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

Harmen H.M. Draisma1,2,3,*, René Pooi1,2,4,*, Michael Kobi5, Rick Jansen5, Ann-Kristin Petersen5, Anika A.M. Vaarhorst4,7,8, Idil Yet9, Toomas Haller10, Ayşe Demirkiran11,12, Tõnu Esko10,13,14,15, Gu Zhu16, Stefan Böhinger17, Marian Beekman17, Jan Bert van Klinken11, Werner Römisch-Margl18, Cornelia Prehn19, Jerzy Adamski19,20,21, Anton J.M. de Craen22, Elisabeth M. van Leeuwen12, Najaf Amin12, Harish Dharuri11, Harm-Jan Westra23, Lude Franke23, Eco J.C. de Geus1,2, Jouke Jan Hottenga1,2, Gonneke Willemsen1,2, Anjali K. Henders16, Grant W. Montgomery16, Dale R. Nyholt16, John B. Whitfield16, Brenda W. Penninx2,25, Tim D. Spector9, Andres Metspalu10, P. Eline Slagboom4,7, Ko Willems van Dijk11,26, Peter A.C. ’t Hoen11, Konstantin Strauch5,27, Nicholas G. Martin16, Gert-Jan B. van Ommen11, Thomas Illig28,29,30, Jordana T. Bell5, Massimo Mangino9, Karsten Suhr18,31,32, Mark I. McCarthy33,34,35, Christian Gieger5,28,36, Aaron Isaacs12, Cornelia M. van Duijn8,12,24,11 & Dorret I. Boomsma1,2,4,22

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 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 (Z-test, P<10^-9) 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.
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 variants 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 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 (TwinUK, 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 hexases) 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- and metabolome-wide significance (Z-test, P < 1.09 × 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 analysis in whole blood and the liver. In whole blood, both cis and trans expression quantitative trait loci (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 Fehmarn-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 Fehmarn-EGCUT samples in Supplementary Data 4 and 5, respectively. Significant (false discovery rate < 0.05) trans-eQTL effects for lead metabolomic SNPs in the Fehmarn-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 (r, Z- or Kruskal-Wallis test) P < 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 × 10^-7) 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 \text{ 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 $-\log_{10}(\text{association } P \text{ value})$. 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). The locus-metabolite association network separates into 12 connected components or disconnected sub-networks, each including metabolites from maximally two chemical classes. This figure was created using Cytoscape.

**Variance explained.** It has been described previously that a relatively small number of genetic variants can explain a relatively large proportion of the variance observed for serum metabolite levels. Therefore, we compared the variance in serum metabolite levels explained by significantly associated SNPs, with the heritability as estimated in a monozygotic twin sample from the NTR cohort ($N = 181$ pairs; see Fig. 3). Among all metabolites, the largest proportion of phenotypic variance was explained for lysoPC a C20:4 (19%); corresponding with 10% of the phenotypic variance), which was associated with a SNP in the FADS1-3 locus. The results of polygenic score (PGS) analyses (see Supplementary Fig. 5 and Supplementary Note 1) suggest different genetic background of variation in serum levels for different metabolites, ranging from close-to-monogenic to highly polygenic.

**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 gene and that displayed genome- and metabolome-wide associations with SNPs in stage 1 GWA meta-analysis in the current study (NATURE COMMUNICATIONS | 6:7208 | DOI: 10.1038/ncomms8208 | www.nature.com/naturecommunications).

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 metabolites tend to match the known function of these genes. SNP rs7582179 in the AGPS gene is associated with the choline plasmalogens PC ae C44:5. Mutations in AGPS (encoding the enzyme alkylglycerone phosphate 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 alkylglycerone phosphate synthase in peroxisomes. The association in the current study of a SNP within the AGPS gene with the serum concentration of the choline plasmalogens 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 AGPS 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 treatment response17,18.

