conducted additional univariate follow-up analyses.

Results

Metabolite classification

In the period of November 2008 to October 2018, 40 GWA and (exome-) sequencing studies identified 242,580 metabolite-SNP or metabolite ratio-SNP associations (see **Supplementary Note 1**). All 242,580 associations may be found at: http://bbmri.researchlumc.nl/atlas/#data,which lists the significant SNP-metabolite associations by study. These associations included 1,804 unique metabolites or ratios, and 49,231 unique SNPs (43,830 after converting all SNPs to NCBI build 37; **Supplementary Data 1**). The Human Metabolome Database (HMDB)[2–5] identifiers of each metabolite were retrieved in order to extract information concerning the metabolite's hydrophobicity and chemical classification (see **Methods**). Excluding the ratios and unidentified metabolites, we classified 953 metabolites into 12 'super classes' (**Table 1**), 43 'classes', or 77 'subclasses' based on the HMDB classification (**Supplementary Data 1**). The majority of the metabolites were classified into the super classes lipids or organic acids. The lipids could be subdivided into 8 classes, with 1 to 95,795 metabolite-SNP associations per class (mean = 17,589; SD = 32,553), and in 32 subclasses, with the number of subclass metabolites-SNP associations ranging from 1 to 40,440 (mean = 4,673; SD = 9,124). The organic acids and derivatives were divided in 9 classes, with the number of metabolite-SNP associations ranging from 1 to 26,832 (mean = 3,374; SD = 8,832). The organic acids and derivatives were also divided into 17 organic acid subclasses, with the number of subclass metabolite-SNP associations ranging from 1 to 26,448 (mean = 1,786; SD = 6,371; **Supplementary Data 1**). Across all four platforms 427 metabolites were assessed. After excluding the ratios (17) and the metabolites of super classes not included in the curated metabolite-SNP association list (8), data were available for 402 metabolites. The full list of metabolites, with their classifications and the quartile values of the untransformed levels, are included in **Supplementary Table 1**. The 402 metabolites were classified as 336 lipids, 53 organic acids, 9 organic oxygen compounds, 3 proteins and one organic nitrogen compound, these super classes were consisted of 12 classes (**Supplementary Table 2**). In this paper we mainly focus on the first two super classes. After quality control (QC), 369 metabolites from these two super classes were retained for analysis.

Characterization of the heritable influences on lipid and organic acid levels

Data of 5,117 participants were available from the following four metabolomics platforms: the Nightingale Health ^{1}H-NMR platform, a UPLC-MS Lipidomics platform, the Leiden ^{1}H-NMR platform, and the Biocrates Absolute-IDQ^{TM} p150 platform. The participants were registered with the Netherlands Twin Register (NTR)[28] and were clustered in 2,445 nuclear families. Metabolomics and SNP data were available for all participants. Background and demographic characteristics for the sample can be found in **Table 2**.

We aimed to assess the variance explained by previously identified metabolite GWA and (exome-) sequencing genetic variants in our (independent) sample. Clearly, our results are conditional on the power of past the studies, as the list of metabolite genetic variants is based on previous GWA and (exome-) sequencing studies, which vary in power. We present the sample size of each past study in **Supplementary Note 1**, and the sample size per metabolite-SNP association in **Supplementary data 1**.

Linear-mixed models including all loci for genetic variants associated with metabolites in a single genetic relatedness matrix (GRM) will contain SNPs that are associated with some metabolites, but not with others, or include many SNPs that are not associated with a given metabolite. We therefore created two GRMs for the loci associated with metabolite hits (see **Methods**): one class-specific and one non-class specific (i.e. GRMs including metabolite loci for all metabolites, except for the target metabolite class). We explored models for the 12 class-specific and the corresponding not-class specific GRMs (**Supplementary Note 2**). These models displayed high degrees of non-convergence (37.9% total), with models including small class-specific GRMs displaying more non-convergence (**Supplementary Table 2**). Therefore, the results in the remainder of this paper were based on the metabolite super classes, i.e. lipids and organic acids.

For the 369 lipids and organic acids, we carried out unconstrained four-variance component analyses (**Figure 1**). In genome-wide complex trait analysis (GCTA)[21] we specified a model in which we partition the metabolite variation into SNP-associated (h^2_{SNP}), pedigree-associated (h^2_{ped}), class-specific metabolite-loci-associated ($h^2_{class\text{-}hits}$), and not-class metabolite-loci-associated ($h^2_{notclass\text{-}hits}$) genetic variation (**Figure 1**). We report the total heritability (h^2_{total}), the proportion attributable to metabolite superclass-specific loci ($h^2_{Class\text{-}hits}$), the proportion of variance attributable to non-superclass metabolite loci ($h^2_{Notclass\text{-}hits}$) and the contribution of known metabolite loci to metabolite levels ($h^2_{Metabolite\text{-}hits}$). The analyses were performed separately for lipids and organic acids, with class-specific and corresponding non-class GRMs (created using the LDAK program[29,30]) in both sets of analyses. The lipid analyses employed a class-specific GRM of 479 lipid loci and a corresponding non-class GRM of 596 loci (**Supplementary Figure 1**). The organic acid analyses included a class-specific GRM of 397 loci and a non-class GRM of 683 loci (**Supplementary Figure 1**). Before the analyses, the metabolite data were normalized (log-normal or inverse rank; **see Methods**). All models included age at blood draw, sex, the first 10 principal components (PCs) from SNP genotype data, genotyping chip and metabolomics measurement batch as covariates.

Supplementary Table 3 includes the estimates from the four-variance genetic component models for all 369 metabolites. The genomic relatedness matrix residual maximum likelihood (GREML) algorithm converged for 361 (97.8%) of the 53 organic acids and 316 lipids (**Supplementary Table 4**). Non-convergence of the GREML algorithm was observed for 6 metabolites (1.6%). The analyses of 2 metabolites (0.5%) were not completed due to non-invertible variance-covariance matrices. The estimates for h^2_{total} of the 309 lipids ranged from 0.11 to 0.66 (mean = 0.47; mean s.e. = 0.04). The estimates for $h^2_{Metabolite\text{-}hits}$ ranged from -0.05 to 0.16 (mean = 0.06; mean s.e. = 0.03; **Table 3**). The 52 organic acids had h^2_{total} estimates ranging from 0.14 to 0.72 (mean = 0.41; mean s.e. = 0.04). The estimates for $h^2_{Metabolite\text{-}hits}$ ranged from -0.08 to 0.11 (mean = 0.01; mean s.e. = 0.02; **Table 3**). On average, for both lipids and organic acids the h^2_{class} was higher than the $h^2_{Notclass}$, with $h^2_{Class\text{-}hits}$ ranging from -0.02 to 0.16 (0.06; mean s.e. = 0.02) for lipids and from -0.04 to 0.14 for organic acids (mean = 0.01; mean s.e. = 0.02). For both lipids and organic acids $h^2_{Notclass\text{-}hits}$ was zero (mean s.e. = 0.02), ranging from -0.06 to 0.12 for lipids and from -0.06 to 0.05 for organic acids (**Table 3**).

Including multiple metabolomics platforms allowed for a comparison of metabolites as measured on multiple platforms. An earlier study showed that 29 out of 43 metabolites present on two platforms to exhibit moderate heritability on both platforms[31]. In the current study, 61 metabolites were measured on multiple platforms (phenotypic correlations provided in **Supplementary Table 5**), with moderate h^2_{total} on each of the platforms and on average a positive correlation of 0.36 between the $\mathrm{h}^2_{\mathrm{total}}$ of the same metabolite assessed on different platforms (**Supplementary Table 5**).

Differential heritability among metabolite classes and lipid-species

Figure 2 shows variation in median heritability among the following classes of organic acids: keto acids, hydroxy acids and carboxylic acids (see **Supplementary Table 1** for metabolites per class). Keto acids, followed by carboxylic acids, had the highest median h^2_{total}, and $h^2_{Class\text{-}hits}$ estimates (**Figure 2**). While hydroxy acids had the highest median $h^2_{Notclass\text{-}hits}$ and $h^2_{Metabolite\text{-}hits}$ estimates, the lowest median h^2_{total}, and $h^2_{Class\text{-}hits}$ estimates were observed for these metabolites (**Figure 2**). To investigate whether heritability differs significantly among classes of organic acids, we applied multivariate mixed-effect meta-regression, corrected for metabolite platform effects (see **Methods**). The multivariate mixed-effect meta-regression models showed that h^2_{total} and $h^2_{Class\text{-}hits}$ for the organic acid classes did not differ significantly. However , significant differences among the organic acid classes, though, were observed with respect to the $h^2_{Metabolite\text{-}hits}$ estimates ($F(4, 47) = 3.44$, FDR-adjusted p-value = 0.03), and the $h^2_{Notclass\text{-}hits}$ estimates ($F(4,47) = 19.95$, FDR-adjusted p-value = $1.25\text{x}10^{-08}$; **Supplementary Table 6**).

The multivariate mixed-effect meta-regressions were also applied to assess the significance of heritability differences among essential and non-essential amino acids (subdivision of carboxylic acids; see **Supplementary Table 7**) and among lipid classes (see **Supplementary Table 1** for metabolites per lipid class). The meta-regression analyses revealed no significant mean differences among essential and non-essential amino acids (**Table 4**; **Supplementary Table 8**). Small but significant median heritability differences were observed among the different classes of lipids (**Figure 3**). For lipid classes the $h^2_{Metabolite\text{-}hits}$ estimates differed significantly ($F(8, 300) = 8.47$; FDR-adjusted p-value = 0.004; **Supplementary Table 6**).

