nature genetics

# A genome-wide perspective of genetic variation in human metabolism

Thomas Illig[1,13], Christian Gieger[1,13], Guangju Zhai[2], Werner Römisch-Margl[3], Rui Wang-Sattler[1], Cornelia Prehn[4], Elisabeth Altmaier[3,5], Gabi Kastenmüller[3], Bernet S Kato[2], Hans-Werner Mewes[3,6], Thomas Meitinger[7,8], Martin Hrabé de Angelis[4,9], Florian Kronenberg[10], Nicole Soranzo[2,11], H-Erich Wichmann[1,12], Tim D Spector[2], Jerzy Adamski[4,9] & Karsten Suhre[3,5]



**Serum metabolite concentrations provide a direct readout of biological processes in the human body, and they are associated with disorders such as cardiovascular and metabolic diseases. We present a genome-wide association study (GWAS) of 163 metabolic traits measured in human blood from 1,809 participants from the KORA population, with replication in 422 participants of the TwinsUK cohort. For eight out of nine replicated loci (*FADS1*, *ELOVL2*, *ACADS*, *ACADM*, *ACADL*, *SPTLC3*, *ETFDH* and *SLC16A9*), the genetic variant is located in or near genes encoding enzymes or solute carriers whose functions match the associating metabolic traits. In our study, the use of metabolite concentration ratios as proxies for enzymatic reaction rates reduced the variance and yielded robust statistical associations with *P* values ranging from $3 \times 10^{-24}$ to $6.5 \times 10^{-179}$. These loci explained 5.6%–36.3% of the observed variance in metabolite concentrations. For several loci, associations with clinically relevant parameters have been reported previously.**

We have previously identified frequent genetic polymorphisms with large effects that alter an individual's metabolic capacities[1]. In that study we described genetic variants in metabolism-related genes that lead to specific and clearly differentiated metabolic phenotypes, which we call 'genetically determined metabotypes'. Knowledge of these genetically determined metabotypes in human populations is central to identifying the contributions and interactions of genetic and environmental factors in the etiology of complex diseases, providing a new paradigm for the study of gene-environment interactions. However, our original GWAS was limited in power because of its modest number of participants ($n = 284$). In an effort to identify new major genetically determined metabotypes of biomedical relevance, we conducted a GWAS to metabolic traits in human serum using cohorts of larger sample sizes.

We genotyped KORA samples using the Affymetrix 6.0 GeneChip array and TwinsUK samples using the Illumina Hap317K chip. We also determined fasting serum concentrations of 163 metabolites, covering a biologically relevant panel of amino acids, sugars, acylcarnitines and phospholipids, using electrospray ionization tandem mass spectrometry (ESI-MS/MS) with the Biocrates AbsoluteIDQ targeted metabolomics technology. A full list of the measured metabolites, the abbreviations used to denote them in this paper, and their biological roles is presented in the Online Methods section. Motivated by our previous finding that use of metabolite concentration ratios as proxies for enzymatic reaction rates reduces the variance and yields robust statistical associations[1,2], we tested all of the 163 metabolite concentrations and also all possible metabolite concentration ratios ($163 \times 162 = 26{,}406$ traits) with a linear additive model for association with all single-nucleotide polymorphisms (SNPs) that passed our selection criteria. The corresponding estimated genome-wide significance level after correction for testing 517,480 SNPs (minor allele frequency (MAF) > 10%) and 26,406 multiple metabolic trait combinations is $P = 3.64 \times 10^{-12}$ (see Online Methods). This hypothesis-free approach highlights pairs of metabolites that are more likely to be coupled either biochemically or physiologically.

We applied a two-step discovery design in the KORA F4 population, followed by a replication step in the TwinsUK cohort. Starting with an initial discovery step based on samples of 1,029 male and female individuals of Southern German origin from the KORA F4 population, we selected all loci with *P* values of association $<10^{-7}$ for metabolite concentrations and $P < 10^{-9}$ for concentration ratios in a GWAS (**Fig. 1**);

[1]Institute of Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany. [2]Department of Twin Research and Genetic Epidemiology, King's College London, London, UK. [3]Institute of Bioinformatics and Systems Biology and [4]Institute of Experimental Genetics, Genome Analysis Center, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany. [5]Faculty of Biology, Ludwig-Maximilians-Universität, Planegg-Martinsried, Germany. [6]Department of Genome Oriented Bioinformatics, Life and Food Science Center Weihenstephan, Technische Universität München, Freising-Weihenstephan, Germany. [7]Institute of Human Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany. [8]Institute of Human Genetics, Klinikum rechts der Isar, Technische Universität München, Munich, Germany. [9]Institute of Experimental Genetics, Life and Food Science Center Weihenstephan, Technische Universität München, Freising-Weihenstephan, Germany. [10]Division of Genetic Epidemiology, Department of Medical Genetics, Molecular and Clinical Pharmacology, Innsbruck Medical University, Innsbruck, Austria. [11]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK. [12]Institute of Medical Informatics, Biometry and Epidemiology, Ludwig-Maximilians-Universität and Klinikum Grosshadern, Munich, Germany. [13]These authors contributed equally to this work. Correspondence should be addressed to K.S. (karsten.suhre@helmholtz-muenchen.de).

