Published in final edited form as:
Nature. ; 477(7362): . doi:10.1038/nature10354.

Human metabolic individuality in biomedical and pharmaceutical research

Karsten Suhre[1,2,3,+], **So-Youn Shin**[#4], **Ann-Kristin Petersen**[#5], **Robert P. Mohney**[6], **David Meredith**[7], **Brigitte Wägele**[1,8], **Elisabeth Altmaier**[1], **CARDIoGRAM**[9], **Panos Deloukas**[4], **Jeanette Erdmann**[10], **Elin Grundberg**[4,11], **Christopher J. Hammond**[11], **Martin Hrabé de Angelis**[12,13], **Gabi Kastenmüller**[1], **Anna Köttgen**[14], **Florian Kronenberg**[15], **Massimo Mangino**[11], **Christa Meisinger**[16], **Thomas Meitinger**[17,18], **Hans-Werner Mewes**[1,8], **Michael V. Milburn**[6], **Cornelia Prehn**[12], **Johannes Raffler**[1,2], **Janina S. Ried**[4], **Werner Römisch-Margl**[1],

[+]corresponding authors: karsten@suhre.fr (KS), ns6@sanger.ac.uk (NS) .

Designed the study: Jerzy Adamski, Christian Gieger, Thomas Illig, David Meredith, Nicole Soranzo, Karsten Suhre
Conducted the experiments: David Meredith, Michael V. Milburn, Robert P. Mohney
Analyzed the data: Jerzy Adamski, Elisabeth Altmaier, Christian Gieger, Gabi Kastenmüller, Anna Köttgen, Florian Kronenberg, Christa Meisinger, David Meredith, Ann-Kristin Petersen, Cornelia Prehn, Johannes Raffler, Janina S. Ried, Werner Römisch-Margl, So-Youn Shin, Karsten Suhre, Brigitte Wägele
Provided material / data / analysis tools: The CARDIoGRAM consortium, Panos Deloukas, Jeanette Erdmann, Elin Grundberg, Christopher J. Hammond, Martin Hrabé de Angelis, Thomas Illig, Massimo Mangino, Thomas Meitinger, Hans-Werner Mewes, Nilesh Samani, Kerrin S. Small, Tim D. Spector, H.-Erich Wichmann, Guangju Zhai
Wrote the paper: Christian Gieger, Nicole Soranzo, Karsten Suhre

DISCLOSURES AND COMPETING FINANCIAL INTERESTS Michael V. Milburn and Robert P. Mohney are employees of Metabolon Inc.

Web links

- GWAS server: http://metabolomics.helmholtz-muenchen.de/gwa/
- SNAP: http://www.broadinstitute.org/mpg/snap/
- NHGRI catalog of published GWAS: http://www.genome.gov/gwastudies/
- eQTL : http://www.sanger.ac.uk/Software/analysis/genevar/
- GRAIL : http://www.broadinstitute.org/mpg/grail/
- IPA : Ingenuity Pathway Analysis : http://ingenuity.com
- OMIM : http://www.ncbi.nlm.nih.gov/omim
- yED network editor : http://www.yworks.com
- BioGPS : http://biogps.gnf.org
- Genecards : http://www.genecards.org
- WikiGenes : http://www.wikigenes.org
- Pharmacogenomics Knowledge Base : http://www.pharmgkb.org
- R statistical analysis system: http://www.r-project.org
- KORA study population : http://www.helmholtz-muenchen.de/kora/
- TwinsUK study : http://www.twinsuk.ac.uk
- Metabolon Inc. : http://www.metabolon.com
- MERLIN: http://www.sph.umich.edu/csg/abecasis/Merlin
- PLINK: http://pngu.mgh.harvard.edu/~purcell/plink
- R: http://www.r-project.org
- SNPTEST: http://www.stats.ox.ac.uk/~marchini/software/gwas/snptest.html

Europe PMC Funders Author Manuscripts

Nilesh J. Samani[19], **Kerrin S. Small**[11], **H.-Erich Wichmann**[20,21,22], **Guangju Zhai**[11], **Thomas Illig**[23], **Tim D. Spector**[11], **Jerzy Adamski**[12], **Nicole Soranzo**[#4,+], and **Christian Gieger**[#5]

[1]Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany

[2]Faculty of Biology, Ludwig-Maximilians-Universität, Planegg-Martinsried, Germany

[3]Department of Physiology and Biophysics, Weill Cornell Medical College in Qatar, Education City - Qatar Foundation, Doha, Qatar

[4]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton UK

[5]Institute of Genetic Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany

[6]Metabolon Inc., Durham, North Carolina, USA

[7]School of Life Sciences, Oxford Brookes University, Headington, Oxford, UK

[8]Department of Genome-oriented Bioinformatics, Life and Food Science Center Weihenstephan, Technische Universität München, Freising-Weihenstephan, Germany