Finally, we explored whether heritability of phosphatidylcholines and TGs increases with a larger number of carbon atoms and/or double bonds in their fatty acyl side chains. To this end we employed both uni- and multivariate mixed-effect meta-regression models separately for the TGs, diacyl phosphatidylcholines (PCaa) and acyl-alkyl phosphatidylcholines (PCae; see **Methods**). The platform specific heritability estimates for each of these lipid species are depicted in **Supplementary Figure 2**. Variation in the number of carbon atoms and double bonds was significantly associated with $h^2_{Metabolite-hits}$ estimates for PCaa's (F(3, 52) = 7.05; FDR-adjusted p-value = 0.009) and PCae's (F(3, 45) = 3.41; FDR-adjusted p-value = 0.05; **Supplementary Table 6**). Phosphatidylcholines with a larger number of carbon atoms showed lower heritability estimates and phosphatidylcholines with a larger number of double bonds had higher heritability estimates (**Supplementary Table 6**). The differences among the phosphatidylcholines with a variable number of carbon atoms and/or double bonds may have contributed to differential h^2_{Class} estimates. Univariate models confirmed the results for the number of double bonds in PCaa's and PCae, though they were not significant after correction for multiple testing (**Supplementary Table 8**).

Discussion

We carried out a comprehensive assessment of GWA-metabolomics studies, and created a repository of all studies reporting on associations of SNPs and blood metabolites in European ancestry samples. We curated 241,965 genome-wide metabolite associations and we classified the associated metabolites into super classes, classes and sub-classes. The complete overview of all blood metabolite-SNP associations is provided in **Supplementary Data 1** (http://bbmri.researchlumc.nl/atlas/#data), with the complete list of references in **Supplementary Note 1**. The information from the repository was used to construct GRMs, which served to identify genetic variance components in the analysis of 369 metabolites. The metabolite data in our study came from a large cohort of twin-families (N = 5,117 clustered in 2,445 families) measured on four metabolomics platforms. We focused on two metabolite super classes. By mapping all metabolites to the Human Metabolome Database (HMDB)[2–5] we were able to classify both the measured metabolites and all previously published metabolites as either lipids or organic acids. In the current study, we sought to elucidate the contribution of known metabolite loci, based on a decade of GWA and (exome-) sequencing studies, to metabolite levels ($h^2_{Metabolite\text{-}hits}$). A unique feature of our study was the ability to disentangle the role of class-specific ($h^2_{Class\text{-}hits}$) and non-class ($h^2_{Notclass\text{-}hits}$) metabolite loci on heritability differences among metabolite classes and lipid species.

To evaluate differences among metabolite classes and lipid species in the estimates for h^2_{total}, we applied multivariate mixed-effect meta-regression models to the estimates of $h^2_{Metabolite\text{-}hits}$, $h^2_{Class\text{-}hits}$, and $h^2_{Notclass\text{-}hits}$. We observed no significant differences in h^2_{total} estimates among the metabolite classes. Consistent with a previous twin-family study[13], none of the heritability estimates differed significantly among essential and non-essential amino acids. We observed significant $h^2_{Metabolite\text{-}hits}$ differences among the different classes of organic acids. Keto acids had significantly lower $h^2_{Metabolite\text{-}hits}$ estimates as compared with carboxylic acids. Class-specific metabolite loci heritability estimates for fatty acyls, lipoproteins and steroids were significantly higher. Similarly, significant heterogeneity in lipid class heritability, with lower h^2_{total} and h^2_{SNP} for phospholipids than for sphingolipids or glycerolipids has been reported[16,18,32]. Lastly, we assessed whether heritability increases with added complexity in lipid species[14,19]. We found that this was the case with respect to $h^2_{Metabolite\text{-}hits}$ estimates in more complex diacyl and acyl-alkyl phosphatidylcholines, but not for more complex TGs. Previous research reported significant higher h^2_{SNP} estimates in polyunsaturated fatty acid containing lipids[18]. Furthermore, loci associated with traditional lipid measures explained 2% to 21% of the variance in lipid levels[18]. Together these results suggest that higher heritability in phosphatidylcholines is driven by a lower number of carbon atoms and higher number of double bonds, e.g., a larger degree of unsaturation.

Evaluating the mean heritability differences among lipids and organic acids, it appears that lipids have higher h^2_{total}, $h^2_{Class\text{-}hits}$ and $h^2_{Metabolite\text{-}hits}$ estimates than organic acids (**Table 3**). Previous twin-family studies indicates that the heritability difference among lipids and organic acid is rarely investigated[12–15].

This is possibly because most metabolomics platforms focus mainly on either lipids or organic acids. Lipid metabolite classes tend to be very well represented on metabolomics platforms, whereas organic acids are unrepresented, and as a consequence, the analysis to obtain $h^2_{\textit{Class-hits}}$ and $h^2_{\textit{Metabolite-hits}}$ estimates of the organic acids will be underpowered due to this imbalance.

Limitations

The current study has several limitations. First, the extent to which our findings generalize to populations of non-European ancestry is unknown. Loci of common human metabolism pathways are most likely to replicate over ethnicities[33]. Second, estimates of the total variance explained may show upward bias when based on data from closely related individuals (e.g., first cousins or closer)[22,23]. This bias is caused by the influence of shared environmental influences, epistatic interactions, or dominance[22,23]. While the results of the current study may suffer of such biases by the inclusion of twins, siblings and parents, the sample also includes many unrelated individuals which will reduce the possible bias (**Supplementary Figure 3**).

Kettunen et al. (2012) investigated 217 metabolites of the Nightingale Health ^{1}H-NMR platform in a classical twin design and reported dominance effects for 6.45% of the metabolites[34]. Tsepsilov et al. (2015) performed GWA study targeting non-additive genetic effects and concluded that most genetic effects on metabolite levels and ratios were in fact additive[35]. Together, these studies suggested that the bias due to dominance effects on metabolite levels will be minor.

Relatively few twin-family studies explicitly investigated the role of shared environmental influences on metabolite levels. Overall, shared environmental influences are reported for a small number of metabolites (e.g., 14.3% of all Nightingale Health ^{1}H-NMR metabolites[34]) and the influence of the shared environment is small-to-moderate (platform and metabolite class-dependent averages range from 0.03 to 0.45[9,16,36–38] with larger estimates deriving from small studies). For studies including parents and offspring, or adult twin and siblings pairs the question arises which effects are captured by the shared environment. Are these the lasting influences of the environment offspring shared with their parents and with each other before they started living independently? Additional research is necessary to elucidate the role of the shared environment on metabolite levels[22].

Third, standard errors of h^2_{SNP} estimates were high. While we have included all h^2_{SNP} estimates in the supplements, we stress that the primary goal of our paper was to investigate the contribution of known metabolite loci in an independent sample rather than obtaining the h^2_{SNP} estimates for metabolites.

Finally, the estimates for $h^2_{metabolite\text{-}hits}$ are based on SNPs of 40 different studies from a decade of GWA and (exome-) sequencing studies. The sample size, and therefore the power, of these studies vary, with some studies conducted with as few as 211 individuals while others included over 24,000 individuals (**Supplementary Note 1**). For underrepresented metabolites the low power may result in downward biased heritability estimates. However, leveraging information from a decade of research in 40 studies and extracting loci for metabolite classes across multiple studies, the number of such metabolites is not large. New[32,39–41] and future studies will increase the number of variants identified as metabolite loci. The investment in UK Biobank[42] is expected to dramatically increase sample sizes for large-scale genomic investigations of the human metabolome and subsequently the number of metabolite loci.

Future directives and conclusions

Mendelian Randomization may benefit from the comprehensive overview of metabolite loci that we identified. The identified loci can serve as instruments in metabolome-wide Mendelian Randomization studies of complex traits. In addition, our work offers valuable insights into the role of common genetic variants in class specific heritability differences among metabolite classes and lipids species. Further research is required to elucidate the contribution of rare genetic variants to metabolite levels, and differences in the contribution of rare genetic variants among metabolite classes. A reasonable approach would be to carry out a similar study in a large sample of whole-genome sequencing (WGS) data. Such an approach, using MAF- and LD-stratified GREML analysis[43], identified additional variance due to rare variants for height and BMI[44].

In conclusion, we contributed to our understanding of the genetic architecture of fasting blood metabolite levels, and of differences in the genetic architecture among metabolite classes. Extending the GREML framework with the inclusion of known metabolite loci allowed us to simultaneously estimate h^2_{total}, and $h^2_{metabolite\text{-}hits}$ (which consists of $\mathrm{h}^2_{\mathrm{Class}\text{-}hits}$ and $h^2_{Notclass\text{-}hits}$) for 361 metabolites. Significant differences in $h^2_{Metabolite\text{-}hits}$ estimates were observed among different classes of lipids and organic acids and for more complex diacyl and acyl-alkyl phosphatidylcholines. Future studies should address the proportion of metabolite variation influenced by heritable and non-heritable lifestyle factors, as this will facilitate the development of personalized disease prevention and treatment of complex disorders.

Methods

Participants

At the Netherlands Twin Register (NTR)[45] metabolomics data for twins and family members as measured in blood samples were available for 6,011 individuals of whom 5,667 were genotyped. The blood samples for the four metabolomics experiments described in this study were mainly collected in participants of the NTR biobank project[28,46]. Blood samples were collected after a minimum of two hours of fasting (1.3%), with the majority of the samples collected after overnight fasting (98.7%). Fertile women were bled in their pill-free week or on day 2-4 of their menstrual cycle. For the current paper, we excluded participants who were not of European ancestry, who were on lipid-lowering medication at the time of blood draw, and who failed to adhere to the fasting protocol. The exact number of exclusions per dataset is listed in **Supplementary Table 9**. After completing the preprocessing of the metabolomics data, the separate subsets (e.g., different collection and measurement waves; see **Supplementary Table 9**) of each platform were merged into a single per platform dataset, retaining a single (randomly chosen) observation per platform when multiple observations were available. **Supplementary Table 10** gives an overview of the overlap in participants among the different platforms, with the overlap among each metabolite that survived quality control (QC) for all four platforms available in **Supplementary Table 11**. The final number of participants included in the study was 5,117, with platform specific sample size ranging from 1,448 to 4,227 individuals clustered in 946 to 2,179 families. Characteristics for the individuals can be found in **Table 2**. **Supplementary Figure 3** depicts the distribution of the relatedness in the sample. Informed consent was obtained from all participants. Projects were approved by the Central Ethics Committee on Research Involving Human Subjects of the VU University Medical Centre, Amsterdam, an Institutional Review Board certified by the U.S. Office of Human Research Protections (IRB number IRB00002991 under Federal-wide Assurance- FWA00017598; IRB/institute codes, NTR 03-180 and EMIF-AD 2014.210).

