Genome-wide association study identifies novel genetic variants contributing to variation in blood metabolite levels

Draisma, Harmen H. M.; Pool, Rene; Kobl, Michael; Jansen, Rick; Petersen, Ann-Kristin; Vaarhorst, Anika A. M.; Yet, Idil; Haller, Toomas; Demirkan, Ayse; Esko, Tonu

Published in:
Nature Communications

DOI:
10.1038/ncomms8208

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2015

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
Genome-wide association study identifies novel genetic variants contributing to variation in blood metabolite levels

Harmen H.M. Draisma1,2,3,*, René Pool1,2,4,*, Michael Kobi5, Rick Jansen5, Ann-Kristin Petersen5, Anika A.M. Vaarhors4,7, Idil Yet9, Toomas Haller10, Ayşe Demirkiran11,12, Tõnu Esko10,13,14,15, Gu Zhu16, Stefan Böhinger17, Marian Beekman17, Jan Bert van Klinken11, Werner Römisch-Margi18, Cornelia Prehn19, Jerzy Adamski20,21, Anton J.M. de Craen22, Elisabeth M. van Leeuwen12, Najaf Amin12, Harish Dharani11, Har-Jun Westra23, Lude Franke23, Eco J.C. de Geus1,2, Jouke Jan Hottenga1,2, Konstantin Strauch5,27, Nicholas G. Martin16, Gert-Jan B. van Ommen18, Cornelia M. van Duijn4,8,12,*** & Dorret I. Boomsma1,2,4,***

Metabolites are small molecules involved in cellular metabolism, which can be detected in biological samples using metabolomic techniques. Here we present the results of genome-wide association and meta-analyses for variation in the blood serum levels of 129 metabolites as measured by the Biocrates metabolomic platform. In a discovery sample of 7,478 individuals of European descent, we find 4,068 genome- and metabolome-wide significant (P < 1.09 × 10^-8) associations between single-nucleotide polymorphisms (SNPs) and metabolites, involving 59 independent SNPs and 85 metabolites. Five of the fifty-nine independent SNPs are new for serum metabolite levels, and were followed-up for replication in an independent sample (N = 1,182). The novel SNPs are located in or near genes encoding metabolite transporter proteins or enzymes (SLC22A16, ARG1, AGPS and ACSL1) that have demonstrated biomedical or pharmaceutical importance. The further characterization of genetic influences on metabolic phenotypes is important for progress in biological and medical research.