[9]The member list of the CARDIoGRAM consortium is provided as **Supplemental Information**

[10]Universität zu Lübeck, Medizinische Klinik II, Lübeck, Germany

[11]Department of Twin Research & Genetic Epidemiology, King's College London, UK

[12]Institute of Experimental Genetics, Genome Analysis Center, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany

[13]Institute of Experimental Genetics, Life and Food Science Center Weihenstephan, Technische Universität München, Freising-Weihenstephan, Germany

[14]Renal Division, University Hospital Freiburg, Germany

[15]Division of Genetic Epidemiology, Department of Medical Genetics, Molecular and Clinical Pharmacology, Innsbruck Medical University, Innsbruck, Austria

[16]Institute of Epidemiology II, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany

[17]Institute of Human Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany

[18]Institute of Human Genetics, Klinikum rechts der Isar, Technische Universität München, Munich, Germany

[19]Department of Cardiovascular Sciences, University of Leicester, and Leicester NIHR Biomedical Research Unit in Cardiovascular Disease, Glenfield Hospital, Leicester, UK

[20]Institute of Epidemiology I, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany

[21]Institute of Medical Informatics, Biometry and Epidemiology, Chair of Epidemiology, Ludwig-Maximilians-Universität, Munich, Germany

[22]Klinikum Grosshadern, Munich, Germany

[23]Unit for Molecular Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany

[#] These authors contributed equally to this work.

SUMMARY

Genome-wide association studies (GWAS) have identified many risk loci for complex diseases, but effect sizes are typically small and information on the underlying biological processes is often lacking. Associations with metabolic traits as functional intermediates can overcome these problems and potentially inform individualized therapy. Here we report a comprehensive analysis of genotype-dependent metabolic phenotypes using a GWAS with non-targeted metabolomics. We identified 37 genetic loci associated with blood metabolite concentrations, of which 25 exhibit effect sizes that are unusually high for GWAS and account for 10-60% of metabolite levels per allele copy. Our associations provide new functional insights for many disease-related associations that have been reported in previous studies, including cardiovascular and kidney disorders, type 2 diabetes, cancer, gout, venous thromboembolism, and Crohn's disease. Taken together our study advances our knowledge of the genetic basis of metabolic individuality in humans and generates many new hypotheses for biomedical and pharmaceutical research.

Understanding the role of genetic predispositions and their interaction with environmental factors in complex chronic diseases is key in the development of safe and efficient therapies, diagnosis and prevention. Genome-wide association studies (GWAS) have identified hundreds of disease risk loci [1]. However, functional information on the underlying biological processes is often lacking [2]. Previously, we have shown the promise of using associations with blood metabolites as functional intermediate phenotypes, the so-called genetically determined metabotypes (GDMs), to understand the potential relevance of these genetic variants for biomedical and pharmaceutical research [3,4]. Building on this early work, we present here the most comprehensive evaluation of genetic variance in human metabolism to date, combining genetics and metabolomics for hypothesis generation in a GWAS. We used an extensive, non-targeted and metabolome-wide panel of small molecules, analyzing >250 metabolites from 60 biochemical pathways in serum samples of 2,820 individuals from two large population-based European cohorts. We identified 37 genetic loci significant at a stringent genome-wide threshold. In contrast to most GWAS, these loci exhibited exceptionally large effect sizes of 10-60% per allele copy in 25 loci. In the majority of cases a protein biochemically related to the associated metabolic traits is encoded at these loci. As a proof of principle validation of new discoveries, we experimentally validated the predicted function of *SLC16A9* as a carnitine efflux transporter. We further cross-referenced these loci with databases of disease-related and pharmaceutically-relevant genetic associations, uncovering hitherto unknown links and providing new hypotheses into the function of these loci. Finally, we made publically available a knowledge-base resource via a web-server to aid future functional studies and biological as well as clinical interpretation of GWAS findings. In summary, this study provides compelling evidence for novel associations of metabolic traits at a wide range of loci of biomedical and pharmaceutical interest, and suggests a powerful new paradigm for dissecting human metabolic and disease pathways.

METHODS

Metabolic profiling was done on fasting serum from participants of the German KORA F4 study (n=1,768) and the British TwinsUK study (n=1,052) using ultrahigh performance liquid-phase chromatography and gas chromatography separation coupled with tandem mass spectrometry [5-7]. We achieved highly efficient profiling (24 minutes/sample) with low median process variability (<12%) of more than 250 metabolites, covering over 60 biochemical pathways of human metabolism (**Supplemental Table 1**). Based on our previous observation that ratios between metabolite concentrations can strengthen the association signal and provide new information about possible metabolic pathways [4,8], we included all pairs of ratios between these metabolites in the genome-wide statistical analysis.