Metabolite profiling

Nightingale Health ^{1}H-NMR platform

Metabolic biomarkers were quantified from plasma samples using high-throughput proton nuclear magnetic resonance spectroscopy (^{1}H-NMR) metabolomics (Nightingale Health Ltd, Helsinki, Finland; formerly Brainshake Ltd.). This method provides simultaneous quantification of routine lipids, lipoprotein subclass profiling with lipid concentrations within 14 subclasses, fatty acid composition, and various low-molecular weight metabolites including amino acids, ketone bodies and glycolysis-related metabolites in molar concentration units. Details of the experimentation and epidemiological applications of the NMR metabolomics platform have been reviewed previously[47,48].

UPLC-MS lipidomics platform

Plasma lipid profiling was performed at the division of Analytical Biosciences at the Leiden Academic Center for Drug Research at Leiden University/Netherlands Metabolomics Centre. The lipids were analyzed with an Ultra-High-Performance Liquid Chromatograph directly coupled to an Electrospray Ionization Quadruple Time-of-Flight high resolution mass spectrometer (UPLC-ESI-Q-TOF; Agilent 6530, San Jose, CA, USA) that uses reference mass correction. For liquid chromatographic separation a ACQUITY UPLC HSS T3 column (1.8µm, 2.1 ∗ 100mm) was used with a flow of 0.4 ml/min over a 16 minute gradient. Lipid detection was done using a full scan in the positive ion mode. The raw MS data were pre-processed using Agilent MassHunter Quantitative Analysis software (Agilent, Version B.04.00). Detailed descriptions of lipid profiling and quantification have been described previously[49,50].

Leiden ^{1}H-NMR platform (for small metabolites)

The Leiden ^{1}H-NMR spectroscopy experiment of EDTA-plasma samples used a 600 MHz Bruker Advance II spectrometer (Bruker BioSpin, Karlsruhe, Germany). The peak deconvolution method used for this platform has been previously described[51].

Biocrates Absolute-IDQ^{TM} p150 platform

The Biocrates Absolute-IDQ^{TM} p150 (Biocrates Life Sciences AG, Innsbruck, Austria) metabolomics platform on serum samples was analyzed at the Metabolomics Facility of the Genome Analysis Centre at the Helmholtz Centre in Munich, Germany. This platform utilizes flow injection analysis coupled to tandem mass spectrometry (MS/MS) and has been described in detail elsewhere[7,52,53].

Metabolomics data preprocessing

Preprocessing of the metabolomics data was done separately for each of the platforms and each measurement batch. Metabolites were excluded from analysis when the mean coefficient of variation exceeded 25% and the missing rate exceeded 5%. Metabolite measurements were set to missing if they were below the lower limit of detection or quantification or could be classified as an outlier (five standard deviations greater or smaller than the mean). Metabolite measurements, which were set to missing because they fell below the limit of detection/quantification were imputed with half of the value of this limit, or when this limit was unknown with half of the lowest observed level for this metabolite. All remaining missing values were imputed using multivariate imputation by chained equations ('mice')[54]. On average, 9 values were imputed for each metabolite (SD = 12; range: 1-151). Data for each metabolite on both ^{1}H-NMR platforms were normalized by inverse normal rank transformation[51,55], while the imputed values of the Biocrates metabolomics platform and the UPLC-MS lipidomics platform were normalized by natural logarithm transformation[14,56], conform previous normalization strategies applied to the data obtained using these platforms. The complete lists with full names of all detected metabolites that survived QC and preprocessing for all platforms can be found in **Supplementary Table 1**, these tables also include the quartile values of the untransformed metabolites.

Genotyping, imputation and ancestry outlier detection

Genotype information was available for 21,001 NTR participants from 6 different genotyping arrays (Affymetrix 6.0 [*N* = 8,640], Perlegen-Affymetrix [*N* = 1,238], Illumina Human Quad Bead 660 [*N* = 1,439], Affymetrix Axiom [*N* = 3,144], Illumnia GSA [*N* = 5,938] and Illumina Omni Express 1M [*N* =238]), as well as sequence data from the Netherlands reference genome project GONL (BGI full sequence at 12x (*N* = 364)[57]. For each genotyping array samples were removed if they had a genotype call rate above 90%, gender-mismatch occurred or if heterozygosity (Plink F statistic) fell outside the range of -0.10 – 0.10. SNPs were removed if they were palindromic AT/GC SNPs with a minor allele frequency (MAF) range between 0.4 and 0.5, if the MAF was below 0.01, if Hardy Weinberg Equilibrium (HWE) had $p < 10^{-5}$, and if the number of Mendelian errors was greater than 20 and the genotype call rate was < 0.95. After QC the six genotyping arrays were aligned to the GONL reference set (V4) and SNPs were removed if the alleles mismatched with this reference panel or the allele frequency different more than 0.10 between the genotyping array and this reference set.

The data from the six genotyping chips were subsequently merged into a single dataset (1,781,526 SNPs). Identity-by-decent (IBD) was estimated with PLINK[58] and KING[59] for all individual pairs based on the ~10.6K SNPs in common across the arrays. Next IBD was compared to expected family relations and individuals were removed in the event of a mismatch. Prior to imputation to the GONL reference data[60,61] the duplicate monozygotic pairs (N = 3,032) or trios (N = 7) and NTR GONL samples (N = 364) were removed and the data was cross-array phased using MACH-ADMIX[62]. Post-imputation the NTR GONL samples and the duplicated MZ pairs and trios were re-turned to the dataset. Filtering of the imputed dataset included the removal of SNPs that were significantly associated with a single genotyping chip ($p < 10^{-5}$), had HWE $p < 10^{-5}$, the Mendelian error rate > mean + 3 SD, or imputation quality (R^2) below 0.90. The final cross-platform imputed dataset included 1,314,639 SNPs, including 20,792 SNPs on the X-chromosome.

The cross-platform imputed data was aligned with PERL based "HRC or 1000G Imputation preparation and checking" tool (version 4.2.5; https://www.well.ox.ac.uk/~wrayner/tools). The remaining 1,302481 SNPs were phased with EAGLE[63] for the autosomes, and SHAPEIT[64] for chromosome X and then imputed to 1000 Genomes Phase 3 (1000GP3 version 5)[65] on the Michigan Imputation server using Minimac3 following the standard imputation procedures of the server[66]. Principal Component Analysis (PCA) was used to project the first 10 PCs of the 1000 genomes references set population on the NTR cross-platform imputed data using SMARTPCA[67]. Ancestry outliers (non-Dutch ancestry; N = 1,823) were defined as individuals with PC values outside the European/British population range[68]. After ancestry outlier removal the first 10 PCs were recalculated.

Curation of metabolite loci

In October 2018 PubMed and Google Scholar were searched to identify published GWA and (exome-) sequencing studies on metabolomics or fatty acid metabolism in blood samples using ^{1}H-NMR, mass spectrometry or gas chromatography-based methods. In the period of November 2008 to October 2018 40 GWA or (exome-) sequencing studies on blood metabolomics in European samples were published (**Supplementary Note 1**). The genome-wide significant ($p < 5x10^{-8}$) metabolite-SNP associations of all studies were extracted, including only those observations for autosomal SNPs and reporting SNP effect sizes and p-values based on the summary statistics excluding NTR samples[55,56]. In the 40 studies, 242,580 metabolite-SNP or metabolite ratio-SNP associations were reported. These associations included 1,804 unique metabolites or ratios and 49,231 unique SNPs (**Supplementary Data 1**). For all metabolites their Human Metabolome Database (HMDB)[3–5], PubChem[69], Chemical Entities of Biological Interest (ChEBI)[70] and International Chemical Identifier (InChiKey)[71] identifiers were retrieved. Information with regards to the 'super class', 'class' and 'subclass' of metabolites was extracted from HMDB. If no HMDB identifier was available and categorization information could not be extracted, 'super class', 'class' and 'subclass' were provided based on expert opinion. Excluding the ratios and unidentified metabolites, 953 metabolites were classified into 12 'super classes', 43 'classes' or 77 'subclasses' (**Supplementary Data 1**). Based on the metabolite identifiers we also extracted the *log(S)* value for each metabolite to assess the hydrophobicity of the metabolites. The *log(S)* value represents the log of the partition coefficient between 1-octanol and water, two fluids that hardly mix. The partition coefficient is the ratio of concentrations in water and in octanol when a substance is added to an octanol-water mixture and hence indicates the hydrophobicity of a compound. Thus, we classified a metabolite as hydrophobic if it is more hydrophobic than 1-octanol, and as hydrophilic otherwise (**Supplementary Data 1**).

The rsIDs or chromosome-base pair positions of the 49,231 unique SNPs were reported by different genome builds or dbSNP maps[72], therefore we lifted all SNPs to HG19 build 37[73], after which 43,830 unique SNPs remained (**Supplementary Figure 1**; **Supplementary Data 1**). All bi-allelic metabolite SNPs were extracted from our 1000GP3 data, which excluded 295 tri-allelic SNPs, and 4,256 SNPs that could not be retrieved from 1000GP3. Next, MAF > 1% (2,067 SNPs removed), $R^2 > 0.70$ (2,002 SNPs) and HWE $P < 10^{-4}$ (72 SNPs) filtering was performed, resulting in 35,138 metabolite SNPs for NTR participants (**Supplementary Figure 1**). Next, we created two 'super class'-specific lists of metabolite loci and two 'not-superclass' lists of metabolite loci. To create a list of loci associated with the 652 unique metabolites classified as 'lipids and lipid-like molecules' (e.g., lipids), we clumped (PLINK version 1.9) all 112,760 lipid-SNP associations using an LD-threshold (r^2) of 0.10 in a 500kb radius in 2,500 unrelated individuals (**Supplementary Figure 1**). Clumping identified 482 lead SNPs, or loci for lipids. An additional 12,169 SNPs were identified as LD-proxies for the lipid-loci (**Supplementary Figure 1**). To obtain the 'not-superclass' list of lipid loci the 12,651 lipid loci and proxies were removed from the list of all metabolite-SNP associations and the resulting list was clumped to obtain the 598 'non-superclass' loci (**Supplementary Figure 1**). The same clumping procedure was applied to the 26,352 organic acid-SNP associations, identifying 398 organic acids loci, 10,781 organic acid LD-proxies and 687 'non-superclass' loci (**Supplementary Figure 1**).