Received 22 May; accepted 10 November; published online 27 December 2009; doi:10.1038/ng.507

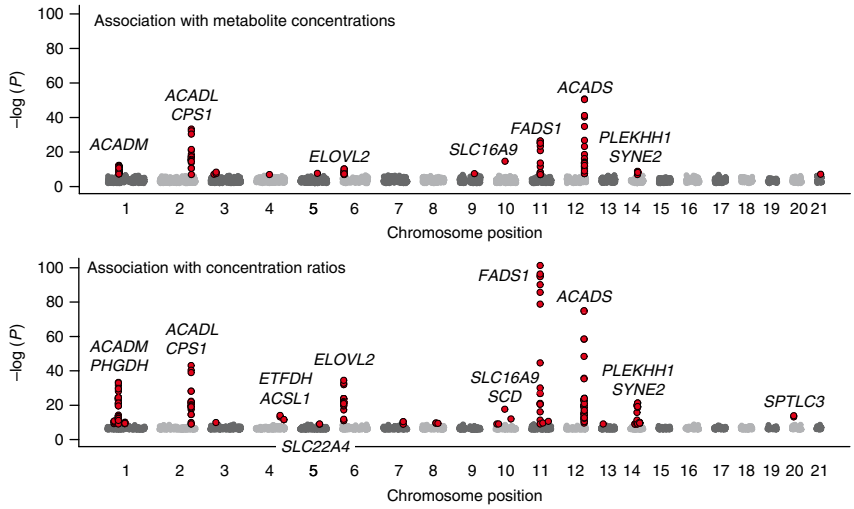


**Figure 1** Manhattan plot of the strength of association with metabolite concentrations (above, data points with $P < 10^{-7}$ are plotted in red) and concentration ratios (below, data points with $P < 10^{-9}$ are plotted in red), based on association with 1,029 samples (step 1 of discovery stage). For each SNP, only the metabolic trait with the lowest $P$ value of association is shown; thus, multiple dots indicate that several SNPs support the association at that locus.

32 loci satisfied these criteria. We then tested one SNP for each locus in a second step in an independent sample of 780 participants selected from the remaining KORA F4 population, using identical genotyping and metabolomics techniques as in the first step. The metabolomics and genotyping experiments for this second step were conducted independently and after completion of the initial study, with an interval of several months. Using data from all 1,809 individuals from the KORA group, we computed joint $P$ values of association. Although this approach is less well powered than a full genome-wide joint analysis, it reflects the historical way in which we selected SNPs for follow-up. The top $P$ values from the discovery step are presented in **Supplementary Table 1**, and a full list is available on request.

We identified 15 loci for which the strength of association increased when additional data were added, and we selected only those for further investigation. All 15 loci had genome-wide significant $P$ values of association that were smaller than $3.64 \times 10^{-12}$ in this joint analysis (**Table 1**; local association plots, box plots by genotype and quantile-quantile plots are presented in **Supplementary Figs. 1–3**). In a third step, for replication in an independent population, we used metabolomics data, measured on our platform, from serum samples of 422 female participants of the TwinsUK cohort. Of the 15 loci tested, 9 were replicated ($P < 0.05$) after Bonferroni correction for 15 tests. Five loci showed signals of association with similar effect-size estimates, but their significance was measured above our threshold, and thus they should be considered as unreplicated. However, this suggestive evidence of an association is supported for four of those five loci (*CPS1*, *SCD*, *SLC22A4* and *PHGDH*) by biological evidence, and two of them (*CPS1* and *SLC22A4*) represent indirect replications of previous studies (**Supplementary Note**). Note also that the TwinsUK study had only limited power because of its smaller sample size. Moreover, all SNPs but one were imputed, and the signal-to-noise ratio in the TwinsUK metabolomics data was about 20% higher than that in the KORA data.

When the functional roles of the genes in these loci are considered, we can draw the most comprehensive view to date of genetic variation in human metabolism. The connections between these genes in the human metabolism are outlined in **Figure 2**. For eight of nine fully replicated genetic polymorphisms, and also for four of the five suggestive loci, the genetic variant is located in or near enzyme-encoding or solute carrier–encoding genes for which the associating metabolic traits match the proteins' function. Many of these polymorphisms affect proteins involved in rate-limiting steps of important enzymatic reactions. Four loci (*FADS1*, *ACADS*, *ACADM* and *ELOVL2*) replicate associations from previous studies, including our own. Two of the suggestive loci (*SLC22A4* and *CPS1*) replicate previous studies on related metabolic traits. For some loci (for example, *SLC16A9* and *PLEKHH1*), new hypotheses on the genes' function can be derived from the associating metabolite pattern. For three loci, an association with clinical end points has previously been reported (*SLC22A4* with Crohn's disease, *FADS1* with hyperactivity and cholesterol and triglyceride levels and *ACADS* as a susceptibility locus for ethylmalonic aciduria).

For several other loci (for example, *ACADM*, *ACADL* and *ETFDH*), loss of function of the corresponding gene leads to severe disorders, indicating that the genetic variants we identify here, or variants in linkage disequilibrium, may induce a related but probably more moderate phenotype. This is in line with findings of a recent GWAS on kidney function that identified *UMOD* to be associated with glomerular filtration rate[3]. Rare mutations in the *UMOD* gene are known to be the cause of monogenic autosomal dominant kidney diseases. Common mutations in the same gene region can be the cause of disease-related phenotypes of less severity on the population level. As discussed previously[1], a ratio between the concentrations of two metabolites that are linked to a substrate-product pair of some enzymatic reaction may constitute an approximation of the conversion rate of that reaction. From the effect size of the association ($\beta'$ in the linear model), one can therefore derive the per-allele difference in metabolic capacities of an individual with respect to the enzymatic reaction considered. For instance, $\beta'$ of the association of rs211718 (*ACADM*) with C12/C10 is 0.12. Assuming an additive-per-copy effect (**Supplementary Fig. 2**), this implies that individuals who are homozygotes of the major allele of *ACADM* burn fatty acids with a chain length of 12 carbons about 24% faster than do carriers of two copies of the minor allele. Similar arguments hold for the other loci ($\beta'$ for all loci is reported in **Supplementary Table 2**).