1 Department of Biological Psychology, VU University Amsterdam, van der Boechorststraat 1, Amsterdam 1081 BT, The Netherlands. 2 The EMM GO + Institute for Health and Care Research, VU University Medical Center, Van der Boechorststraat 7, Amsterdam 1081 BT, The Netherlands. 3 Neuroscience Campus Amsterdam, De Boelelaan 1085, Amsterdam 1081 HV, The Netherlands. 4 BBMRI-NL: Infrastructure for the Application of Metabolomics Technology in Epidemiology (RP4), S4-P, Postbus 9600, Leiden 2300 RC, The Netherlands. 5 Institute of Genetic Epidemiology, Helmholtz Zentrum München—German Research Center for Environmental Health, Ingolstädter Landstraße 1, Neufahrn 85764, Germany. 6 Department of Psychiatry, VU University Medical Center, Neuroscience Campus Amsterdam, VUmc, A.J. Ernststraat 187, Amsterdam 1081 H L, The Netherlands. 7 Department of Molecular Epidemiology, Leiden University Medical Center, PO Box 9600, Leiden 2300 RC, The Netherlands. 8 Netherlands Consortium for Healthy Aging, Leiden University Medical Center, Leiden, The Netherlands. 9 Department of Twin Research and Genetic Epidemiology, Institute of Health and Biomedical Innovation, Queensland University of Technology, 60 Musk Ave, Kelvin Grove QLD 4059, Queensland, Australia. 10 Estonian Genome Center, University of Tartu, 23b Riiu street, Tartu 51010, Estonia. 11 Department of Human Genetics, Leiden University Medical Center, S4-P, PO Box 9600, Leiden 2300 RC, The Netherlands. 12 Genetic Epidemiology Unit, Department of Epidemiology, Erasmus Medical Center, P.O. Box 2040, Rotterdam 3000 CA, The Netherlands. 13 Divisions of European Science and Genetics and Center for Basic and Translational Obesity Research, Boston Children’s Hospital, 300 Longwood Ave, Boston MA02115, Massachusetts, USA. 14 Medical and Population Genetics Program, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA. 15 Department of Genetics, Harvard Medical School, 77 Avenue Louis Pasteur, NR8 0330, Boston MA 02115, Massachusetts, USA. 16 Department of Genetics and Computational Biology, QIMR Berghofer Medical Research Institute, 300 Herston Road, Brisbane 4006, Australia. 17 Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, PO Box 9600, Leiden 2300 RC, The Netherlands. 18 Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München, Ingolstädter Landstraße 1, Neufahrn 85764, Germany. 19 Institute of Experimental Genetics, Genome Analysis Center, Helmholtz Zentrum München, Ingolstädter Landstraße 1, Neufahrn 85764, Germany. 20 German Center for Diabetes Research at Helmholtz Zentrum München, Ingolstädter Landstraße 1, Neufahrn 85764, Germany. 21 Lehrstuhl für Experimentelle Genetik, Technische Universität München, Freising-Weihenstephan 85350, Germany. 22 Department of Genetology and Genetics, Leiden University Medical Center, PO Box 9600, Leiden 2300 RC, The Netherlands. 23 Department of Genetics, CB50, University Medical Center Groningen, University of Groningen, P.O. Box 30001, Groningen 9700 RB, The Netherlands. 24 Statistical and Genomic Medicine, Imperial College, London, W2 1PG, United Kingdom. 25 Department of Biomedical Engineering, Queensland University of Technology, 60 Musgrave St, Kelvin Grove QLD 4059, Queensland, Australia. 26 Department of Psychiatry, VU University Medical Center, A.J. Ernststraat 187, Amsterdam 1081 H L, The Netherlands. 27 Department of Endocrinology, Leiden University Medical Center, S4-P, PO Box 9600, Leiden 2300 RC, The Netherlands. 28 Institute of Medical Informatics, Biometry and Epidemiology, Chair of Genetic Epidemiology, Ludwig-Maximilians-Universität, Munich 80337, Germany. 29 Research Unit of Molecular Epidemiology, Helmholtz Zentrum München, Ingolstädter Landstraße 1, Neufahrn 85764, Germany. 30 Hannover Unified Biobank, Hannover Medical School, Hannover 30625, Germany. 31 Institute for Human Genetics, Hannover Medical School, Carl-Neuberg-Strasse 1, Hannover 30625, Germany. 32 Faculty of Biology, Ludwig-Maximilians-Universität, Planegg-Martinsried 82152, Germany. 33 Department of Physiology and Biophysics, Weill Cornell Medical College in Qatar (WCM-Q), PO Box 24144, Education City—Qatar Foundation, Doha, Qatar. 34 Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK. 35 Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Churchill Hospital, Headington OX3 7JE, UK. 36 Oxford National Institute for Health Research Biomedical Research Centre, The John Radcliffe Hospital, Old Road, Headington OX3 7JE, UK. 37 Institute of Epidemiology II, Helmholtz Zentrum München—German Research Center for Environmental Health, Ingolstädter Landstraße 1, Neufahrn 85764, Germany. * These authors contributed equally to this work. ** These authors jointly supervised this work. Correspondence and requests for materials should be addressed to H.H.M.D. (email: h.h.m.draisma@vu.nl).
Metabolite levels in human blood reflect the physiological state of the body, and may differ between individuals because of variation in genetic makeup and environmental exposure. The study of the genetic contribution to variation in metabolite levels is an important basis for improved aetiological understanding, prevention, diagnosis and treatment of complex disorders. Modern high-throughput metabolomics enables the cost-effective measurement of large metabolite panels in blood samples obtained from many individuals. The data generated by such metabolomic experiments have been combined with genotypic data in several recent genome-wide association (GWA) studies. Indeed, the combined investigation of large numbers of genetic variants and large numbers of metabolic traits is beginning to draw a systems-wide overview of genetic influences on human metabolism. However, the heritability estimates from twin and family studies suggest that additional genetic variants influencing variation in serum metabolite levels remain to be found in GWA studies.

In the current study, we set out to further characterize the genetic contribution to variation in human blood metabolite levels. We perform GWA and meta-analyses for the concentrations of 129 serum metabolites in seven independent cohorts, with replication analyses in one additional cohort. To functionally characterize the significant single-nucleotide polymorphism (SNP)-metabolite associations, we integrate the results of the GWA meta-analyses with those from gene expression analysis in whole blood and the liver. Finally, we compare the variance explained by significantly associated SNPs with heritability estimates for each metabolite.

We identify 4,068 significant SNP-metabolite associations, involving 59 independent SNPs and 85 different metabolites. Five of the fifty-nine independent SNPs are novel for serum metabolite levels. The newly found SNP-metabolite associations may lead to a better understanding of cardiovascular and metabolic disease, and may have implications for chemotherapy. Our findings contribute to the understanding of human metabolism.

### Results

**Discovery meta-analysis of GWA scans.** Primary genetic association analyses were carried out in seven cohorts (TwinsUK, KORA, EGCUT, LLS, QIMR, ERF and NTR) with a combined sample size of 7,478 individuals. Characteristics of the study participants included in the analyses (all of European descent) are given in Supplementary Table 1. Within each cohort, SNP genotypes were imputed and analysed for association with the concentrations of each metabolite, assuming a linear model of association and correcting for population stratification (see Methods and Supplementary Table 2). Supplementary Tables 3–5 and Supplementary Data 1 list the characteristics of the 129 metabolites (18 acylcarnitines, 14 amino acids, 82 glycerophospholipids, 14 sphingolipids and hexose) that were measured in the serum samples from all study participants using the Biocrates platform. The cohort-level GWA results were pooled in inverse variance-weighted, fixed-effects meta-analysis. The values of the genomic control lambda applied to the individual cohort-level results for each metabolite before meta-analysis) varied between 0.976 and 1.081 across all metabolites and cohorts (see Supplementary Table 6), suggesting little residual influence on the GWA results of population stratification and other potential confounders. A three-dimensional Manhattan plot providing an overview of the association P values in the discovery phase for all metabolites is given in Fig. 1; two-dimensional Manhattan plots and quantile-quantile plots for each metabolite separately are given in Supplementary Figs 1 and 2, respectively. Overall, 4,068 SNP-metabolite associations reached genome-wide significance (Z-test, \( P < 1.09 \times 10^{-9} \)), which reduced to 123 associations involving 59 independent SNPs and 85 different metabolites. Of these 123 associations (listed in Supplementary Data 2), 4 represented secondary association signals according to approximate conditional analysis. Regional association plots, showing the association signals in the regions surrounding the lead metabolomic SNPs, are given for all 123 associations in Supplementary Fig. 3. SNPs representing independent association signals were aggregated into 31 genomic loci, which are listed in Supplementary Data 3. Figure 2 depicts all associations between loci and metabolites as detected in the discovery phase.