To reduce the computational and data storage burden associated with meta-analyzing over 37,000 metabolites and ratios, we applied a staged approach for selection of promising association signals (**Supplemental Figure 1**). In the initial screening stage we assessed associations of approximately 600,000 genotyped SNPs with over 37,000 metabolic traits (concentrations and their ratios) by fitting linear models separately in both cohorts to log-transformed metabolic traits, adjusting for age, gender and family structure (**Supplemental Figure 2 & Supplemental Table 2**). Next, we selected all association signals having suggestive evidence for association with a metabolic trait in both cohorts ($p<10^{-6}$ in both cohorts or $p<10^{-3}$ in one and $p<10^{-9}$ in the other). For each of these loci, we then re-assessed the amount of association signals through fixed-effects inverse variance meta-analysis of the two cohorts for all 37,000 available traits using imputed SNPs relative to HapMap2 data (see **Online Methods** for details). The SNP/trait combination yielding the smallest P-value in this meta-analysis was finally selected for each locus. To account for multiple testing we applied conservative Bonferroni correction leading to an adjusted threshold for genome-wide significance of $p < 2.0\times10^{-12}$.

RESULTS

We identified a total of 37 independent loci that reached genome-wide significance in the meta-analysis (Table 1, **Supplemental Tables 3&4**). 23 of these loci describe new genetic associations with metabolic traits, and 14 replicate and extend our knowledge of known GDMs, including 10 from our own studies [3,4]. We used information on SNP location within genes, known gene function and regional association plots **(Supplemental Figure 2**) to prioritize plausible candidate genes within associated loci. In most cases our annotation was further supported by a statistical analysis of association of gene relationships in published literature [9] (**Supplemental Table 5**). Associations with additional metabolic traits at the 37 loci presented in Table 1 may capture further biochemical information and are provided as **Supplemental Table 6.** At 30 loci the sentinel SNP mapped to a protein that was biochemically linked to the associating metabolites, for instance because responsible for their synthesis, degradation or metabolism. We next extensively searched literature and databases (**see web-links**) to identify which of these 37 loci were previously reported as associated with a clinical endpoint, a medically relevant intermediate phenotype, or a pharmacogenetic effect. Associations of metabolites at disease loci can be used to gain novel information on possible metabolic changes associated with biological processes underlying that association (Figure 1, Table 1, **Supplemental Table 7**). In 15 cases such a relationship could be identified based on the association of the lead SNP or a proxy ($r^2\geq0.8$) with the disease-associated SNPs, including cardiovascular disease, kidney disease, Crohn's disease, gout, cancer, pharmacogenomics, and predisposing risk factors for diabetes and cardiovascular disease. Except for three loci, all SNPs are common with minor allele frequencies over 10%. In 25 cases the effect size per allele copy is larger than 10%, and up to 60% in the case of the *ACADS* locus.

Overlap with chronic disease loci

Many genetic risk loci for heart disease, kidney failure, diabetes and other complex disorders have been identified by GWAS. However, the etiology of these common diseases is complex and testable hypotheses are needed in order to develop new avenues for diagnosis and therapy. Associations of known disease risk loci with metabolic traits allow identifying new and potentially relevant biological processes and pathways. Below we report some examples from our study that illustrate this idea, with the full association dataset being freely available for further analysis and reference at http://www.gwas.eu.

Detoxification and kidney failure

N-acetylation is an important mechanism to detoxify numerous nephrotoxic medications and environmental toxins. A reduced ability to detoxify such substances could lead to impaired kidney function. A key GDM is the N-acetylase *NAT8* locus, which was reported to associate with kidney function [10,11]. Here we found a highly-significant association of variation at the *NAT8* locus with N-acetylornithine. Using this information we then asked whether N-acetylornithine concentrations were associated with kidney function. We found a clear association in both our studies with estimated glomerular filtration rate (eGFR), whereby higher levels of N-acetylornithine were correlated with lower eGFR ($p_{KORA}=7.6\times10^{-4}$, $p_{TwinsUK}=3.6\times10^{-8}$ after adjusting for age and gender). In accord with the genetic effect of the *NAT8* polymorphism in the chronic kidney disease (CKD) association, the risk allele associated here with higher N-acetylornithine concentrations. Although causality cannot be inferred from this kind of association studies, the role of ornithine acetylation in the etiology of CKD warrants further exploration.

Diabetes

GCKR is a major pleiotropic risk locus associated with diabetes- and cardiometabolic-related traits, such as fasting glucose and insulin [12], triglyceride levels [13], and CKD [11]. Here we identified a highly significant association of this locus with mannose to glucose ratios. Fasting mannose is lower in carriers of the risk allele, as opposed to glucose. Interestingly, we also observed a 3.3% increase in lactate concentrations per copy of the risk allele at the same locus. Little is known about the physiological role of mannose other than its use in protein glycosylation. Mannose enters the cell via a specific transporter that is insensitive to glucose [14], and hepatic glycogen breakdown is implicated in the maintenance of plasma mannose concentrations [15]. These observations and the association with *GCKR* observed here, which is even stronger than that of glucose with *GCKR*, suggest the need for further investigations on the role of mannose as a differential biomarker or even as a point of intervention in diabetes care.