The SNPs identified in this study can now be examined in GWAS with clinical parameters. As an example, in our previous study[1] we suggested *FADS1* to be a risk locus for perturbed blood lipid parameters. This was supported by the observed association with different phospholipids and the fact that two published GWAS investigating lipid levels reported $P$ values of association for the *FADS1* locus with levels of low-density lipoprotein (LDL), high-density lipoprotein (HDL) and total cholesterol that ranged between $1.89 \times 10^{-4}$ and $6.07 \times 10^{-5}$ (refs. 4,5). These associations had not been included in the list of potential candidates for replication in those studies, as their $P$ values taken alone were not sufficiently small in the context of a classical GWAS. Three better-powered GWAS with lipid parameters have only recently confirmed this prediction[6–8], thereby proving that a combination of a GWAS using metabolomic phenotypic traits with data from previous GWAS can identify new candidate SNPs associated with known phenotypes of clinical relevance. To facilitate further

**Table 1 Summary of the 15 loci identified in the KORA population**

| SNP | SNP type | Locus | Chr | Coded/ noncoded allele | Position | MAF | Strongest association | *n* KORA | β′ KORA | $\eta^2$ KORA | *P* value KORA | *n* TwinsUK | β′ TwinsUK | *P* value TwinsUK | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| rs174547 | Intronic | *FADS1* | 11 | T/C | 61,327,359 | 30.4% | PC aa C36:3 / PC aa C36:4 | 1,806 | 0.151 | 36.3% | $6.5 \times 10^{-179}$ | 422 | 0.156 | $2.0 \times 10^{-26}$ | * |
| rs2014355 | Intronic | *ACADS* | 12 | T/C | 119,659,907 | 27.7% | C3 / C4 | 1,790 | –0.218 | 21.5% | $5.1 \times 10^{-96}$ | 416 | –0.229 | $2.1 \times 10^{-24}$ | * |
| rs211718 | Upstream | *ACADM* | 1 | C/T | 75,879,263 | 30.5% | C12 / C10 | 1,804 | 0.120 | 14.6% | $1.3 \times 10^{-63}$ | 299 | 0.133 | $2.8 \times 10^{-11}$ | * |
| rs2286963 | Coding | *ACADL* | 2 | T/G | 210,768,295 | 36.5% | C9 / C10:2 | 1,806 | 0.219 | 13.8% | $3.1 \times 10^{-60}$ | 421 | 0.312 | $2.4 \times 10^{-19}$ | * |
| rs9393903 | Intronic | *ELOVL2* | 6 | G/A | 11,150,895 | 24.6% | PC aa C40:3 / PC aa C42:5 | 1,803 | 0.087 | 9.8% | $2.3 \times 10^{-42}$ | 419 | 0.076 | $4.0 \times 10^{-5}$ | * |
| rs2216405 | Downstream | *CPS1* | 2 | A/G | 211,325,139 | 18.5% | Glycine / PC ae C38:2 | 1,792 | 0.129 | 7.1% | $1.9 \times 10^{-30}$ | 420 | 0.094 | 0.014 | |
| rs7156144 | Upstream | *PLEKHH1* | 14 | G/A | 67,049,466 | 41.4% | PC ae C32:1 / PC ae C34:1 | 1,799 | –0.042 | 6.6% | $1.7 \times 10^{-28}$ | 405 | –0.023 | 0.0024 | * |
| rs11158519 | Intronic | *SYNE2* | 14 | G/A | 63,434,338 | 14.5% | PC ae C38:1 / PC aa C28:1 | 1,763 | –0.083 | 6.5% | $1.5 \times 10^{-27}$ | 394 | –0.098 | 0.0075 | |
| rs168622 | Upstream | *SPTLC3* | 20 | G/T | 12,914,089 | 37.5% | SM (OH) C24:1 / SM C16:0 | 1,796 | 0.061 | 5.8% | $5.2 \times 10^{-25}$ | 421 | 0.061 | $2.9 \times 10^{-4}$ | * |
| rs8396 | Downstream | *ETFDH* | 4 | T/C | 159,850,267 | 29.8% | C14:1-OH / C10 | 1,778 | 0.102 | 5.6% | $3.5 \times 10^{-24}$ | 421 | 0.114 | $3.1 \times 10^{-7}$ | * |
| rs7094971 | Intronic | *SLC16A9* | 10 | A/G | 61,119,570 | 13.5% | C0 | 1,786 | –0.091 | 4.6% | $3.8 \times 10^{-20}$ | 421 | –0.089 | $9.3 \times 10^{-4}$ | * |
| rs2046813 | Upstream | *ACSL1* | 4 | T/C | 186,006,153 | 32.2% | PC ae C44:5 / PC ae C42:5 | 1,804 | 0.033 | 4.1% | $3.6 \times 10^{-18}$ | 409 | 0.002 | 0.91 | |
| rs603424 | Upstream | *SCD* | 10 | G/A | 102,065,469 | 19.4% | C14 / C16:1 | 1,805 | 0.054 | 4.0% | $1.5 \times 10^{-17}$ | 422 | 0.023 | 0.10 | |
| rs272889 | Intronic | *SLC22A4* | 5 | G/A | 131,693,277 | 38.5% | Valine / C5 | 1,809 | –0.075 | 3.3% | $7.9 \times 10^{-15}$ | 422 | –0.052 | 0.0098 | |
| rs541503 | Upstream | *PHGDH* | 1 | T/C | 120,009,820 | 37.9% | Ornithine /Serine | 1,809 | 0.058 | 2.7% | $3.0 \times 10^{-12}$ | 419 | 0.029 | 0.087 | |