Construction of genetic relationship matrices

In total six weighted genetic relationship matrixes (GRMs) were constructed, which were corrected for uneven and long-range LD between the SNPs (LDAK version 4.9[29,30]). In **Supplementary Note 3** the use of weighted versus unweighted GRMs is compared using simulations. Two of the GRMs used the cross-platform imputed dataset as backbone and the other four GRMs were based on SNPs extracted from the 1000GP3 imputed data. Before calculating the first GRM, the autosomal SNPs of the cross-platform imputed dataset were filtered on MAF (<1%) and all lipid and organic acid loci, their LD-proxies and 50kb surrounding both types of SNPs were removed (see **curation of metabolite loci**; **Supplementary Figure 1**). The LDAK GRM was created after removal of the 50kb surrounding the lipid and organic acid loci and their LD-proxies (as obtained by the clumping procedure as described above) and included 434,216 SNPs (**Supplementary Figure 1**). The *V(G1)* variance component in the genomic relatedness matrix residual maximum likelihood (GREML) analyses is based on this GRM (see **heritability analyses**; **Figure 1**). The *V(G2)* variance component in the GREML analyses is based on the LDAK GRM including all autosomal SNPs with a MAF greater than 1% included on the cross-platform imputed dataset (447,794 SNPs), where ancestry outliers were removed, and genome sharing was set to zero for all individual pairs sharing less than 0.05 of their genome[22] (**Figure 1**). Depending on the metabolite the *V(G3)* variance component in the GREML analyses was either based on an LDAK GRM of the 1000GP3 extracted lipid loci (479 SNPs) or the organic acid loci (397 SNPs), as obtained after the clumping procedure as described above (**Supplementary Figure 1**; **Figure 1**). Finally, depending on the metabolite either the ‘not-lipid’ LDAK GRM (596 SNPs) or the ‘not-organic acid’ LDAK GRM (683 SNPs) provided the *V(G4)* variance component in the GREML analyses (**Supplementary Figure 1**; **Figure 1**). The not-class metabolite loci on which the LDAK GRMs were build were obtained by the clumping procedure as described above (**Supplementary Figure 1**). **Supplementary Data 1** indicates for each listed SNP if it was included in any of the class-specific or not-class LDAK GRMs.

Statistical analyses

Heritability analyses

Mixed linear models[22], implemented in the genome-wide complex trait analysis (GCTA) software package (version 1.91.7)[21], were applied to compare three models including a variable number of covariates. **Supplementary Table 12** gives the three different models, full descriptions of the covariates and model comparison have been given in **Supplementary Note 4**. The most parsimonious model was chosen for further analyses (full results in **Supplementary Table 13**). This final model included the first

10 genetic PCs for the Dutch population, genotyping chip, sex and age at blood draw as covariates. For metabolites of the Nightingale Health ^{1}H-NMR and Biocrates platform, measurement batch was included as covariate.

The final four-variance component model, including four GRMs, allows for the estimation of the proportion of variation explained by superclass-specific significant metabolite loci and non-superclass significant metabolite loci. The first two variance components in the 4-variance component model (**Figure 1**), *V(G1)* and *V(G2)* allow for the estimation of the additive genetic variance effects captured by genome-wide SNPs (h^2_g) and the additive genetic effects associated with pedigree (h^2_{ped})[22,74], and *V(G3)* and *V(G4)* capture the additive genetic effect associated with class-specific ($h^2_{class\text{-}hits}$) and not-class ($h^2_{notclass\text{-}hits}$) metabolite loci. Based on the 4-variance component model, three additional heritability estimates can be calculated: the total variance explained by significant metabolite loci ($h^2_{Metabolite\text{-}hits}$) consists of the sum of $\frac{V(G3)}{Vp}$ and $\frac{V(G4)}{Vp}$, where *Vp* is the phenotypic variance, h^2_{SNP} is defined as the sum of $\frac{V(G1)}{Vp}$, $\frac{V(G3)}{Vp}$ and $\frac{V(G4)}{Vp}$, and the total variance explained (h^2_{total}) is defined as the sum of $\frac{V(G1)}{Vp}$, $\frac{V(G2)}{Vp}$, $\frac{V(G3)}{Vp}$ and $\frac{V(G4)}{Vp}$ (**Figure 1**). We note that the total variance explained by genetic factors may also include influences of the shared environment, dominance and epistasis, which may result in upward bias of the h^2_{total} estimates[22,23]. This bias is expected to arise by the presence of closely related participants, who may share these effects, in addition to the additive genetic effects. To calculate the standard errors (s.e.'s) for the composite variance estimates, we have randomly sampled 10,000 new variances from the parameter variance-covariance matrices of the *V(G1), V(G3)* and *V(G4)* GRMs for each metabolite. Random sampling was performed in R by creating 10,000 multivariate normal distributions (mvrnorm function in MASS package version 7.3-50[75]) based on the original means and variance/covariance matrices. The s.e.'s of the specific ratio of interest were then based on the standard deviation of the ratio of interest across 10,000 samples. The four-variance component models included variance components that were not constrained to be positive, thus allowing for negative h^2_{SNP} and $h^2_{Metabolite\text{-}hits}$ estimates. All four-variance component models applied the --reml-bendV flag where necessary to invert the variance-covariance matrix *V* if *V* was not positive definite, which may occur when variance components are negative[76]. Finally, we calculated the log likelihood of a reduced model with either *V(G3)*, *V(G4)* or both dropped from the full model and calculated the LRT and p-value (**Supplementary Table 3**).

Mixed-effect meta-regression analyses

To investigate differences in heritability estimates among metabolites of different classes we applied mixed-effect meta-regression models as implemented in the 'metafor' package (version 2.0-0) in R (version 3.5.1)[77]. Here we tested for the moderation of heritability estimates by metabolite class and metabolomics platform on all 361 successfully analyzed metabolites. We included a matrix combining the phenotypic correlations (**Supplementary Table 14**) and the sample overlap (**Supplementary Table 11**) between the metabolites as random factor to correct for dependence among the metabolites and participants. This matrix includes the sample size of the metabolite on the diagonal, with the off-diagonal computed by $\frac{N_{1,2}}{\sqrt{n_1 * n_2}} * r$ (**Supplementary Table 15**), where $N_{1,2}$ is the sample overlap between the metabolites, n_1 is the sample size of metabolite one, n_2 is the sample size of metabolite two and *r* is the phenotypic (Spearman's rho) correlation between the metabolites. In all mixed-effect meta-regression analyses we obtained the robust estimates based on a sandwich-type estimator, clustered by the metabolites included in the models to correct for the sample overlap among the different metabolites[78]. First, we used multivariate mixed-effect meta-regression models to simultaneously estimate the effect of metabolite class and metabolomics platform on the h^2_{total}, h^2_{SNP} and the $h^2_{Metabolite\text{-}hits}$, as well as the $h^2_{Class\text{-}hits}$ and $h^2_{Notclass\text{-}hits}$ estimates. Subsequently, to separately assess the effect of the number of carbon atoms or double bonds in the fatty acyls chains of phosphatidylcholines and triglycerides univariate models were fitted, as follow-up. To account for multiple testing the p-values were adjusted with the with the False Discovery Rate (FDR)[79] using the 'p.adjust' function in R. Multiple testing correction was done separately for the univariate and the multivariate models.

Data availability

The curated list of all published metabolite-SNP associations is included in **Supplementary Data 1** and is publicly available through the BBMRI – omics atlas (http://bbmri.researchlumc.nl/atlas/#data). All information on the metabolites in this study are in **Supplementary Table 1**; with full summary statistics for the four-variance component models included in **Supplementary Table 3**. The Nightingale Health metabolomics data may be requested through BBMRI-NL (https://www.bbmri.nl/Omics-metabolomics). All (other) data may be accessed, upon approval of the data access committee, through the Netherlands Twin Register (ntr.fgb@vu.nl). A reporting summary for this Article is available as Supplementary Information file.

Funding

This work was performed within the framework of the BBMRI Metabolomics Consortium funded by BBMRI-NL, a research infrastructure financed by the Dutch government (NWO, no. 184.021.007 and 184.033.111). The European Network of Genomic and Genetic Epidemiology (ENGAGE) contributed to funding to perform the Biocrates Absolute-IDQ^{TM} p150 metabolomics measurements (European Union Seventh Framework Program: FP7/2007-2013, grant number 201413). Analyses were supported by the Netherlands Organization for Scientific Research: Netherlands Twin Registry Repository: researching the interplay between genome and environment (480-15-001/674); the European Union Seventh Framework Program (FP7/2007-2013): ACTION Consortium (Aggression in Children: Unravelling gene-environment interplay to inform Treatment and InterventiON strategies; grant number 602768).

Genotyping was made possible by grants from NWO/SPI 56-464-14192, Genetic Association Information Network (GAIN) of the Foundation for the National Institutes of Health, Rutgers University Cell and DNA Repository (NIMH U24 MH068457-06), the Avera Institute, Sioux Falls (USA) and the National Institutes of Health (NIH R01 HD042157-01A1, MH081802, Grand Opportunity grants 1RC2 MH089951 and 1RC2 MH089995) and European Research Council (ERC-230374). EMIF-AD has received support from the EU/EFPIA Innovative Medicines Initiative Joint Undertaking EMIF grant agreement n°115372. DIB acknowledges her KNAW Academy Professor Award (PAH/6635). M.Bartels is supported by an ERC consolidator grant (WELL-BEING 771057 PI Bartels).