Table 1 | Novel SNPs for serum metabolite levels identified in the current study.

| Lead metabolomic SNP | Lead metabolite | Cytoband | P value in discovery phase | EA/NEA | EAF in discovery phase | Beta in discovery phase | Total N in discovery phase | Nearest gene | Beta in replication phase | P value in replication phase | EAF in replication phase | Replicated | Locus name |
|----------------------|----------------|----------|---------------------------|--------|------------------------|------------------------|---------------------------|-------------|--------------------------|---------------------------|--------------------------|-------------|-----------|
| rs2210538             | C18:2          | 6q21     | 5.03 x 10^-11             | A/G    | 80.8%                  | 0.086                  | 6,574                     | SLC22A16    | 0.099                    | 2.65 x 10^-13             | 75.9%                    | *           | SLC22A16  |
| rs7657817             | Orn            | 6q23.2   | 1.32 x 10^-11             | T/C    | 98.0%                  | -0.156                 | 2,991                     | ARGI        | -0.123                   | 1.40 x 10^-4             | 97.5%                    | *           | ARGI      |
| rs7236012             | Orn            | 6q23.2   | 6.4 x 10^-12              | T/C    | 83.9%                  | 0.045                  | 7,476                     | ARGI        | -0.044                   | 1.57 x 10^-3             | 84.5%                    | *           | ARG1      |
| rs7582179             | C18:0          | 2q31.2   | 4.07 x 10^-10             | A/G    | 16.8%                  | -0.048                 | 5,360                     | AGPS        | -0.021                   | 0.147                     | 16.5%                    | *           | AGPS      |
| rs7700133             | PC ae C44:5    | 4q35.1   | 3.35 x 10^-11             | T/C    | 30.5%                  | 0.036                  | 7,476                     | CENPU       | 0.038                    | 1.12 x 10^-3             | 30.9%                    | *           | ACSL1     |

Abbreviations: EA/NEA, effect allele/non-effect allele. EAF, frequency of EA; SNP, single-nucleotide polymorphism.

Lead metabolite, metabolite 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 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.

Figure 3 | Decomposition of variation in serum metabolite levels. This figure displays the proportions of variance in serum metabolite level explained by significantly associated SNPs; heritability not explained by significantly associated SNPs and unexplained (environmental) variance. Seventy-six metabolites are included for which both heritability estimates (monozygotic twin correlations taken from reference13; N = 181 pairs; Pearson correlation) were available, and that displayed genome- and metabolome-wide associations with SNPs in stage 1 GWA meta-analysis in the current study (N = [1,588, 7,478]). Proportion of variance explained by significantly associated SNPs was estimated as Pearson’s phi coefficient squared. Metabolites are grouped according to biochemical class.
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. 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 ARGL 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. 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. In this context, it is interesting to note that the large proportion of explained heritability we observed for the large proportion of explained heritability we observed for lysoPC aC20:4 (19%) was caused exclusively by the association between a SNP and the concentration of a metabolite, based on the approximate distribution. Throughout the manuscript, names of lipids detected by the Biocrates AbsoluteIDQ p150 platform are abbreviated as follows: acylcarnitines, Cx:y; hydroxyacylcarnitines, (OH)Cx:y; diacylglycerolylcarboxylates, Ca:y:DC; sphingomyelins, SmCx:y; N-hydroxyacyl sphingolipids, SCx:y; phosphatidylcholines, PC (aa = dacyl, a = acyl-alkyl). 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 phenotypes, where the standard deviation (s.d.) 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 metabolite concentration data for each cohort were compared with the large proportion of explained heritability we observed for

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; ECUT, 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; LLS, Medical Ethical Committee of the Leiden University Medical Center; ERM, Human Research Ethics Committee, counterpart 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 metabolomic 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). The method of AbsoluteIDQ p150 kit has been proven to be in accordance with FDA Guideline ‘Guidance for Industry—Bioanalytical Method Validation (May 2001)’, which implies proof of reproducibility within a given error range. For all cohorts, metabolite measurements were carried out at the Metabolics Platform of the Genome Analysis Center at the Helmholtz Zentrum München, Germany as per the manufacturer’s instructions. 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 micromolar. 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 day of the quality control of the Biocrates metabolite concentration measurement data was performed by each participating cohort as follows: 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), was calculated as: where the standard deviation (s.d.) 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 large proportion of explained heritability we observed for

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 phenotypes, where the standard deviation (s.d.) 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 large proportion of explained heritability we observed for

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 phenotypes, where the standard deviation (s.d.) 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 large proportion of explained heritability we observed for

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 phenotypes, where the standard deviation (s.d.) 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 large proportion of explained heritability we observed for

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 phenotypes, where the standard deviation (s.d.) 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 large proportion of explained heritability we observed for

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 phenotypes, where the standard deviation (s.d.) 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 large proportion of explained heritability we observed for