Loci that were replicated in the TwinsUK cohort ($P < 0.05$ after correction for 15 tests) are indicated by *. β′ is the estimation of the effect per allele copy, normalized by the mean of the metabolic trait. $\eta^2$ is the proportion of the variance that can be explained by the genotype in the linear model. Gene variants are reported on the positive strand of NCBI build 36. Additional information on the 50 strongest-associating metabolic traits for each locus is provided in **Supplementary Table 2**. Additional *P* values computed using log-scaled metabolite concentration ratios and also from the first discovery step alone are given in **Supplementary Table 3**. Abbreviations for metabolite names are as follows: PC, phosphatidylcholine; aa, acyl-acyl-PC; ae, acyl-alkyl-PC; a, *lyso*-PC; SM, sphingomyelin; C, acylcarnitine. Compositions of fatty acid side chains are denoted as follows: *x*:*y*, where *x* is the number of carbons and *y* is the number of double bonds; DC, with decarboxyl group; OH, with hydroxyl group (OH); M, with methyl group. A full list of the metabolites determined in this study is provided in the Online Methods section. A detailed discussion of the individual loci is provided as a **Supplementary Note**.

studies, we provide a list of all high-scoring associations of this study to allow similar use by other consortia (**Supplementary Table 1**).

The association data presented in this study can be used to learn more about previously reported associations to related traits. Using the *Catalog of Published Genome-Wide Association Studies* (ref. 9, see URLs; accessed 14 April 2009), we identified three such loci that also associate with metabolic traits in our study (**Table 2**). For instance, SNP rs964184 in the apolipoprotein cluster *APOA1-APOC3-APOA4-APOA5* associates strongly with blood triglyceride levels ($P < 10^{-60}$)[4]. We find that the same SNP associates with ratios between different phosphatidylcholines (for example, PC aa C36:2/PC aa 38:1 (see Online Methods for an explanation of metabolite denotation); $P = 1.8 \times 10^{-10}$), which are biochemically connected to triglycerides by the intermediary of only a few enzymatic reaction steps. SNP rs1260326 (a polymorphism resulting in a P446L substitution) in *GCKR*, which encodes the glucose kinase regulator protein, inversely modulates fasting glucose ($P = 8 \times 10^{-13}$) and triglyceride levels ($P = 1 \times 10^{-4}$) and reduces type 2 diabetes risk in the DESIR prospective general French population[10]. This locus associates with different ratios between plasmalogens and phosphatidylcholines (for example,

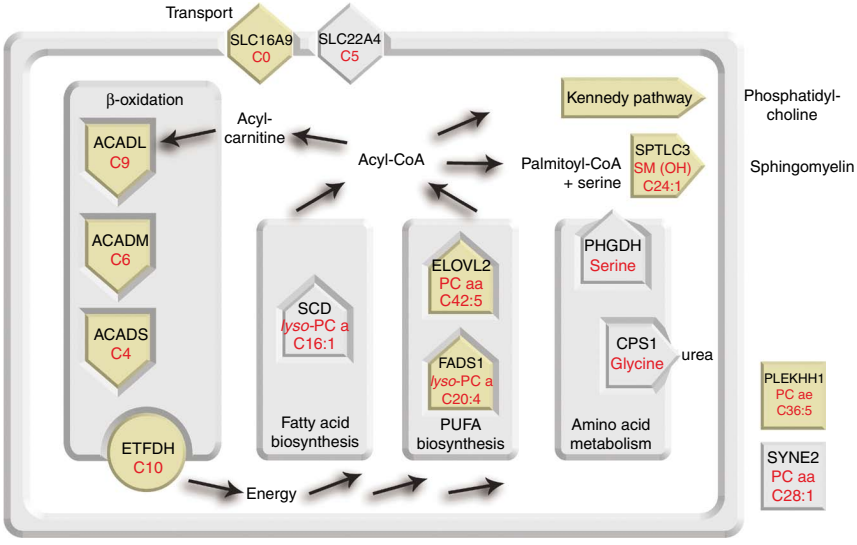


**Figure 2** A systemic view of genetic variation in human metabolism, as identified in this study. Eight of nine replicated genetic polymorphisms (beige) and also four of five suggestive loci (gray) are located in or near genes encoding enzymes that are central to the different processes in human lipid metabolism, including β-oxidation (*ACADS*, *ACADM* and *ACADL*), polyunsaturated fatty acid biosynthesis (*FADS1* and *ELOVL2*), fatty acid synthesis (*SCD*), breakdown of fats and proteins to energy (*ETFDH*) and biosynthesis of phospholipids (*SPTLC3*). Two SNPs are located in or near genes encoding carrier proteins (*SLC22A4* and *SLC16A9*), and two SNPs involve enzymes that are related to amino acid metabolism (*PHGDH* and *CPS1*). Only for two genetic variants does the attribution of a metabolic function remain elusive (*PLEKHH1* and *SYNE2*). For each locus, the most strongly associating single metabolite is indicated in red.