Five independent SNPs had not been associated with variation in serum metabolite levels in previous GWA studies (see Table 1). To further interpret the association of the remaining 54 SNPs with serum metabolite concentrations, we compared our findings with those from 11 published GWA studies for which at least one of the included metabolites overlapped with the current study. The identified associations of known SNPs with metabolites that were significant in discovery stage meta-analysis in the current study and that had not been reported in those previous studies are highlighted in Fig. 2 and in Supplementary Data 2.

**Replication analysis.** Replication analyses were performed in an independent sample (N = 1,182) from the KORA S4 cohort (hereafter KORA S4 replication sample) for the five new SNPs for serum metabolite levels that had been found in the discovery phase meta-analysis. The associations with their most strongly associated metabolite were replicated for four of these five novel SNPs; the only non-replicated association was that between SNP rs7582179 and metabolite PC ae C44:5. Although the effect sign was concordant between the discovery set and the KORA S4 replication sample (Table 1), this association was significant in the discovery phase for the NTR and KORA cohorts only (see Supplementary Fig. 4).

**Integration with gene expression analysis results.** We integrated the results of the metabolomics discovery stage GWA meta-analysis with the results of gene expression analyses in whole blood and the liver. In whole blood, both cis and trans expression quantitative trait locus (cis-eQTL and trans-eQTL, respectively) analyses were performed in two different samples originating from the United Kingdom, the Netherlands and Estonia: the Dutch NTR-NESDA sample (N = 5,071) and the Fehrmann-EGCUT sample comprising data from three cohorts that were meta-analysed (total N = 2,360; see Methods and Supplementary Methods). The results of cis-eQTL analysis for lead metabolomic SNPs showing overlap with cis-eQTL SNPs are given for the NTR-NESDA and Fehrmann-EGCUT samples in Supplementary Data 4 and 5, respectively. Significant (false discovery rate \(< 0.05\)) trans-eQTL effects for lead metabolomic SNPs in the Fehrmann-EGCUT sample are listed in Supplementary Data 6. We did not detect trans-eQTL effects for the lead metabolomic SNPs in the NTR-NESDA sample. Thirty-five lead metabolomic SNPs identified cis-eQTLs in at least one of the searched tissues (i.e., whole blood and/or the liver) with a \( t \)- or Kruskal-Wallis test \( P \textless 0.001 \), defining a total of 67 SNP-gene pairs and 28 different genes (see Supplementary Data 7). The cis-eQTL analysis results were used to support the annotation of likely causal genes to loci that displayed significant association with variation in serum metabolite concentrations in the discovery stage meta-analysis (see Supplementary Data 3). Of the 28 genes, 14 were predicted to be causal on the basis of our annotation and the other 14 were predicted to be non-causal.
Figure 1 | Manhattan plots for all metabolites targeted by the Biocrates AbsoluteIDQ p150 kit (\(N = 1,497,7,478\)). These plots graphically display the \(P\) values for significant (Z-test \(P < 1.09 \times 10^{-5}\)) SNP-metabolite associations in the discovery phase in the current study. (a) A three-dimensional view; orthogonal projections are given in (b) and (c). SNPs are arranged according to genomic location along the 'Chromosome' axes. The ordering of the metabolites along the 'Metabolite index' axes is equal in both (a) and (c), and equal to that in Supplementary Table 3. In (a) and (b), all data points are displayed semi-transparent and therefore opaque regions in the plot indicate clusters of significant associations. In (b), loci are identified by most plausible causal gene or, if no plausible genes found, by nearest gene. Where multiple plausible genes could be identified at the locus (possibly for different metabolites), the gene names are separated by an underscore ("_") in the locus name. In (c), the size of the markers scales linearly with \(-\log_{10}(P\ value)\). This Figure is also supplied as a movie (see Supplementary Movie 1).
Figure 2 | Associations between loci and metabolites detected in stage 1 meta-analysis in the current study (*N* = [1,588, 7,478]). Loci significantly associated with at least one metabolite are depicted as grey circles. Biochemical classes (see Supplementary Table 3) of the metabolites (hexagons) are indicated by node colours: green, acylcarnitines; blue, amino acids; purple, glycerophospholipids; yellow, sphingolipids. Arrows point from each locus to the associated metabolite(s); arrow widths scale linearly with -log_{10}(association *P* value). Grey arrows denote previously known associations; red arrows denote associations that were newly discovered on the basis of stage 1 meta-analysis in the current study (that is, either associations with new SNPs for serum metabolite levels, or an association of a known SNP with a new metabolite with respect to 11 previous GWA studies for serum metabolite levels). It has been described previously that a large proportion of the variance observed for serum metabolite levels is explained by significantly associated SNPs, which was carried out for each of 129 metabolites measured in the serum underpinnings of variation in circulating metabolite levels in humans. To this end, we employed a well-established targeted metabolomics platform (Biocrates) in combination with genome-wide SNP genotyping and imputation in eight independent cohorts of European descent. By meta-analysis of GWA analyses carried out for each of 129 metabolites measured in the serum samples of all individual study participants, the current study identified 123 significant SNP-metabolite associations between 59 independent SNPs and 85 different metabolites. Five of the independent SNPs were new for variation in serum metabolite levels.