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 phenotypes, where the standard deviation (s.d.) 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 large proportion of explained heritability we observed for

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 phenotypes, where the standard deviation (s.d.) 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 large proportion of explained heritability we observed for
used the monozygotic twin correlations for each metabolite, based on data from 181 pairs from the NTR cohort. 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, 0.01 and so on). In the target sample, PGSs are 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 analysis 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, PGSs 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 analysis below the following thresholds were included: $P < 1.0 \times 10^{-8}$, $P < 1.0 \times 10^{-7}$, $P < 1.0 \times 10^{-6}$, $P < 1.0 \times 10^{-5}$, $P < 1.0 \times 10^{-4}$, $P < 1.0 \times 10^{-3}$, $P < 5.0 \times 10^{-2}$, $P < 0.1$, $P < 0.2$, $P < 0.3$, $P < 0.4$, $P < 0.5$, $P < 0.6$, $P < 0.7$, $P < 0.8$, $P < 0.9$, and $P < 1.0$. For each lead SNP 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 $\leq$ MAF $< 0.50$) were excluded because these SNPs are potentially ambiguous and therefore may lead to spurious association in the case of strand flips. From the imputed SNP genotype data for the KORA S4 target sample, the 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.

$$y = \beta_0 + \beta_1 \text{PGS} + \beta_2 \text{age} + \beta_3 \text{sex} + \beta_4 \text{study}$$

where $y$ represents log(metabolite) values, and the study-specific covariates adjusts include education for, for example, population stratification (and thus $\beta_4$ can be a vector). The Biocartes metabolite values were obtained and preprocessed using the same methods as described for the primary GWA in the Section ‘Biocartes metabolite quantification’. The proportion of variance explained by the PGSs was assessed by computing the raw $R^2$ values for the full model (that is, a model including the genetic score as a covariate) with the raw $R^2$ values when fitting a ‘reduced’ model that did not include the genetic score as a covariate. The significance of the association of the PGSs with serum metabolite levels was estimated on the basis of the $P$ value of association for $\beta_1$ according to the full model.

**EQT analysis.** 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. For each lead metabolomic SNP, we retrieved all SNPs with $R^2 > 0.8$ in the 1000 Genomes Project pilot phase (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.tweelingregister.org/engagiebiocartesgwa

References

1. Suhr, K. & Geiger, C. Genetic variation in metabolic phenotypes: study designs and applications. *Nat. Rev. Genet.* 13, 759–769 (2012).
2. Suhr, K. et al. Human metabolic individuality in biomedical and pharmaceutical research. *Nature* 477, 54–60 (2011).
3. Nicholson, G. et al. A genome-wide metabolic QTL analysis in Europeans implicates two loci shaped by recent positive selection. PLoS Genet. 7, e1002270 (2011).

4. Illig, T. et al. A genome-wide perspective of genetic variation in human metabolism. Nat. Genet. 42, 137–141 (2010).

5. Gieger, C. et al. Genetics meets metabolomics: a genome-wide association study of metabolic profile in humans. PLoS Genet. 4, e1000282 (2008).

6. Demirkan, A. et al. Genome-wide association study identifies novel loci associated with circulating phospho- and sphingolipid concentrations. PLoS Genet. 10, e1002974 (2014).

7. Hicks, A. A. et al. Genetic determinants of circulating sphingolipid concentrations in European populations. PLoS Genet. 5, e1000672 (2009).

8. Tukiainen, T. et al. Detailed metabolic and genetic characterization reveals new associations for 30 known lipid loci. Hum. Mol. Genet. 21, 1444–1455 (2012).

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

10. 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).

11. Shin, S. Y. et al. An atlas of genetic influences on human blood metabolites. Nat. Genet. 46, 543–550 (2014).

12. Yu, B. et al. Genetic determinants influencing human serum metabolome among African Americans. PLoS Genet. 10, e1004212 (2014).

13. Draisma, H. H. et al. Familial resemblance for serum metabolite concentrations. Twin Res. Hum. Genet. 16, 948–961 (2013).

14. Itzkovitz, B. et al. Functional characterization of novel mutations in GNPAT and AGPS, causing rhizomelic chondrodysplasia punctata (RCDP) types 2 and 3. J. Med. Genet. 43, 183–197 (2012).