**Table 2 Summary of loci associated with metabolite traits that were replicated in the TwinsUK study**

| Locus | Gene function | Metabolites | Notes | Reference |
|---|---|---|---|---|
| *FADS1* | Fatty acid desaturase 1 | *Lyso*-PC a C20:4, PC aa C38:4, PC aa C36:4 | Risk locus for high cholesterol[6–8]; linked to hyperactivity[23]; covariate in association breastfeeding-IQ[24]; C20:4 fatty acids are products of *FADS1* | Replication of KORA[1] and indirect replication of INCHIANTI[25], which measured selected PUFAs |
| *ACADS* (*SCAD*) | Acyl-coenzyme A dehydrogenase, C2 to C3 short chain | C4 | Predisposition allele that can cause *ACADS* deficiency if additional factors are present[26]; C4 fatty acids are substrates of *ACADS* | Replication of KORA[1] and a target gene study in a US population[27] |
| *ACADM* (*MCAD*) | Acyl-coenzyme A dehydrogenase, C4 to C12 straight chain | C6*, C8, C10, C10:1 | C4 to C12 fatty acids are substrates of *ACADM* | Replication of KORA[1] |
| *ACADL* (*LCAD*) | Acyl-coenzyme A dehydrogenase, long chain | C9 | Coding SNP; patented in a method for diagnosing the risk of thermolabile phenotype in influenza-associated encephalopathy[14]; long-chain fatty acids (including C9) are substrates of *ACADL* | – |
| *ELOVL2* | Elongation of very long chain fatty acids (FEN1/Elo2, SUR4/Elo3, yeast)–like 2 | PC aa C42:5, PC aa C42:6 | PUFA products and/or substrates of *ELOVL2* are incorporated into phosphatidylcholines, such as PC aa C42:5 and PC aa C42:6 | Indirect replication of INCHIANTI[25], which measured selected PUFAs |
| *PLEKHH1* | Pleckstrin homology domain–containing, family H (with MyTH4 domain) member 1 | PC ae C36:5, PC ae C32:2 | Gene function unknown; our results suggest that this locus may be functionally related to the metabolism of plasmalogens, such as PC ae C36:5 and PC ae C32:2 | – |
| *SPTLC3* | Serine palmitoyltransferase, Long-chain base subunit 3 | SM (OH) C24:1 | The SPT complex catalyzes the rate-limiting step of sphingomyelin (SM) biosynthesis; *SPTLC3* is the regulatory subunit; association with ratios between sphingolipids indicates that this polymorphism may modify the cellular mechanism to adjust SPT activity to tissue-specific requirements of sphingolipid synthesis | – |
| *ETFDH* | Electron-transferring flavoprotein dehydrogenase | C10, C8 | Breakdown of fats and proteins to energy (β-oxidation of substrates such as C8 and C10 fatty acids); genetic variant in this locus also involves changes in acylcarnitine hydroxylation and carboxylation | – |
| *SLC16A9* (*MCT9*) | Solute carrier family 16, member 9 (monocarboxylic acid transporter 9) | C0 | Hypothesis: carnitine (C0) is a monocarboxylic acid and may therefore be the substrate of this 'orphan' transporter | – |

Metabolites with the lowest *P* values of association for these loci with respect to metabolite concentrations are reported; note that concentration ratios with the lowest *P* values of association are given in **Table 1**, and data for the 50 strongest associations are provided in **Supplementary Table 2**. Loci that have already been published elsewhere are indicated as 'replication'. The term 'indirect replication' is used when the study in question measures not the same but related metabolites. Commonly used but nonstandard gene names are mentioned in parentheses.



PC ae C34:2/PC aa C32:2; $P = 3.2 \times 10^{-8}$), thereby providing new avenues for further investigation of the functional background of this association. SNP rs10830963 in the gene encoding melatonin receptor (*MTNR1B*) associates with fasting glucose[11]. The same SNP associates in this study with tryptophan/phenylalanine ratios ($P = 5.7 \times 10^{-6}$). This is notable because phenylalanine is a precursor of melatonin, indicating a functional relationship between this pathway and the regulation of glucose homeostasis. We expect the list of loci with parallel association of clinically relevant parameters and metabolic traits to grow as new GWAS become available, and we therefore provide our association data for such use (**Supplementary Table 1**).

The SNPs identified in this study can also be used in clinical studies for association with response to drug treatment. One published example is a common polymorphism in the gene encoding dihydropyrimidine dehydrogenase (*DPYP*) that associates strongly with fluoropyrimidine-related toxicity in cancer patients[12]. Carriers of this variant could benefit from individual dose adjustment of the fluoropyrimidine drug or from alternative therapies. It is now possible to use the here-identified SNPs in association studies with phenotypes that are specific to a disease, such as the development of particular complications during the course of a disease or treatment. One published example is a SNP in the gene encoding carnitine palmitoyltransferase II (*CPT-II*) that is a predisposing factor for influenza-associated encephalopathy (thermolabile phenotype)[13]. Indeed, the SNP that we identified here in *ACADL* has been patented by others[14], along with the *CPT-II* polymorphism, as a method of diagnosing the risk of a thermolabile phenotype.

In summary, this study allowed us to draw a systemic perspective of the genetic variation that is found in human metabolism. In contrast to most GWAS with clinically relevant end points, it seems that for metabolic traits most of the associations are linked to genetic variants in genes with a matching metabolic function (**Fig. 2**). The use of metabolite concentration ratios results in a pronounced sharpening of the association with strongly decreased *P* values when compared to an analysis of single metabolites. Moreover, as we show with examples (*FADS1*, the *APO* cluster, *MTNR1B* and *GCKR*), it allows for comparisons to GWAS with clinically relevant end points. Our study demonstrates the exciting potential of metabolomics to unravel the genetics of human metabolism. The genome-wide perspective on genetic variation in human metabolism presented here will continue to improve as more extensive metabolite panels become available for use on a genome-wide scale, with, among others, additional studies that include nutritional challenges of the participants. We believe that the introduction of metabolomics into the field of molecular epidemiology provides a new hypothesis-driven approach to GWAS.

## METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

*Note: Supplementary information is available on the Nature Genetics website.*

**ACKNOWLEDGMENTS**

The KORA (Kooperative Gesundheitsforschung in der Region Augsburg) research platform and the MONICA (Monitoring trends and determinants on cardiovascular diseases) Augsburg studies were initiated and financed by the Helmholtz Zentrum München–National Research Center for Environmental Health, which is funded by the German Federal Ministry of Education, Science, Research and Technology and by the State of Bavaria. Part of this work was financed by the German National Genome Research Network (NGFNPlus: 01GS0823) and by grants from the 'Genomics of Lipid-associated Disorders–GOLD' of the 'Austrian Genome Research Programme GEN-AU'. Computing resources have been made available by the Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities (HLRB project h1231) and the DEISA Extreme Computing Initiative (project PHAGEDA). Part of this research was supported within the Munich Center of Health Sciences (MC Health) as part of LMUinnovativ. The TwinsUK study was funded by the Wellcome Trust, European Community's Seventh Framework Programme (FP7/2007-2013)/grant agreement HEALTH-F2-2008-201865-GEFOS and (FP7/2007-2013), ENGAGE project grant agreement HEALTH-F4-2007-201413 and the FP-5 GenomEUtwin Project (QLG2-CT-2002-01254). The study also received support from the Department of Health via the UK National Institute for Health Research (NIHR) comprehensive Biomedical Research Centre award to Guy's & St. Thomas' NHS Foundation Trust in partnership with King's College London (T.D.S.).
The project also received support from a UK Biotechnology and Biological Sciences Research Council (BBSRC) project grant (G20234). We acknowledge the funding and support of the US National Eye Institute (NEI) via a US National Institutes of Health (NIH) and Center for Inherited Disease Research (CIDR) genotyping project (principal investigator T. Young). We acknowledge the contributions of P. Lichtner, G. Eckstein, G. Fischer, T. Strom and all other members of the Helmholtz Zentrum München genotyping staff in generating the SNP data set, T. Halex and A. Sabunchi to the metabolomics measurements, and of all members of the field staffs who were involved in the planning and conduct of the MONICA and KORA Augsburg studies. The KORA group consists of H.-E.W. (speaker), A. Peters, C. Meisinger, T.I., R. Holle, J. John and their co-workers who are responsible for the design and conduct of the KORA studies. For the TwinsUK study, we thank the staff from the Genotyping Facilities at the Wellcome Trust Sanger Institute for sample preparation, quality control and genotyping led by L. Peltonen and P. Deloukas, Le Centre National de Génotypage (France), led by M. Lathrop, for genotyping, Duke University, North Carolina, USA, led by D. Goldstein, for genotyping and the Finnish Institute of Molecular Medicine, Finnish Genome Center, University of Helsinki, led by A. Palotie. Genotyping was also performed by the CIDR as part of an NEI and NIH project grant. Finally, we thank all participants of the KORA and the TwinsUK studies.

**AUTHOR CONTRIBUTIONS**

T.I., T.D.S., J.A. and K.S. designed the experiment; T.I., C.G., T.M. and H.-E.W. contributed genetics data and analysis from the KORA study; G.Z., B.S.K., N.S. and T.D.S. contributed genetics data and analysis from the TwinsUK study; W.R.-M., R.W.-S., C.P., G.K., H.-W.M., M.H.d.A., T.D.S., J.A. and K.S. contributed to the metabolomics experiments; C.G., G.Z., E.A. and K.S. analyzed the data; C.G., F.K., N.S. and K.S. wrote the manuscript; all authors contributed their critical reviews of the manuscript during its preparation.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Gieger, C. *et al.* Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. *PLoS Genet.* **4**, e1000282 (2008).
2. Altmaier, E. *et al.* Bioinformatics analysis of targeted metabolomics—uncovering old and new tales of diabetic mice under medication. *Endocrinology* **149**, 3478–3489 (2008).
3. Köttgen, A. *et al.* Multiple loci associated with indices of renal function and chronic kidney disease. *Nat. Genet.* **41**, 712–717 (2009).
4. Kathiresan, S. *et al.* Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nat. Genet.* **40**, 189–197 (2008).
5. Willer, C.J. *et al.* Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat. Genet.* **40**, 161–169 (2008).
6. Aulchenko, Y.S. *et al.* Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. *Nat. Genet.* **41**, 47–55 (2009).
7. Kathiresan, S. *et al.* Common variants at 30 loci contribute to polygenic dyslipidemia. *Nat. Genet.* **41**, 56–65 (2009).
8. Sabatti, C. *et al.* Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat. Genet.* **41**, 35–46 (2009).
9. Hindorff, L.A., Junkins, H.A., Mehta, J.P. & Manolio, T.A. *A Catalog of Published Genome-Wide Association Studies* (Office of Population Genomics, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, USA, accessed 14 April 2009). <http://www.genome.gov/26525384>.
10. Vaxillaire, M. *et al.* The common P446L polymorphism in *GCKR* inversely modulates fasting glucose and triglyceride levels and reduces type 2 diabetes risk in the DESIR prospective general French population. *Diabetes* **57**, 2253–2257 (2008).
11. Prokopenko, I. *et al.* Variants in *MTNR1B* influence fasting glucose levels. *Nat. Genet.* **41**, 77–81 (2009).
12. Gross, E. *et al.* Strong association of a common dihydropyrimidine dehydrogenase gene polymorphism with fluoropyrimidine-related toxicity in cancer patients. *PLoS One* **3**, e4003 (2008).
13. Chen, Y. *et al.* Thermolabile phenotype of carnitine palmitoyltransferase II variations as a predisposing factor for influenza-associated encephalopathy. *FEBS Lett.* **579**, 2040–2044 (2005).
14. Kido, H., Kinoshita, M., Mizuguchi, H. & Takahashi, N. Method of diagnosing the risk of thermolabile phenotype diseases by using gene. Japanese patent PCT/JP2005/021294 (2007).