**Discussion**

We set out to enhance the current understanding of the genetic underpinnings of variation in circulating metabolite levels in humans. To this end, we employed a well-established targeted metabolomics platform (Biocrates) in combination with genome-wide SNP genotyping and imputation in eight independent cohorts of European descent. By meta-analysis of GWA analyses carried out for each of 129 metabolites measured in the serum samples of all individual study participants, the current study identified 123 significant SNP-metabolite associations between 59 independent SNPs and 85 different metabolites. Five of the independent SNPs were new for variation in serum metabolite levels.
Proportion of variance explained by significantly associated SNPs was estimated as Pearson’s phi coefficient squared. Metabolites are grouped according to and that displayed genome- and metabolome-wide associations with SNPs in stage 1 GWA meta-analysis in the current study (Figure 3 | Decomposition of variation in serum metabolite levels). Seventy-six metabolites tend to match the known function of these genes. The five new SNPs for serum metabolite levels are also which in most cases encoded a metabolite transporter protein or enzyme. The five new SNPs for serum metabolite levels are also

<table>
<thead>
<tr>
<th>Lead metabolomic SNP</th>
<th>Lead metabolite</th>
<th>Cytoband</th>
<th>P value in discovery phase</th>
<th>EA/NEA</th>
<th>EAF in discovery phase</th>
<th>Beta in discovery phase</th>
<th>Total N in discovery phase</th>
<th>Nearest gene</th>
<th>Beta in replication phase</th>
<th>P value in replication phase</th>
<th>EAF in replication phase</th>
<th>Replicated Locus name</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs12210538</td>
<td>C18:2</td>
<td>6q21</td>
<td>(5.03 \times 10^{-21})</td>
<td>A/G</td>
<td>80.8%</td>
<td>0.086</td>
<td>6,574</td>
<td>SLC22A16</td>
<td>0.099</td>
<td>2.65 \times 10^{-13}</td>
<td>75.9%</td>
<td>*</td>
</tr>
<tr>
<td>rs17657817</td>
<td>Orn</td>
<td>6q23.2</td>
<td>(1.32 \times 10^{-11})</td>
<td>T/C</td>
<td>98.0%</td>
<td>0.156</td>
<td>2,991</td>
<td>ARG1</td>
<td>-0.123</td>
<td>1.40 \times 10^{-4}</td>
<td>97.5%</td>
<td>*</td>
</tr>
<tr>
<td>rs2246012</td>
<td>Orn</td>
<td>6q23.2</td>
<td>(6.43 \times 10^{-12})</td>
<td>T/C</td>
<td>83.9%</td>
<td>0.045</td>
<td>7,476</td>
<td>ARG1</td>
<td>0.044</td>
<td>1.57 \times 10^{-3}</td>
<td>84.5%</td>
<td>*</td>
</tr>
<tr>
<td>rs7582179</td>
<td>PC ae C44:5</td>
<td>2q31.2</td>
<td>(4.07 \times 10^{-10})</td>
<td>A/G</td>
<td>16.8%</td>
<td>-0.048</td>
<td>5,360</td>
<td>AGPS</td>
<td>0.021</td>
<td>0.147</td>
<td>16.5%</td>
<td>*</td>
</tr>
<tr>
<td>rs7700133</td>
<td>PC ae C44:5</td>
<td>4q35.1</td>
<td>(3.35 \times 10^{-11})</td>
<td>T/C</td>
<td>30.5%</td>
<td>0.036</td>
<td>7,476</td>
<td>CENPU</td>
<td>0.038</td>
<td>1.12 \times 10^{-3}</td>
<td>30.9%</td>
<td>*</td>
</tr>
</tbody>
</table>

| Abbreviations: EA/NEA, effect allele/non-effect allele. EAF, frequency of EA; SNP, single-nucleotide polymorphism. | Lead metabolite, metabololite displaying strongest association with SNP in discovery phase GWA meta-analysis in current study. P values in discovery and replication phases were calculated by Z- and t-tests, respectively. Loci that were replicated in the KORA S4 replication sample (\(P < 0.05\) after Bonferroni correction for 5 tests) are indicated by *. Loci are identified by most plausible causal gene or, if no plausible genes found, by nearest gene. Where multiple plausible genes could be identified at the locus (possibly for different metabolites), the gene names are separated by an underscore (\(_\) ) in the locus name. |