Acknowledgements

We thank all twins and family members for their participation. We thank P. M. Visscher (University of Queensland) for his helpful comments and C. V. Dolan (Vrije Universiteit Amsterdam) for critically reading and commenting on the final version of the manuscript. Preliminary analyses of this paper were included in a presentation at the 46th Annual Meeting of the Behavioral Genetics Association (BGA) in June 2016 (abstract in Behav. Genet. (2016) 46:785-786), and a presentation at the 49th Annual Meeting of the BGA in June 2019 (abstract forthcoming).

Author contributions

Nightingale Health metabolomics data: HES, MBeekman, PES and CMvD. Leiden ^{1}H-NMR metabolomics data: KWvD and AV. UPLC-MS lipidomics data: ACH and TH. EMIF-AD data: AdB and PJV. Genotype data: JJH, AA and IOF. NTR Biobank data: GW and EJCdG. Metabolomics pre-processing: RP, HHMD and FAH. Statistical analyses: FAH and MGN. Wrote the paper: FAH, JvD, MBartels, MGN and DIB. All authors critically read and commented on the manuscript.

Competing interests statement

The authors declare no competing financial interests.

BBMRI Metabolomics Consortium

Cohort Collection and Sample Management Group:

M. Beekman[1], H.E.D. Suchiman[1], N. Amin[2], J.W. Beulens[3,4], J.A. van der Bom[5-8], N. Bomer[9], A. Demirkan[2], J.A. van Hilten[10], J.M.T.A. Meessen[11], R. Pool[12], M.H. Moed[1], J. Fu[13,14], G.L.J. Onderwater[15], F. Rutters[3], C. So-Osman[10], W.M. van der Flier[3,16], A.A.W.A. van der Heijden[17], A. van der Spek[2], F.W. Asselbergs[18], E. Boersma[19], P.M. Elders[20,21], J.M. Geleijnse[22], M.A. Ikram[2,23,24], M. Kloppenburg[8,25], I. Meulenbelt[1], S.P. Mooijaart[26], R.G.H.H. Nelissen[27], M.G. Netea[28,29], B.W.J.H. Penninx[21,30], C.D.A. Stehouwer[31,32], C.E. Teunissen[33], G.M. Terwindt[15], L.M. 't Hart[1,3,21,34,35], A.M.J.M. van den Maagdenberg[36], P. van der Harst[8], I.C.C. van der Horst[37], C.J.H. van der Kallen[31,32], M.M.J. van Greevenbroek[31,32], W.E. van Spil[38], C. Wijmenga[13], A.H. Zwinderman[39], A. Zhernikova[13], J.W. Jukema[40]

Database & Catalogue: J.J.H. Barkey Wolf[1], M. Beekman[1], D. Cats[1], H. Mei[1,41], M. Slofstra[13], M. Swertz[13]

Quality Control: E.B. van den Akker[1,42,43], J.J.H. Barkey Wolf[1], J. Deelen[1,44], M.J.T. Reinders[42,43]

Steering Committee: D.I. Boomsma[21,45], C.M. van Duijn[2], P.E. Slagboom[1]

Affiliations:

[1] Department of Molecular Epidemiology, Leiden University Medical Center, Leiden, The Netherlands.

[2] Department of Epidemiology, Erasmus MC University Medical Center, Rotterdam, The Netherlands.

[3] Department of Epidemiology and Biostatistics, Amsterdam University Medical Center, Amsterdam, the Netherlands.

[4]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands.

[5]Centre for Clinical Transfusion Research, Sanquin Research, Leiden, The Netherlands.

[6]Jon J van Rood Centre for Clinical Transfusion Research, Leiden University Medical Centre, Leiden, The Netherlands.

[7]TIAS, Tilburg University, Tilburg, The Netherlands.

[8]Department of Clinical Epidemiology, Leiden University Medical Centre, Leiden, The Netherlands.

[9]Department of Cardiology, University Medical Center Groningen, University of Groningen, Groningen, the Netherlands.

[10]Center for Clinical Transfusion Research, Sanquin Research, Leiden, the Netherlands.

[11]Department of Orthopedics, Leiden University Medical Centre, Leiden, The Netherlands.

[12]Department of Biological Psychology, Vrije Universiteit, Amsterdam, the Netherlands.

[13]Department of Genetics, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands

[14]Department of Pediatrics, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands

[15]Department of Neurology, Leiden University Medical Center, Leiden, the Netherlands.

[16]Department of Neurology and Alzheimer Center, Neuroscience Campus Amsterdam, VU University Medical Center, Amsterdam, The Netherlands.

[17]Department of General Practice, The EMGO Institute for Health and Care Research, VU University Medical Center, Amsterdam, The Netherlands.

[18]Department of Cardiology, Division Heart and Lungs, University Medical Center Utrecht, Utrecht, The Netherlands Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands.

[19]Thorax centre, Erasmus Medical Centre, Rotterdam, the Netherlands.

[20]Department of General Practice and Elderly Care Medicine, VU University Medical Center, Amsterdam, The Netherlands.

[21]Amsterdam Public Health research institute, VU University Medical Center, Amsterdam, The Netherlands.

[22]Division of Human Nutrition and Health, Wageningen University, Wageningen, The Netherlands.

[23]Department of Radiology, Erasmus University Medical Center Rotterdam, Rotterdam, The Netherlands.

[24]Department of Neurology, Erasmus University Medical Center Rotterdam, Rotterdam, The Netherlands.

[25]Department of Rheumatology, Leiden University Medical Center, The Netherlands.

[26]Department of Internal Medicine, Division of Gerontology and Geriatrics, Leiden University Medical Centre, Leiden, The Netherlands.

[27]Department of Orthopaedics, Leiden University Medical Center, Leiden, The Netherlands.

[28]Department of Internal Medicine, Radboud Center for Infectious Diseases, Radboud University Medical Center, Nijmegen, Netherlands.

[29]Department for Genomics & Immunoregulation, Life and Medical Sciences Institute (LIMES), University of Bonn, Bonn, Germany.

[30]Department of Psychiatry, VU University Medical Center, Amsterdam, The Netherlands.

[31]Department of Internal Medicine, Maastricht University Medical Center (MUMC+), Maastricht, the Netherlands.

[32]School for Cardiovascular Diseases (CARIM), Maastricht University, Maastricht, the Netherlands.

[33]Neurochemistry Laboratory, Clinical Chemistry Department, Amsterdam University Medical Center, Amsterdam Neuroscience, the Netherlands.

[34]Department of Cell and Chemical Biology, Leiden University Medical Center, Leiden, the Netherlands.

[35]Department of General practice, Amsterdam University Medical Center, Amsterdam, the Netherlands.

[36]Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands.

[37]Department of Critical Care, University Medical Center Groningen, Groningen, The Netherlands.

[38]UMC Utrecht, Department of Rheumatology & Clinical Immunology, Utrecht, The Netherlands.

[39]Department of Clinical Epidemiology, Biostatistics, and Bioinformatics, Academic Medical Centre, University of Amsterdam, Amsterdam, the Netherlands.

[40]Department of Cardiology, Leiden University Medical Center, Leiden, The Netherlands.

[41]Sequencing Analysis Support Core, Leiden University Medical Center, Leiden, The Netherlands

[42]Leiden Computational Biology Center, Leiden University Medical Center, Leiden, The Netherlands.

[43]Department of Pattern Recognition and Bioinformatics, Delft University of Technology, Delft, The Netherlands.

[44]Max Planck Institute for Biology of Ageing, Cologne, Germany.

[45]Netherlands Twin Register, Department of Biological Psychology, Vrije Universiteit, Amsterdam, The Netherlands.

References

1. Patti, G. J., Yanes, O. & Siuzdak, G. Innovation: Metabolomics: the apogee of the omics trilogy. *Nat. Rev. Mol. Cell Biol.* **13**, 263–269 (2012).

2. Wishart, D. S. *et al.* HMDB: The human metabolome database. *Nucleic Acids Res.* **35**, 521–526 (2007).

3. Wishart, D. S. *et al.* HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Res.* **37**, D603-10 (2009).

4. Wishart, D. S. *et al.* HMDB 3.0-The Human Metabolome Database in 2013. *Nucleic Acids Res.* **41**, 801–807 (2013).

5. Wishart, D. S. *et al.* HMDB 4.0: The human metabolome database for 2018. *Nucleic Acids Res.* **46**, D608–D617 (2018).

6. Kuehnbaum, N. L. & Britz-McKibbin, P. New advances in separation science for metabolomics: resolving chemical diversity in a post-genomic era. *Chem. Rev.* **113**, 2437–68 (2013).

7. Mittelstrass, K. *et al.* Discovery of sexual dimorphisms in metabolic and genetic biomarkers. *PLoS Genet.* **7**, e1002215 (2011).

8. Chaleckis, R., Murakami, I., Takada, J., Kondoh, H. & Yanagida, M. Individual variability in human blood metabolites identifies age-related differences. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 4252–4259 (2016).

9. Menni, C. *et al.* Targeted metabolomics profiles are strongly correlated with nutritional patterns in women. *Metabolomics* **9**, 506–514 (2013).

10. Kastenmüller, G., Raffler, J., Gieger, C. & Suhre, K. Genetics of human metabolism: an update. *Hum. Mol. Genet.* **24**, R93–R101 (2015).

11. Gieger, C. *et al.* Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. *PLoS Genet.* **4**, e1000282 (2008).

12. Nicholson, G. *et al.* Human metabolic profiles are stably controlled by genetic and environmental variation. *Mol. Syst. Biol.* **7**, 525 (2011).

13. Shah, S. H. *et al.* High heritability of metabolomic profiles in families burdened with premature cardiovascular disease. *Mol. Syst. Biol.* **5**, 258 (2009).

14. Draisma, H. H. M. *et al.* Familial resemblance for serum metabolite concentrations. *Twin Res. Hum. Genet.* **16**, 948–61 (2013).

15. Rhee, E. P. *et al.* A genome-wide association study of the human metabolome in a community-based cohort. *Cell Metab.* **18**, 130–143 (2013).

16. Frahnow, T. *et al.* Heritability and responses to high fat diet of plasma lipidomics in a twin study. *Sci. Rep.* **7**, 1–11 (2017).