## ONLINE METHODS

**Study population.** The KORA S4 survey, an independent population-based sample from the general population living in the region of Augsburg, Southern Germany, was conducted in 1999–2001. The standardized examinations applied in the survey (4,261 participants, response 67%) have been described in detail (ref. 15 and the references therein). A total of 3,080 subjects participated in a follow-up examination of S4 in 2006–2008 (KORA F4), comprising individuals who, at that time, were aged 32–81 years. Informed consent has been given. The study has been approved by the local ethical committee. For the first genome-wide screening step, 1,048 blood samples of KORA F4 participants were metabolically characterized. For 1,029 samples (374 males and 406 females) of this group, genome-wide genotype data were also available. In a second step, 972 samples were metabolically characterized in an independent experimental batch. For 780 (374 males and 406 females) of these samples, genome-wide genotype data were available. For the joint analysis in KORA, metabolomics and genotype data for a total of 1,809 individuals were available. No evidence of population stratification has been found in multiple published analyses using the KORA cohort.

**Genotyping and imputation.** In KORA F4, we carried out genotyping using the Affymetrix 6.0 GeneChip array. Analysis in the discovery stage was performed solely on genotyped SNPs. As is current standard for GWAS, we excluded all X chromosome–linked SNPs for the following reasons: (i) the X chromosome must be treated differently from the autosomes; (ii) it cannot be predicted which allele is active; (iii) testing males separately results in different sample sizes and power. Imputation of SNPs in the HapMap CEU population was performed using IMPUTE[16] for use in the regional association plots. In the discovery stage we limited our analysis to SNPs with a moderate-to-high MAF (>10%) and a high genotyping quality (call rate >95%) and with respect to Hardy-Weinberg equilibrium ($P_{HWE} > 0.001$). A total of 517,480 SNPs satisfied all of these criteria.

**Blood sampling.** We collected blood samples for metabolic analysis between 2006 and 2008 in parallel with the KORA F4 examinations. To avoid variation due to circadian rhythm, blood was drawn in the morning between 8:00 a.m. and 10:00 a.m. after a period of overnight fasting. Material was drawn into serum gel tubes, gently inverted two times and then allowed to rest for 30 min at room temperature (18–25 °C) to obtain complete coagulation. The material was then centrifuged for 10 min (2,750*g* at 15 °C). Serum was divided into aliquots and kept for a maximum of 6 h at 4 °C, after which it was deep frozen to −80 °C until analysis.

**Metabolite measurements.** Liquid handling of serum samples (100 µl) was performed with a Hamilton Star (Hamilton Bonaduz AG) robot, and samples were prepared for quantification using the AbsoluteIDQ kit (BIOCRATES Life Sciences AG). Sample analyses were done on API 4000 Q TRAP LC/MS/MS System (Applied Biosystems) equipped with a Schimadzu Prominence LC20AD pump and a SIL-20AC autosampler. The complete analytical process was performed using the MetIQ software package, which is an integral part of the AbsoluteIDQ kit. We did not apply any data correction, nor did we remove any data points. The experimental metabolomics measurement technique is described in detail by US patent US 2007/0004044 (ref. 17; see URLs) and in the manufacturer's manuals. A summary of the method can be found in refs. 18,19, and a comprehensive overview of the field and the related technologies is given in ref. 20. Briefly, a targeted profiling scheme is used to quantitatively screen for known small-molecule metabolites using multiple reaction monitoring, neutral loss and precursor-ion scans. Quantification of the metabolites of the biological sample is achieved by reference to appropriate internal standards. The method has been proven to conform with 21CFR (Code of Federal Regulations) Part 11, which implies proof of reproducibility within a given error range (see **Supplementary Table 4** for data). It has been applied in different academic and industrial applications[1,2,21]. Concentrations of all analyzed metabolites are reported as micromolar concentrations.

**Metabolite panel.** In total, we detected 163 different metabolites (**Supplementary Table 4**). The metabolomics data set contains 14 amino acids, hexose (H1), free carnitine (C0), 40 acylcarnitines (C*x:y*), hydroxylacylcarnitines (C(OH)*x:y*), and dicarboxylacylcarnitines (C*x:y*-DC), 15 sphingomyelins (SM*x:y*) and *N*-hydroxylacyloylsphingosylphosphocholine (SM (OH)*x:y*), 77 phosphatidylcholines (PC, aa = diacyl, ae = acyl-alkyl) and 15 *lyso*-phosphatidylcholines. Lipid side chain composition is abbreviated as C*x:y*, where *x* denotes the number of carbons in the side chain and *y* the number of double bonds. For example, "PC ae C33:1" denotes an acyl-alkyl phosphatidylcholine with 33 carbons in the two fatty acid side chains and a single double bond in one of them. Full biochemical names are provided in **Supplementary Table 4**. The precise position of the double bonds and the distribution of the carbon atoms in different fatty acid side chains cannot be determined with this technology. In some cases, the mapping of metabolite names to individual masses can be ambiguous. For example, stereochemical differences are not always discernible, and neither are isobaric fragments. In such cases, possible alternative assignments are indicated.

**Statistical analysis.** For statistical analysis we included only SNPs with MAF of at least 10% to avoid spurious associations due to small numbers. We used additive genetic models assuming a trend per copy of the minor allele to specify the association between genotype categories and each of the 163 metabolite concentrations, as well as all possible metabolite concentration ratios ($163 \times 162 = 26.406$ traits). No further adjustment was performed. The linear regression algorithm implemented in the statistical analysis system R was used in the GWAS, and SPSS for Windows (Version 17.0, SPSS Inc.) was used for statistical analysis on a case-by-case level. Motivated by our previous observation that the use of ratios may lead to a strong reduction in the overall variance and a corresponding improvement in the *P* values of association[2], we also computed all possible pairs of metabolite concentration ratios for those cases and used those ratios as quantitative traits. We present in the main text the results for the untransformed ratios for ease of interpretation. We used a conservative estimate of a genome-wide significance level (using a Bonferroni correction) that, based on a nominal level of 0.05, is $5.93 \times 10^{-10}$ (0.05) / ($163 \times 517{,}480$). To estimate whether deviation from normality of metabolite ratios may have biased our results, we tested associations for both untransformed and log-scaled ratios; we detected no significant differences (**Supplementary Table 3**). The reported *P* values of step 1 of the discovery stage were not corrected for genomic inflation, as the genomic control inflation factors λ are small, ranging from 1.00 to 1.03, both in KORA and in TwinsUK.