Consistent with previous reports, for the majority of all 59 independent SNPs we were able to annotate a likely causal gene, which in most cases encoded a metabolite transporter protein or enzyme. The five new SNPs for serum metabolite levels are also all located nearby such genes, and their associations with metabololites tend to match the known function of these genes. SNP rs7582179 in the AGPS gene is associated with the choline plasmalogen PC ae C44:5. Mutations in AGPS (encoding the enzyme alkylglycerophosphate synthase) are known to cause rhizomelic chondrodysplasia punctata type 3 (RCDP3; OMIM:600121 (ref. 14)), a rare autosomal recessive disorder that is fatal, with death occurring often early in childhood. Clinically, RCDP3 is characterized by significantly delayed and abnormal physical and mental development, with shortness of the proximal limb bones (‘rhizomelia’ being one of the hallmarks. RCDP3 has been shown to result from reduced production of plasmalogens (a type of ether phospholipids) by alkylglycerophosphate synthase in peroxisomes. The association in the current study of a SNP within the AGPS gene with the serum concentration of the choline plasmalogen PC ae C44:5 is therefore perfectly concordant with the known gene-disease link between AGPS and RCDP3. PC ae C44:5 also associated significantly with the new SNP rs7700133 located near the ACSL1 gene, encoding long-chain acyl CoA synthetase 1. Previous studies have shown links between genetic and transcriptional variation of ACSL1 and the metabolic syndrome15,16. The new SNP rs12210538, located within the SLC22A16 gene, associated with the two acylcarnitines C18:1 and C18:2. This gene encodes a carnitine transporter that mediates the uptake of anticancer drugs such as bleomycin and doxorubicin into tumour cells, and its activity correlates with treatment response17,18. Significant associations were found for two SNPs (rs17657817 and rs2246012) located inside the ARG1 gene (coding for the enzyme arginase) with serum concentrations of the amino acid ornithine that participates in the urea cycle. Importantly, the global arginine bioavailability ratio (that is, the ratio of arginine to ornithine and citrulline19) is of interest as a
potential biomarker for endothelial dysfunction, which is a known risk factor for the development of cardiovascular disease. The two newly identified SNPs associated with serum ornithine levels might now be used as instrumental variables in cost-effective Mendelian randomization studies in large samples of individuals, to investigate the possible causal relationship among ornithine, endothelial dysfunction and subsequent cardiovascular disease. Among three meta-analyses of coronary artery disease and myocardial infarction as carried out by the CARDIoGRAMplusC4D Consortium, the association with SNP rs2246012 was the strongest in the CARDIoGRAM GWA study (P = 0.002) involving 22,233 cases and 64,762 controls.28 This suggests that the link between genetic variation at SNP rs2246012 and variation in serum ornithine levels as identified in the current study will indeed be useful to further establish the possible link between the ARGG gene and cardiovascular disease.

We compared the significant SNP-metabolite associations from the current study with those reported in 11 previous publications that employed high-resolution methods to assay the serum metabolome. For several SNPs that were associated with variation in metabolite levels in the previous studies, we identified new associations with individual metabolites. These new associations strengthen the evidence for the associations of these known SNPs with specific metabolites, demonstrating their extended effect on phenotypes that are closely related to the metabolites with which their association was discovered initially.4,23 Also, it has been demonstrated that GWA studies with the more refined metabolic phenotypes provided by metabolomics often yield effect sizes that are larger than those observed in GWA studies of composite measures such as high-density lipoprotein cholesterol, suggesting that these more refined metabolic phenotypes provide better intermediate traits.8,9 In this context, it is interesting to note that the large proportion of explained heritability we observed for lysoPC a C20:4 (19%) was caused exclusively by the association with a single SNP in the FADS1–3 locus, an observation that is in line with the results from previous studies.4,5,9

In conclusion, the results obtained in the current study contribute to the understanding of the genetic background of variation in serum metabolite levels, and are important for further progress in biomedical and pharmaceutical research.

Methods

Participants

The meta-analysis included GWA data from 7,478 participants from seven cohorts originating from five countries (the Netherlands, Germany, Australia, Estonia and the United Kingdom). The independent test sample (‘KORA S4 replication sample’) consisted of 1,182 additional KORA participants. The following local research ethics committees approved the individual studies: KORA, Ethics Committee of the Bavarian Medical Association (Bayerische Landesärztekammer); NTR, Central Ethics Committee on Research Involving Human Subjects of the VU University Medical Center, Amsterdam; ECGUT, Ethics Review Committee on Human Research of the University of Tartu; TwinsUK, St Thomas’ Hospital ethics committee; ERF, medical ethics board of the Erasmus MC Rotterdam, the Netherlands; LLIS, Medical Ethical Committee of the Leiden University Medical Center, MRIM Human Research Ethics Committee. Local research ethics committee approval and consent was obtained from all participants. Sample characteristics for all cohorts included in this study and detailed study sample descriptions are given in Supplementary Table 1 and in the Supplementary Methods, respectively.

Biocrates metabolite quantification.