17. Kaess, B. *et al.* The lipoprotein subfraction profile: heritability and identification of quantitative trait loci. *J. Lipid Res.* **49**, 715–723 (2008).

18. Bellis, C. *et al.* Human Plasma Lipidome Is Pleiotropically Associated With Cardiovascular Risk Factors and Death. *Circ. Cardiovasc. Genet.* **7**, 854–863 (2014).

19. Draisma, H. H. M. Analysis of Metabolomics Data from Twin Families. (Leiden, 2011).

20. Reeds, P. J. Dispensable and Indispensable Amino Acids for Humans. *J. Nutr.* **130**, 1874S-1876S

(2000).

21. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: A tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).

22. Zaitlen, N. *et al.* Using Extended Genealogy to Estimate Components of Heritability for 23 Quantitative and Dichotomous Traits. *PLoS Genet.* **9**, (2013).

23. Zuk, O., Hechter, E., Sunyaev, S. R. & Lander, E. S. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 1193–1198 (2012).

24. Newgard, C. B. Metabolomics and Metabolic Diseases: Where Do We Stand? *Cell Metab.* **25**, 43–56 (2017).

25. Onderwater, G. L. J. *et al.* Large-scale plasma metabolome analysis reveals alterations in HDL metabolism in migraine. *Neurology* **0**, 10.1212/WNL.0000000000007313 (2019).

26. Nedic Erjavec, G. *et al.* Short overview on metabolomic approach and redox changes in psychiatric disorders. *Redox Biol.* **14**, 178–186 (2018).

27. van der Lee, S. J. *et al.* Circulating metabolites and general cognitive ability and dementia: Evidence from 11 cohort studies. *Alzheimer's Dement.* 1–16 (2018). doi:10.1016/j.jalz.2017.11.012

28. Willemsen, G. *et al.* The Netherlands Twin Register biobank: a resource for genetic epidemiological studies. *Twin Res. Hum. Genet.* **13**, 231–45 (2010).

29. Speed, D., Hemani, G., Johnson, M. R. & Balding, D. J. Improved heritability estimation from genome-wide SNPs. *Am. J. Hum. Genet.* **91**, 1011–1021 (2012).

30. Speed, D., Cai, N., Johnson, M. R., Nejentsev, S. & Balding, D. J. Reevaluation of SNP heritability in complex human traits. *Nat. Genet.* (2017). doi:10.1038/ng.3865

31. Yet, I. *et al.* Genetic influences on metabolite levels: A comparison across metabolomic platforms. *PLoS One* **11**, (2016).

32. Tabassum, R. *et al.* Genetics of human plasma lipidome: Understanding lipid metabolism and its link to diseases beyond traditional lipids. *bioRxiv* (2018). doi:10.1101/457960

33. Yousri, N. A. *et al.* Whole-exome sequencing identifies common and rare variant metabolic QTLs in a Middle Eastern population. *Nat. Commun.* **9**, 1–13 (2018).

34. Kettunen, J. *et al.* Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nat. Genet.* **44**, 269–276 (2012).

35. Tsepilov, Y. A. *et al.* Nonadditive effects of genes in human metabolomics. *Genetics* **200**, 707–718 (2015).

36. Tukiainen, T. *et al.* Detailed metabolic and genetic characterization reveals new associations for 30 known lipid loci. *Hum. Mol. Genet.* **21**, 1444–55 (2012).

37. Yet, I. *et al.* Genetic influences on metabolite levels: A comparison across metabolomic platforms. *PLoS One* **11**, (2016).

38. Tremblay, B. L., Guénard, F., Lamarche, B., Pérusse, L. & Vohl, M. C. Familial resemblances in human plasma metabolites are attributable to both genetic and common environmental effects. *Nutr. Res.* **61**, 22–30 (2019).

39. Gallois, A. *et al.* A comprehensive study of metabolite genetics reveals strong pleiotropy and heterogeneity across time and context. *bioRxiv* (2018). doi:http://dx.doi.org/10.1101/461848

40. Wittemans, L. B. L. *et al.* Assessing the causal association of glycine with risk of cardio-metabolic diseases. *Nat. Commun.* **10**, 1–13 (2019).

41. Demirkan, A. *et al.* Genome-wide association study of plasma lipids. *bioRxiv* (2019). doi:http://dx.doi.org/10.1101/621334

42. Sudlow, C. *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Med.* **12**, 1–10 (2015).

43. Yang, J. *et al.* Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat. Genet.* **47**, 1114–1120 (2015).

44. Wainschtein, P. *et al.* Recovery of trait heritability from whole genome sequence data. *bioRxiv* (2019). doi:http://dx.doi.org/10.1101/588020

45. Boomsma, D. I. *et al.* Netherlands Twin Register: from twins to twin families. *Twin Res. Hum. Genet.* **9**, 849–57 (2006).

46. Willemsen, G. *et al.* The Adult Netherlands Twin Register: twenty-five years of survey and biological data collection. *Twin Res. Hum. Genet.* **16**, 271–81 (2013).

47. Soininen, P., Kangas, A. J., Würtz, P., Suna, T. & Ala-Korpela, M. Quantitative Serum Nuclear Magnetic Resonance Metabolomics in Cardiovascular Epidemiology and Genetics. *Circ. Cardiovasc. Genet.* **8**, 192–206 (2015).

48. Würtz, P. *et al.* Quantitative Serum Nuclear Magnetic Resonance Metabolomics in Large-Scale Epidemiology: A Primer on -Omic Technology. *Am. J. Epidemiol.* **186**, 1–13 (2017).

49. Gonzalez-Covarrubias, V. *et al.* Lipidomics of familial longevity. *Aging Cell* **12**, 426–434 (2013).

50. Dane, A. D. *et al.* Integrating metabolomics profiling measurements across multiple biobanks. *Anal. Chem.* **86**, 4110–4114 (2014).

51. Demirkan, A. *et al.* Insight in Genome-Wide Association of Metabolite Quantitative Traits by Exome Sequence Analyses. *PLoS Genet.* **11**, e1004835 (2015).

52. Goek, O. N. *et al.* Serum metabolite concentrations and decreased GFR in the general population. *Am. J. Kidney Dis.* **60**, 197–206 (2012).

53. Römisch-Margl, W. *et al.* Procedure for tissue sample preparation and metabolite extraction for high-throughput targeted metabolomics. *Metabolomics* **8**, 133–142 (2012).

54. Buuren, S. van & Groothuis-Oudshoorn, K. mice: Multivariate Imputation by Chained Equations in R. *J. Stat. Softw.* **45**, (2011).

55. Kettunen, J. *et al.* Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. *Nat. Commun.* **7**, 11122 (2016).

56. Draisma, H. H. M. *et al.* Genome-wide association study identifies novel genetic variants contributing to variation in blood metabolite levels. *Nat. Commun.* **6**, 7208 (2015).

57. Boomsma, D. I. *et al.* The Genome of the Netherlands: design, and project goals. *Eur. J. Hum. Genet.* **22**, 221–227 (2014).

58. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

59. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).

60. Fedko, I. O. *et al.* Estimation of Genetic Relationships Between Individuals Across Cohorts and Platforms: Application to Childhood Height. *Behav. Genet.* **45**, 514–528 (2015).

61. Deelen, P. *et al.* Improved imputation quality of low-frequency and rare variants in European samples using the 'Genome of the Netherlands'. *Eur. J. Hum. Genet.* **22**, 1321–1326 (2014).

62. Liu, E. Y., Li, M., Wang, W. & Li, Y. MaCH-Admix: Genotype Imputation for Admixed Populations. *Genet. Epidemiol.* **37**, 25–37 (2013).

63. Loh, P. R., Palamara, P. F. & Price, A. L. Fast and accurate long-range phasing in a UK Biobank cohort. *Nat. Genet.* **48**, 811–816 (2016).

64. Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9**, 179–81 (2012).

65. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

66. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).

67. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).

68. Abdellaoui, A. *et al.* Population structure, migration, and diversifying selection in the Netherlands. *Eur. J. Hum. Genet.* **21**, 1277–1285 (2013).

69. Kim, S. *et al.* PubChem 2019 update: Improved access to chemical data. *Nucleic Acids Res.* **47**, D1102–D1109 (2019).

70. Hastings, J. *et al.* ChEBI in 2016: Improved services and an expanding collection of metabolites.

Nucleic Acids Res. **44**, D1214–D1219 (2016).

71. Heller, S. R., McNaught, A., Pletnev, I., Stein, S. & Tchekhovskoi, D. InChI, the IUPAC International Chemical Identifier. *J. Cheminform.* **7**, 1–34 (2015).

72. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–11 (2001).

73. Haeussler, M. *et al.* The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res.* **47**, D853–D858 (2019).

74. Xia, C. *et al.* Pedigree- and SNP-Associated Genetics and Recent Environment are the Major Contributors to Anthropometric and Cardiometabolic Trait Variation. *PLoS Genet.* **12**, 1–25 (2016).

75. Venables, W. N. & Ripley, B. D. *Modern applied statistics with S*. (Springer, 2002).

76. Hayes, J. F. & Hill, W. G. Modification of Estimates of Parameters in the Construction of Genetic Selection Indices (' Bending '). *Biometrics* **37**, 483–493 (1981).

77. Viechtbauer, W. Conducting Meta-Analyses in R with the metafor Package. *J. Stat. Softw.* **36**, 1–48 (2010).

78. Hedges, L. V., Tipton, E. & Johnson, M. C. Robust variance estimation in meta-regression with dependent effect size estimates. *Res. Synth. Methods* **1**, 39–65 (2010).

79. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B* **57**, 289–300 (1995).

Figures

Figure 1. Overview of the 4-variance component models, including the GRMs underlying each variant component and all heritability estimates obtained from the models.