**Replication in the TwinsUK study.** The TwinsUK cohort is an adult twin British registry. These unselected twins were recruited from the general population through national media campaigns in the United Kingdom and were shown to be comparable to age-matched population singletons in terms of disease-related and lifestyle characteristics[22]. Ethics approval was obtained from the Guy's and St. Thomas' Hospital Ethics Committee. Written informed consent was obtained from every participant in the study. We genotyped a total of 2,277 individuals of European ancestry (1,073 singletons and 602 dizygotic twins) from the TwinsUK registry using the Illumina Hap317K chip. We applied a strict quality control at both individual and SNP levels. We excluded 51 individuals because of their non-European ancestry and 3,366 SNPs because of a MAF of <1%, a call rate of <95% if the MAF was <5% or a call rate of <99% if the MAF was >5%, or $P_{HWE} < 1 \times 10^{-4}$. After the quality control, 305,811 autosomal SNPs available from 2,226 individuals (1,046 singletons and 590 dizygotic twins) were available and used for imputation. The imputation was carried out using the IMPUTE software[16]. NCBI build 36 was used for strand reference. We have made available data for 2.5 million autosomal imputed from these 2,226 individuals. We selected 422 unrelated individuals from those genotyped for the metabolomics assay. For the TwinsUK study, blood samples were taken after at least 6 h of fasting. The samples were immediately inverted three times, followed by 40 min resting at 4 °C to obtain complete coagulation. The samples were then centrifuged for 10 min at 2,000*g*. Serum was removed from the centrifuged brown-topped tubes as the top, yellow, translucent layer of liquid. Four aliquots of 1.5 ml were placed into skirted microcentrifuge tubes and then stored in a −45 °C freezer until sampling. Metabolite measurements were performed using the same metabolomics platform and following an identical protocol as for the KORA study at the Genome Analysis Centre of the Helmholtz Zentrum München. For the purpose of the replication for the KORA study, the data on the metabolites, the




doi:10.1038/ng.507

ratios of the concentration and the genotypes were extracted from the available database, and the association between the concentration of the metabolites or the ratios and the corresponding SNPs was tested using a linear regression model in STATA version 10 (StatCorp LP).

**URLs.** R statistical analysis system, http://www.r-project.org/; KORA, http://www.helmholtz-muenchen.de/kora/; TwinsUK, http://www.twinsuk.ac.uk; Online access to patent US 2007/0004044 (describing Biocrates technology), http://www.freepatentsonline.com/20070004044.html.


15. Wichmann, H.E., Gieger, C. & Illig, T. KORA-gen–resource for population genetics, controls and a broad spectrum of disease phenotypes. *Gesundheitswesen* **67** (Suppl. 1), S26–S30 (2005).
16. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39**, 906–913 (2007).
17. Ramsay, S.L., Stoeggl, W.M., Weinberger, K.M., Graber, A. & Guggenbichler, W. Apparatus and method for analyzing a metabolite profile. US Patent 2007/0004044 (2007).
18. Weinberger, K.M. Metabolomics in diagnosing metabolic diseases. *Ther. Umsch.* **65**, 487–491 (2008).
19. Weinberger, K.M. & Graber, A. Using comprehensive metabolomics to identify novel biomarkers. *Screening Trends in Drug Discovery* **6**, 42–45 (2005).
20. Wenk, M.R. The emerging field of lipidomics. *Nat. Rev. Drug Discov.* **4**, 594–610 (2005).
21. Wang-Sattler, R. *et al.* Metabolic profiling reveals distinct variations linked to nicotine consumption in humans—first results from the KORA study. *PLoS One* **3**, e3863 (2008).
22. Andrew, T. *et al.* Are twins and singletons comparable? A study of disease-related and lifestyle characteristics in adult women. *Twin Res.* **4**, 464–477 (2001).
23. Brookes, K.J., Chen, W., Xu, X., Taylor, E. & Asherson, P. Association of fatty acid desaturase genes with attention-deficit/hyperactivity disorder. *Biol. Psychiatry* **60**, 1053–1061 (2006).
24. Caspi, A. *et al.* Moderation of breastfeeding effects on the IQ by genetic variation in fatty acid metabolism. *Proc. Natl. Acad. Sci. USA* **104**, 18860–18865 (2007).
25. Tanaka, T. *et al.* Genome-wide association study of plasma polyunsaturated fatty acids in the InCHIANTI Study. *PLoS Genet.* **5**, e1000338 (2009).
26. Gregersen, N. *et al.* Identification of four new mutations in the short-chain acyl-CoA dehydrogenase (*SCAD*) gene in two patients: one of the variant alleles, 511C→T, is present at an unexpectedly high frequency in the general population, as was the case for 625G→A, together conferring susceptibility to ethylmalonic aciduria. *Hum. Mol. Genet.* **7**, 619–627 (1998).
27. Nagan, N. *et al.* The frequency of short-chain acyl-CoA dehydrogenase gene variants in the US population and correlation with the $C_4$-acylcarnitine concentration in newborn blood spots. *Mol. Genet. Metab.* **78**, 239–246 (2003).