Targeted metabolic measurements were performed using electrospray–flow injection analysis–tandem mass spectrometry methods and the Biocrates AbsoluteIDQ p150 kit (BIOCRATES Life Sciences AG), which enables quantification of a total of 163 metabolites (see Supplementary Table 3 for an overview of all metabolites targeted by this kit)24. The method of AbsoluteIDQ p150 kit has been proven to be in accordance with FDA Guideline ‘Guidance for Industry—Biobaktical Method Validation (May 2001)’25, which implies proof of reproducibility within a given error range. For all cohorts, metabolite measurements were carried out at the Metabonomics Platform of the Genome Analysis Center at the Helmholtz Zentrum München, Germany as per the manufacturer’s instructions.24,26,27. In brief, the used metabolomic measurement technique is based on a targeted profiling scheme that is used to quantitatively screen for known small-molecule metabolites by multiple reaction monitoring, neutral loss and precursor-ion scans. Internal standards served as reference for the calculation of all metabolite concentrations, which are reported as absolute concentrations. Data evaluation for quantification of metabolite concentrations and quality assessment have been performed with the MetIDQ software package, which is an integral part of the AbsoluteIDQ kit. Stability of the assay was assessed using the measurement results of five aliquots of the same reference blood sample on every plate. The quality control of the Biocrates metabolite concentration measurement data was performed by each participating cohort as follows:27 for each cohort, metabolite profile measurements for all individuals were performed on multiple plates. For each metabolite i and plate j, the coefficient of variation (CV, j) was calculated as: CV, i = s_{ij} / \mu_{ij}, where the standard deviation (s.) and mean were calculated over all reference measurements per plate j (five per plate). Summary statistics for the metabolite concentration data for each cohort were compared with the corresponding detection limits (DL) and lower limit of quantification (LOQ) established by the manufacturer of the AbsoluteIDQ p150 kit (BIOCRATES). A metabolite was excluded from further analyses for a particular cohort if its concentration measurement data did not meet all of the following criteria: (i) mean CV, over all plates <25%; (ii) <5% missing values; (iii) median ± lower limit of quantification (for metabolites reported as absolute concentrations) or ± limit of detection (for semiquantitatively measured metabolates). Outlying metabolite concentration values (data points) and outlying samples were also removed, and the missing data points were imputed with the ‘K’ package ‘mice’.27 The resulting concentration values for each metabolite were natural log-transformed in order to attain a normal distribution. Throughout the analysis, the Biocrates AbsoluteIDQ p150 platform are abbreviated as follows: acylcarnitines, Cx:y; hydroxyacylcarnitines, C(OH)x:y; dicarboxylacylcarnitines, Cx:y–DC; sphingomyelins, SMx:y; N-hydroxylacylcylsphingosylphosphocholine, SM (OH)x:y; phosphatidylcholines, PC (aa = dacyl, ae = acyl-acyl). Lipid side chain notation is abbreviated as x:y, where x denotes the number of carbons in the side chain and y the number of double bonds.

Association analyses and meta-analyses

Genome-wide SNP genotyping was performed in each cohort with standard genotyping technologies (see Supplementary Table 2 and Supplementary Methods). For the samples contributing to the stage 1 (discovery) meta-analysis, imputation was conducted with reference to the HapMap phase 2 build 24 and phase 3 release 22 or 23 database (HapMap phase 2 release 23 for two European ancestry cohorts). Two independent analysts carried out additive model fixed-effects meta-analysis of association data for imputed autosomal SNPs, using two different software packages (METAL30 and GWAMA31). For a given SNP, cohort-specific effect size estimates were weighted inversely with their variance. Throughout the manuscript, we report the P values as resulting from Z-tests of association as carried out by METAL. A P value equal to 5.0 × 10−8 was adopted as the threshold for genome-wide suggestive association. For each SNP and the concentration of a metabolite, based on the approximate number of independent SNPs in samples of European ancestry.27 To obtain a threshold for significant association, taking account of the number of metabolites tested and their intercorrelations, the threshold value for suggestive association was divided by the number of independent metabolites (Meffl). In the metabolomic data as employed in this study, the method of Meffl was estimated by Li and Ji33. The value of Meffl was estimated as 46 in both ERF and NTR, rendering the P value threshold for genome- and metabolome-wide significance in the present study to be equal to 5.0 × 10−8/46 = 1.09 × 10−10.

Definition of loci and secondary signals analysis

Independent signals of association were identified in the GWA meta-analysis. Results for each metabolite separately, using the linkage disequilibrium-based ‘clumping’ procedure as implemented in PLINK34. This procedure takes all SNPs that show a P value of association with a phenotype below a threshold (‘--clump-p1’), and forms clumps of these ‘index’ SNPs together with all other SNPs that are in linkage disequilibrium with one of the index SNPs. Quantile was controlled by the parameter ‘--clump-qsq’ and distance was controlled by parameter ‘--clump-kb’ to these index SNPs. For the current study, we used the following parameter settings: ‘--clump-p1’, 5.0 × 10−8; ‘--clump-k2’, 20 kb.
0.1. ‘—clump-kb’, 1,000. As input for the ‘clumping’ procedure, we used association
P values from the discovery phase meta-analysis results and linkage disequilibrium
patterns as estimated from the HapMap 2 Build 36 release 24 reference set. For
each metabolite separately, secondary association signals at a locus were verified by
approximate conditional analysis as implemented in GCTA55. In this analysis, the
association for the secondary association signal ‘top SNP’ (that is, the SNP with the
lowest P value of association with variation in serum metabolite level at the
secondary association signal) was conditioned on the top SNP on the locus and
metabolite under consideration. As input for the approximate conditional analysis,
we used the discovery phase GWA meta-analysis results and the imputed SNP
haplotype data from the NTR cohort as a reference for LD structure.55 We report
only secondary signal SNPs for which the P value remained < 10^{-9} in this
approximate conditional analysis. In the current manuscript, the term ‘lead
metabolomic SNP’ refers to a top SNP at a locus or secondary association signal for
one or more metabolites. We identified genomic loci significantly associating with
metabolite levels by grouping lead metabolomic SNPs located within 1 Mb from
each other over all metabolites.