15. Gertow, K. et al. Fatty acid handling protein expression in adipose tissue, fatty acid composition of adipose tissue and serum, and markers of insulin resistance. Eur. J. Clin. Nutr. 60, 1406–1413 (2006).

16. Phillips, C. M. et al. Gene-nutrient interactions with dietary fat modulate the association between genetic variation of the ACSL1 gene and metabolic syndrome. J. Lipid. Res. 51, 1793–1800 (2010).

17. Aouida, M., Poulin, R. & Ramotar, D. The human carotid transporter SLC22A16 mediates high affinity uptake of the anticancer polyamine analogue bleomycin-A5. J. Biol. Chem. 285, 6275–6284 (2010).

18. Bray, J. et al. Influence of pharmacogenetics on response and toxicity in breast cancer patients treated with doxorubicin and cyclophosphamide. Br. J. Cancer 104, 1089–1099 (2011).

19. Tang, W. H., Wang, Z., Cho, L., Brennan, D. M. & Hazen, S. L. Diminished global arginine bioavailability and increased arginine catabolism as metabolic profile of increased cardiovascular risk. J. Am. Coll. Cardiol. 53, 2061–2067 (2009).

20. Tripolt, N. J. et al. Multifactorial risk factor intervention in patients with Type 2 diabetes improves arginine bioavailability ratios. Diabet. Med. 29, e365–e368 (2012).

21. Brion, M.-J. A., Benyamin, B., Visscher, P. M. & Davey Smith, G. Beyond the single SNP: emerging developments in Mendelian randomization in the ‘omics’ era. Curr. Epidemiol. Rep. 1, 228–236 (2014).

22. Schunkert, H. et al. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. Nat. Genet. 43, 333–338 (2011).

23. Dhuriri, H. et al. Genetics of the human metabolome, what is next? Biochem. Biophys. Acta 1842, 1923–1931 (2014).

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

25. US Department of Health and Human Services, F. a. D. A., Center for Drug Evaluation and Research (CDER), Center for Veterinary Medicine (CVM). Guidance for Industry. Bioanalytical Method Validation. (2001).

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

27. Goek, O. N. et al. Serum metabolite concentrations and decreased GFR in elderly cohorts. Br. J. Cancer 104, 1093–1099 (2011).

28. Berndt, S. I. et al. Genetic risk profiles for depression and anxiety in adult and elderly cohorts. Mol. Psychiatry 16, 773–783 (2011).

29. Burcill, S. M. et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 460, 748–752 (2009).

30. Demirkan, A. et al. Genetic risk profiles for depression and anxiety in adult and elderly cohorts. Mol. Psychiatry 16, 773–783 (2011).

31. Berendes, H. et al. Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. Nat. Genet. 45, 501–512 (2013).

32. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13, 2498–2504 (2003).
Author contributions
Jointly supervised the research: A.I., J.T.B., M.M., B.W.v.D., P.A.C.H., K. Strauch, G.J.B.v.O., T.I., J.T.B., M.M., M.I.M., C.G., A.I. and D.I.B. Contributed reagents/materials/analysis tools: N.A., A.J.M.C., P.E.S., H.-J.W., L.F., J.J.H., G.W. and D.R.N. Conceived and designed the experiments: C.M.v.D., C.G. and D.I.B. Performed statistical analysis: E.M.v.L., S.B., H.H.M.D., R.P., M.K., R.J., A.-K.P. and J.B.v.K. Performing metabolomics measurements (sample preparation, mass spectrometric measurements, quality assurance): C.P. and J.A. Measurements and data quality control: W.R. M. Wrote the paper: A.D., C.M.v.D., H.H.M.D., R.P., R.J., A.-K.P., J.B.v.K., E.I.C.G., D.R.N., K.W.v.D., P.A.C.H., K.Suhre, C.M.v.D. and D.I.B. Principal investigator: T.D.S., A.M., P.E.S., K.W.v.D., P.A.C.H., K.Strauch, N.G.M., G.J.B.v.O., T.I., C.M.v.D. and D.I.B. Analysed data: I.Y., A.A.M.V., M.B., H.H.M.D., R.P., M.K., R.J., A.-K.P., T.H., G.Z., A.D., T.E., J.B.v.K., H.D. and K.Suhre. Genotyping: T.E. Directed wet lab: G.W.M. Directed biochemistry: J.B.W. Managed samples: A.K.H.

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/

How to cite this article: 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 doi: 10.1038/ncomms8208 (2015).