Overview of the SNP-filtering and GRM construction can be found in **Supplementary Figure 1** and is explained in details in the **Methods**. This figure describes which GRMs (black boxes) are used to calculate which variance components (orange boxes) by drawing black arrows from the GRMs to the variance components. The variance components give rise to the four different heritability estimates: h^2_{ped}, h^2_{g}, $h^2_{Class\text{-}hits}$, and $h^2_{Notclass\text{-}hits}$ (see **Methods**). The orange arrows indicate how the various variance components are summed to obtain estimates for $h^2_{metabolite\text{-}hits}$, h^2_{SNP} and h^2_{total} (see **Methods**).

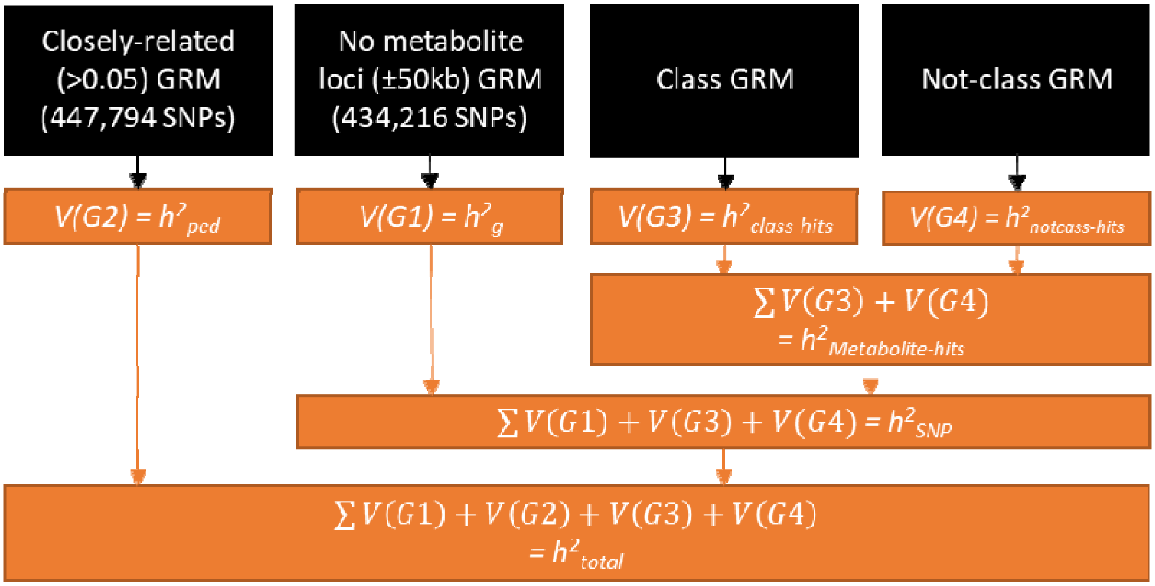

Closely-related (>0.05) GRM (447,794 SNPs)
No metabolite loci (±50kb) GRM (434,216 SNPs)
Class GRM
Not-class GRM
$V(G2) = h^2_{ped}$
$V(G1) = h^2_{g}$
$V(G3) = h^2_{class\ hits}$
$V(G4) = h^2_{notcass\text{-}hits}$
$\sum V(G3) + V(G4) = h^2_{Metabolite\text{-}hits}$
$\sum V(G1) + V(G3) + V(G4) = h^2_{SNP}$
$\sum V(G1) + V(G2) + V(G3) + V(G4) = h^2_{total}$

Figure 2. Heritability of all 52 carboxylic acids and derivatives successfully analyzed across all four metabolomics platforms by class.

Box- and dotplots of the h^2_{total}, and $h^2_{Metabolite\text{-}hits}$ for all 52 successfully analyzed ‘carboxylic acids and derivatives’ by class. The left-hand side of the figure is a close-up of the -0.08 – 0.15 part of the heritability range, focusing on the $h^2_{Class\text{-}hits}$ and $h^2_{Notclass\text{-}hits}$ estimates. The boxes denote the 25th and 75th percentile (bottom and top of box), and median value (horizontal band inside box). The whiskers indicate the values observed within up to 1.5 times the interquartile range above and below the box. **Supplementary Table 3** provides the estimates for each of the individual metabolites.

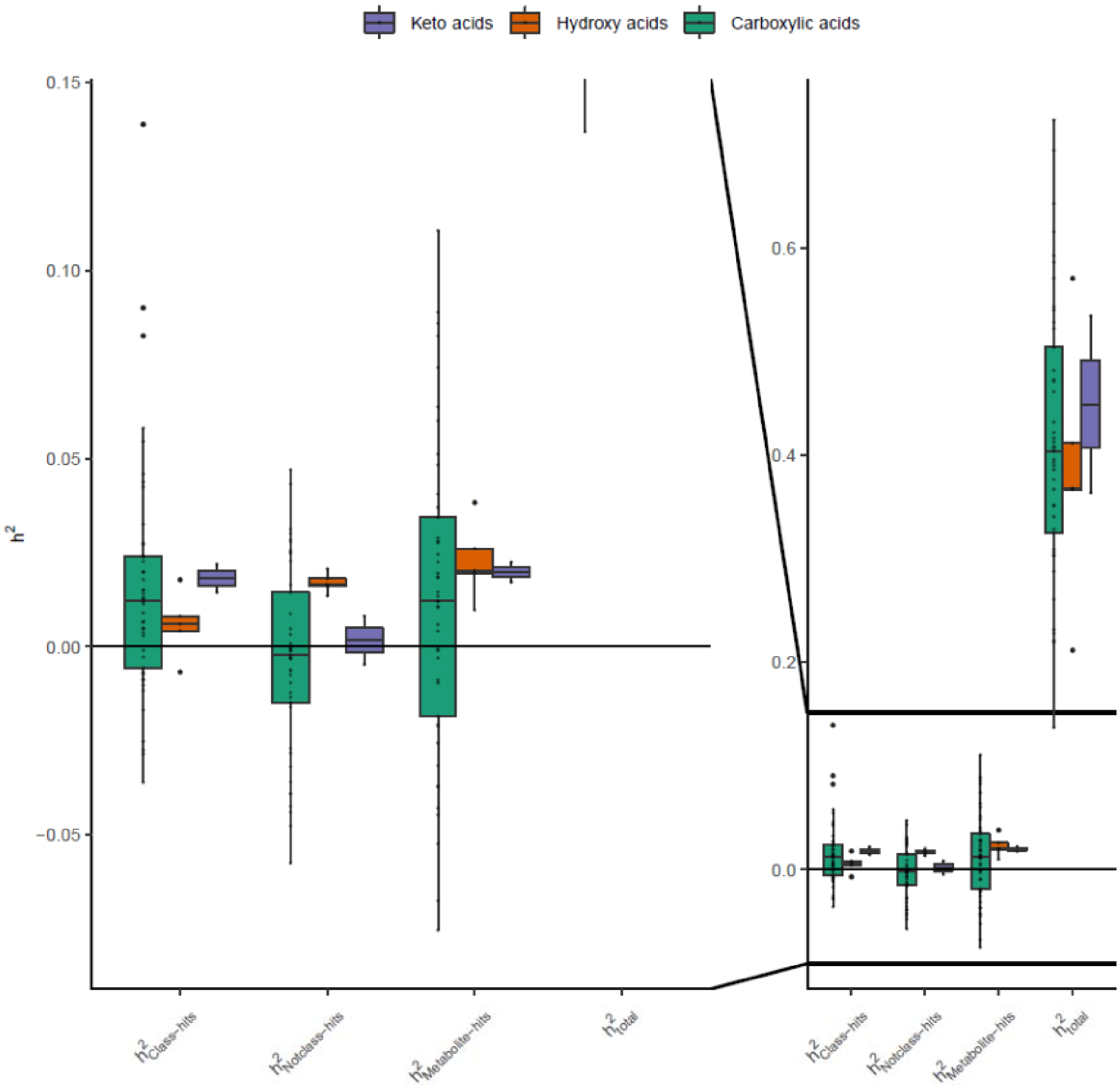
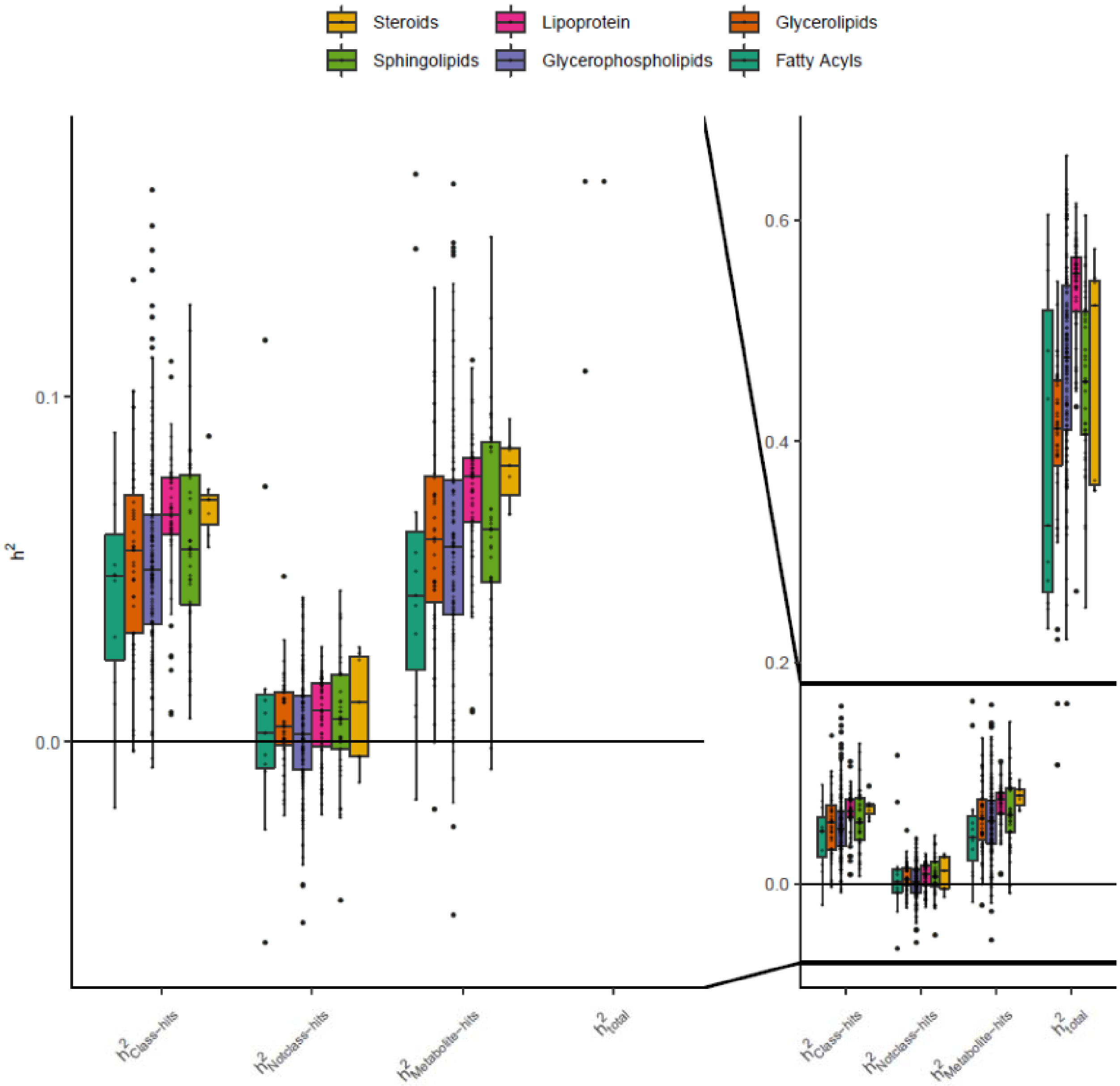