Identification of new SNP–metabolite associations. The approach used in the
current study to identify novel associations between SNPs and serum metabolite
developed is described in Supplementary Fig. 6 and in the Supplementary Methods.
In brief, we applied two complementary methods: the first method identified novel
SNPs associated with variation in serum metabolite levels, and the second method
identified novel SNP–metabolite associations with respect to 11 previous GWA
studies that included at least one metabolite that was also included in the
current study.

Replication analyses. Replication analyses were performed in the KORA S4
replication sample for the associations of the five new SNPs for serum metabolite
levels (listed in Table 1) with their lead metabolites (that is, the metabolites that
were most strongly associated with these SNPs in the discovery phase
meta-analysis).

Lookup of association with cardiovascular disease. Data on coronary artery
disease/myocardial infarction have been contributed by CARDIoGRAMplusC4D
investigators and have been downloaded from www.CARDIOGRAMPLUSC4D.
ORG. We performed a lookup of the associations with SNP rs2246012 in the results from all three meta-analyses as provided on this website23,37,38.

Association and Manhattan plots. For each lead metabolomic SNP, the
LocusZoom39 tool was used to generate association plots in the region between
500 kb before the locus minimum position and 500 kb after the locus maximum
position. Manhattan plots for each metabolite were generated based on the
discovery phase meta-analysis results using in-house developed Python40 code.

SNP annotation. To facilitate the manual process of selecting plausible candidate
genes for each locus, we used an automated workflow developed in-house to
generate reports containing the associated protein, enzyme, metabolic reaction,
pathway and network relationships of each locus. The report included the
locations of each lead metabolomic SNP, SNPs within this window that were published in GWAS
Catalog41 or in GTEx-eQTL (http://www.ncbi.nlm.nih.gov/gtex/GTEX2) were also
listed. In detail, the reports created by our workflow were based on the
NCBI-Gene (http://www.ncbi.nlm.nih.gov/gene), GTEx-eQTL, GWAS Catalog,
ConsensusPathDB42, UniProtKB43, OMIM44, Gene Ontology45, TCDD46,
ExPASy47 and KEGG database48. These databases had been downloaded earlier
from the respective File Transfer Protocol servers and have been integrated offline
in MATLAB (R2009a, The Mathworks Inc.). Overlap of lead metabolomic SNPs
with cis-eQTL SNPs was also used as evidence to support the annotation of likely
causal genes to loci. In case no biologically plausible gene could be found, the locus
was given the name of the nearest gene; a similar approach was followed in the
study by Shin et al.11

Variance explained. We estimated the proportion of phenotypic variance explained
by each independent association signal (lead SNP for a locus or secondary association signal) as Pearson’s ϕ coefficient squared:

ϕ^2 = \frac{s_i^2}{r^2} (2)

where χ^2 = χ^2 = (\hat{β}_i \hat{SE}(\hat{β}_i))^2 / N is the sample size in the discovery phase meta-
analysis for the SNP–metabolite association under consideration; \hat{β}_i is the ordinary
least-squares estimate of β_i (that is, the regression coefficient for the SNP as
estimated in the discovery phase meta-analysis); and \hat{SE}(\hat{β}_i) is its standard error.

For each metabolite, we added up the proportions of variance in metabolite level explained by independent association signals to estimate the total proportion of phenotypic variance explained. We also estimated for each metabolite the proportion of heritability of metabolite level variability explained by approximately
independent association signals. As an estimate of heritability for this analysis we
used the monozygotic twin correlations for each metabolite, based on data from 181 pairs from the NTR cohort13. The proportion of heritability in metabolite level variation explained by independent association signals was estimated by dividing
the proportion of phenotypic variance explained by independent association
signals, by the monozygotic twin correlation. The total proportion of heritability
explained for a particular metabolite was estimated by adding up the proportions of variance explained by all approximately independent association signals for that
metabolite.

PGS analysis. We investigated evidence for the polygenic nature of variation in
serum metabolite levels by building a multi-SNP predictor from the meta-analysis
results for each metabolite to predict the levels of the same metabolite in an
independent target cohort (KORA S4 replication sample). Such a multi-SNP
predictor, or PGS, reflects the weighted sum of multiple SNPs associated with a phenotype
and forms the basis to select SNPs based on liberal significance thresholds (for example, 0.001 and so on). In the target
data, PGSs were calculated for each individual for each set of SNPs by multiplying
the number of effect alleles per SNP (0, 1 or 2) with the beta from the
meta-analysis, summed over all SNPs in the set of SNPs. We performed the PGS analysis49 using the regression coefficients (betas) from the discovery phase
meta-analyses as weights. Analyses were performed for the 127 metabolites for
which concentration data were available both in the discovery sample and in the
KORA S4 target sample. For each of these metabolites, SNPs representing
approximately independent association signals were selected in the discovery phase
meta-analysis results using the PLINK clumping procedure. SNPs with P values
of association with metabolite concentration levels in the discovery phase-meta-analysis below the following thresholds were included: P < 1.0 × 10^{-5}; P < 1.0 × 10^{-7};
P < 1.0 × 10^{-9}; P < 1.0 × 10^{-10}; P < 1.0 × 10^{-12}; P < 5.0 × 10^{-7}; P < 0.1; P < 0.2; P < 0.3; P < 0.4; P < 0.5; P < 0.6; P < 0.8; P < 0.9; P < 1.0.
For each clump of SNPs, the index SNP was taken for possible
inclusion in the score computation. From the resulting set of SNPs eligible for
inclusion in the PGS analysis, A/T and G/C SNPs for which (0.35 < MAF < 0.50) were excluded because these SNPs are potentially ambiguous and therefore may
lead to spurious association in the case of strand flips50. From the imputed SNP
generated for each of the KORA S4 target sample, the best set of SNPs corresponding with
the remaining clump index SNPs were selected. A PGS was constructed for each
individual in the KORA S4 replication sample using the ‘—score’ procedure as
implemented in PLINK v. 1.07 (http://pngu.mgh.harvard.edu/~purcell/plink/
profile.shtml). The resulting PGS was included as a covariate in a multiple linear
regression analysis that was similar to the regression that was carried out in the
primary single SNP-based GWA analysis:

y = α + β_GPG + β_1age + β_2sex + β_3(study − specific covariates) (3)

where y represents log(metabolite) values, and the study-specific covariates includes adjustments for, for example, population stratification (and thus β_3 can be a vector). The Biocrates metabolite values were obtained and preprocessed using the
same methods as described for the primary GWA in the Section ‘Biocrates
metabolome quantification’. The proportion of variation explained by the PGSs was
assessed by comparing the raw (that is, unadjusted) with the PGS
reduced (that is, a model including the genetic score as a covariate) with the raw
R^2 values for the full model.

EQTL analyses. Data from two independent samples originating from the United
Kingdom, the Netherlands and Estonia (the Dutch NTR-NESDA sample and the
Fehrmann-EGCUT sample) were used for cis- and trans-eQTL mapping in whole
blood. Details of these analyses, and of the integration of the cis-eQTL analysis
results with the metabolomics genome-wide meta-analysis, are provided in the Supplementary Methods. We also assessed the overlap of lead
metabolomic SNPs with cis-eQTL signals in the liver as catalogued in the
GTEx-eQTL database, following the method of Shin et al.11 for each lead
metabolomic SNP, we retrieved all SNPs with r^2 > 0.8 in the 1000 Genomes Project
Phased (CEU population). All cis-eQTLs within a 1-Mb window centred on the
lead SNP were retrieved from the GTEx-eQTL database, and the best eQTL P value
was noted. The cis-eQTL results for which overlap with lead metabolomic SNPs
was shown and that displayed association P values < 0.001 are given in
Supplementary Data 7.

Meta-analysis. The full meta-analysis results for all metabolites are available at
www.twielingenregister.org/engagebiocategorieswa

References
1. Suhr, K. & Gieger, C. Genetic variation in metabolic phenotypes: study designs
2. Suhr, K. et al. Human metabolic individuality in biomedical and

Acknowledgements
We thank all participants. This research was supported by the European Community's Seventh Framework Programme (FP7/2007–2013), ENGAGE Consortium, grant agreement HEALTH-F4-2007-201413. The Cooperative Health Research in the Region Augsburg (KORA) study was supported by the Helmholtz Zentrum München—German Research Center for Environmental Health, which is funded by the German Federal Ministry of Education and Research (BMBF) and by the State of Bavaria. KORA research was supported within the Munich Center of Health Sciences (MC-Health), Ludwig-Maximilians-Universität, as part of LMUinnovativ. This work was financed by the German National Genome Research Network (NGFN; NGFNPlus, project number 01GS0843). K. Suhre is supported by Biomedical Research Program funds at the Wellcome Trust Sanger Institute and Medical Research Council (W099/000815). We acknowledge help from the Estonian Genome Center of Tartu University (EGUCT) received support from the Estonian Government IUT20–60, IUT24–6, Estonian Research Roadmap through the Estonian Ministry of Education and Research (3.2.0341.11–0312), Center of Excellence in Genomics (EXCE-GEN), Development Fund from the University of Tartu (SP1G4AREG), an EFSF New Horizons Grant and the US National Institute of Health (ROI1K075787). We acknowledge the EGUCT personnel, especially Mr V. Soot. Data analyses were carried out in part in the High Performance Computing Center of the University of Tartu. TwinsUK receives support from the National Institute for Health Research (NIHR) Clinical Research Facility at Guy’s & St Thomas’ Hospital Foundation Trust and NIHR Biomedical Research Centre based at Guy’s and St Thomas’ NHS Foundation Trust and King’s College London. The Netherlands Twin Register (NTR) was funded by the Netherlands Organization for Scientific Research (NWO: MagW/ZoonMW grants 904–61–090, 985–10–002, 904–61–193, 480–04–004; 400–05–717; Middelgroet–911–09–032; 56–464–14192), Center for Medical Systems Biology (CMSB); Netherlands Bioinformatics Centre/BioAssist/RK(2008.024); Biobanking and Biomolecular Resources Research Infrastructure (BBMRI)-NL (164.021.067) and the European Science Council (ERC-280374). H.L.M.D. is supported by an EMBO Fellowship. This work was carried out on the Dutch national e-infrastructure with the support from the SURF Foundation. The Erasmus Ruschen Family (EUF) is grateful to Petra Veraart for her help in genealogy, Jeannette Vergeer for her help in genealogy, andQIMR cohort, serum metabolite level assessment was co-funded by the European Community’s Seventh Framework Programme (FP7/2007–2013), ENGAGE project, grant agreement HEALTH-F4-2007-201413 and NHMRC-European Union Collaborative Research Grant 496739.
Author contributions

Additional information
Accession codes. The metabolomic data described in this manuscript are uploaded to MetaboLights and are available under the accession code MTBLS192.

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/