Keto acids
Hydroxy acids
Carboxylic acids
h²
0.15
0.10
0.05
0.00
−0.05
0.6
0.4
0.2
0.0
h²Class-hits
h²Notclass-hits
h²Metabolite-hits
h²Total
h²Class-hits
h²Notclass-hits
h²Metabolite-hits
h²Total

Figure 3. Heritability of all 309 lipids successfully analyzed across all four metabolomics platforms by class.

Box- and dotplots of the h^2_{total}, and $h^2_{Metabolite\text{-}hits}$ for all 309 successfully analyzed lipids by class. The left-hand side of the figure is a close-up of the -0.06 – 0.17 part of the heritability range, focusing on the $h^2_{Class\text{-}hits}$ and $h^2_{Notclass\text{-}hits}$ estimates. The boxes denote the 25th and 75th percentile (bottom and top of box), and median value (horizontal band inside box). The whiskers indicate the values observed within up to 1.5 times the interquartile range above and below the box. **Supplementary Table 3** provides the estimates for each of the individual metabolites.


Steroids
Lipoprotein
Glycerolipids
Sphingolipids
Glycerophospholipids
Fatty Acyls
h^2
0.1
0.0
0.6
0.4
0.2
0.0
$h^2_{Class-hits}$
$h^2_{Notclass-hits}$
$h^2_{Metabolite-hits}$
h^2_{total}
$h^2_{Class-hits}$
$h^2_{Notclass-hits}$
$h^2_{Metabolite-hits}$
h^2_{total}

Tables

Table 1. Overview of the number of unique metabolites, for which significant SNP-metabolite associations have been published, per Human Metabolome Database[3–5] ‘super class’.

See **Supplementary Data 1** for an overview of the exact metabolites classified per ‘super class’, ‘class’ and ‘subclass’, as well as the SNPs associated with each metabolite.

Super class	**Number of unique metabolites**
Lipids and lipid-like molecules (e.g., lipids)	662
Organic acids and derivatives (e.g., organic acids)	182
Organoheterocyclic compounds	45
Organic oxygen compounds	19
Nucleosides, nucleotides, and analogues	12
Benzenoids	12
Organic nitrogen compounds	11
Phenylpropanoids and polyketides	4
Proteins	3
Organic compounds	1
Trichlorophenols	1
Organooxygen compounds	1

Table 2. Participant characteristics after preprocessing per metabolomics platform.

This table gives an overview of the number of individuals (N) per platform, specifies the number of families these individuals belong to and the percentage of females and twins in each dataset. In addition, for each platform the mean and standard deviation (SD) of the age at blood draw in years, the body-mass-index (BMI), the cholesterol level in mmol/l, the low-density lipoprotein cholesterol (LDL) levels in mmol/l and the high-density lipoprotein cholesterol (HDL) levels in mmol/l are given.

Metabolomics platform	**N**	**N families**	**Age* (mean ± SD)**	**Female (%)**	**Twins (%)**	**BMI (mean ± SD)**	**Cholesterol$^{\$}$ (mean ± SD)**	**LDL$^{\$}$ (mean ± SD)**	**HDL$^{\$}$ (mean ± SD)**
All Participants	5,117	2,445	42.1 ± 14.2	62.8%	63.4%	24.8 ± 4.1	4.9 ± 1.2	3.0 ± 1.0	1.7 ± 1.0
Nightingale Health ^{1}H-NMR	4,227	2,179	40.7 ± 13.7	67.3%	69.7%	24.6 ± 4.0	4.9 ± 1.2	3.0 ± 1.0	1.7 ± 1.0
UPLC-MS Lipidomics	2,324	1,251	39.0 ± 12.9	66.6%	89.2%	24.4 ± 4.1	5.0 ± 1.0	3.0 ± 0.9	1.4 ± 0.4
Leiden ^{1}H-NMR	2,324	1,323	37.6 ± 12.5	67.0%	89.0%	24.2 ± 4.1	4.6 ± 1.3	2.7 ± 1.0	2.0 ± 1.4
Biocrates	1,448	946	45.7 ± 15.3	43.8%	39.6%	25.2 ± 3.9	4.6 ± 1.5	2.8 ± 1.1	2.3 ± 1.7

* Age at blood draw in years; $^{\$}$ levels in mmol/l.

Table 3. Summary of the heritability estimates of the four-variance component models for the 309 lipids and the 52 organic acids analyzed across all four metabolomics platforms.

The mean, median and range of the total heritability (h^2_{total}), heritability based on the 479 significant metabolite loci for the lipids or the 397 significant metabolite loci for the organic acids ($h^2_{Class\text{-}hits}$), the 596-683 significant metabolite loci not belonging to these classes ($h^2_{Notclass\text{-}hits}$) and the total heritability explained by metabolite loci (e.g., sum of $h^2_{Class\text{-}hits}$ and $h^2_{Notclass\text{-}hits}$: $h^2_{Metabolite\text{-}hits}$), as well as their standard errors (s.e.'s), are depicted for all 361 successfully analyzed metabolites as included on all platforms. **Supplementary Table 1** denotes which metabolites belong to each class and **Supplementary Table 3** provides the estimates for each of the individual metabolites.

		Lipids and lipid-like molecules		**Organic acids and derivatives**	
		estimate	**s.e.**	**estimate**	**s.e.**
$\mathbf{h^2_{total}}$	*mean*	0.47	0.04	0.41	0.04
	median	0.47	0.03	0.40	0.03
	range	(0.11 - 0.66)	(0.02 - 0.07)	(0.14 - 0.72)	(0.02 - 0.07)
$\mathbf{h^2_{Metabolite\text{-}hits}}$	*mean*	0.06	0.03	0.01	0.02
	median	0.06	0.03	0.02	0.02
	range	(-0.05 - 0.16)	(0.01 - 0.04)	(-0.08 - 0.11)	(0.01 - 0.04)
$\mathbf{h^2_{Class\text{-}hits}}$	*mean*	0.06	0.02	0.01	0.02
	median	0.06	0.02	0.01	0.02
	range	(-0.02 - 0.16)	(0.01 - 0.03)	(-0.04 - 0.14)	(0.01 - 0.03)
$\mathbf{h^2_{Notclass\text{-}hits}}$	*mean*	0.00	0.02	0.00	0.02
	median	0.01	0.02	0.00	0.02
	range	(-0.06 - 0.12)	(0.01 - 0.03)	(-0.06 - 0.05)	(0.01 - 0.03)

Table 4. Summary of the heritability estimates of the four-variance component models for the 17 essential and the 14 non-essential amino acids analyzed across all four metabolomics platforms.

The mean, median and range of the total heritability (h^2_{total}), and heritability based on the 397 significant metabolite loci for the organic acids ($h^2_{Class\text{-}hits}$), the 683 significant metabolite loci not belonging to this class ($h^2_{Notclass\text{-}hits}$) and the total heritability explained by metabolite loci (e.g., sum of $h^2_{Class\text{-}hits}$ and $h^2_{Notclass\text{-}hits}$: $h^2_{Metabolite\text{-}hits}$), as well as their standard errors (s.e.'s), are depicted for all 31 successfully analyzed essential and non-essential amino acids as included on all platforms. **Supplementary Table 1** denotes which metabolites belong to each class and **Supplementary Table 3** provides the estimates for each of the individual metabolites.

		Essential amino acids		**Non-essential amino acids**	
		estimate	**s.e.**	**estimate**	**s.e.**
h^2_{total}	*mean*	0.42	0.04	0.39	0.04
	median	0.40	0.03	0.39	0.04
	range	(0.23 - 0.64)	(0.02 - 0.07)	(0.22 - 0.69)	(0.03 - 0.07)
$h^2_{Metabolite\text{-}hits}$	*mean*	0.00	0.02	0.02	0.03
	median	0.00	0.02	0.01	0.03
	range	(-0.05 - 0.05)	(0.01 - 0.03)	(-0.07 - 0.11)	(0.01 - 0.04)
$h^2_{Class\text{-}hits}$	*mean*	0.01	0.02	0.03	0.02
	median	0.00	0.02	0.01	0.02
	range	(-0.03 - 0.05)	(0.01 - 0.02)	(-0.03 - 0.14)	(0.01 - 0.03)
$h^2_{Notclass\text{-}hits}$	*mean*	-0.01	0.02	0.00	0.02
	median	-0.01	0.02	0.00	0.02
	range	(-0.06 - 0.04)	(0.01 - 0.03)	(-0.04 - 0.03)	(0.01 - 0.03)