Venous thromboembolism

With the mass-spectrometry method used here, different forms of the abundant fibrinogen A-alpha peptides can be detected. Fibrinogen plays a role in blood clot formation. Its active form, the fibrinogen A-alpha chain ADSGEGDFXAEGGGVR can be phosphorylated at Serine-3 to ADpSGEGDFXAEGGGVR [16]. The ratio between the concentrations of these fibrinogen A-alpha peptides provides a measure for fibrinogen A-alpha phosphorylation (FAaP). Increased levels of FAaP have been observed under different physiological and pathophysiological conditions [17]. Here, three loci (*ABO, ALPL, FUT2*) associated with FAaP. Intriguingly, these three genes are functionally linked: *ABO* and *FUT2* are involved in determining the blood group, and the *ABO* locus is associated with blood levels of the phosphatase ALPL [18]. The association of *ALPL* with FAaP may be explained by either a genotype-dependent dephosphorylation of fibrinogen by ALPL, or a genotype-dependent change in the phosphorus pool available for FAaP. Variants in the *ABO* gene are associated with many different outcomes, including venous thromboembolism (VTE) [19]. The association of *ABO* with FAaP, and thus modified blood coagulation properties, provides a functional explanation for the reported association of *ABO* with VTE risk. Moreover, if FAaP is at the basis of VTE, then *FUT2* and *ALPL* should also be investigated as VTE risk genes, which is a hypothesis that may now be tested in the respective patient groups.

Coronary artery disease

We have shown previously [4] that strong associations with metabolic traits can point to interesting associations in GWAS with clinical endpoints that otherwise would not be

considered as relevant. A recent meta-analysis with lipid traits [20] identified several genetic loci also affecting risk of CAD in the CARDIoGRAM study [21] using a similar strategy. Six of such loci are also reported here (*ABO, NAT2, CPS1, NAT8, ALPL, KLKB1*), albeit some of them showed only weak evidence for association ($p < 0.01$) with CAD in the CARDIoGRAM study (**Supplemental Table 8**). Although not statistically strong, the biochemical function of the associated metabolic traits identified here may support a possible role in heart disease. For instance, *NAT8* may be linked to CKD via ornithine acetylation (see above). *KLKB1* controls blood pressure via the bradykinin pathway. In this study a genetic variant in *KLKB1* associated with bradykinin concentrations and we also confirmed the expected directional association of bradykinin with hypertension in both our studies ($p_{KORA}=1.7\times10^{-9}$, $p_{TwinsUK}=0.0495$, with covariates age and gender). *ABO* and *ALPL* associated with FAaP, and it may therefore be speculated that genetically determined differences in FAaP and resulting blood coagulation properties may be at the basis of these associations with CAD. Furthermore, our associations suggest that the role of FAaP as a biomarker for acute myocardial infarction, and the combined additive genetic effect of *ABO, ALPL,* and *FUT2* loci (**Supplemental Figure 4**) on CAD risk, should be investigated in greater detail.

New biological and functional insights derived from this study

Genome-wide association studies uncover merely statistically significant associations and thereby are only able to generate biological hypotheses. While it is clear that providing experimental validation of all associations is beyond what can be achieved in a single study, we nevertheless attempted to show that in principle this is possible. The association of SNP rs7094971 in *SLC16A9* (*MCT9*) with carnitine suggested that this metabolite is the substrate of this hitherto uncharacterized monocarboxylic acid transporter. We therefore tested [^{3}H]-carnitine uptake by *SLC16A9*-expressing *Xenopus* oocytes. As shown in Figure 2, our data shows that SLC16A9 is a sodium- and pH-independent carnitine efflux transporter, possibly responsible for carnitine efflux from absorptive epithelia into the blood. Another prominent example is the highly significant association of increased urate levels and their clinical complication of gout with variants in the *SLC2A9* gene [22], the former of which we also observe here. Although previously annotated as a glucose transporter, *SLC2A9* was later shown [23] to encode a high-capacity urate transporter. Similar characterization experiments by specialists in the related fields shall be motivated and guided by our association data. Among the 37 GDMs reported here, we suggest that the associations with coarsely-characterized enzyme and transporter genes that are known disease risk loci may warrant further experimental investigation, for instance in experiments using isotope-labeled derivates of the associated metabolites reported here as putative target substrates. For the reasons detailed above we deem *NAT8* to be a prime candidate for such a study.

Pharmacogenomics

Using the *Pharmacogenomics Knowledge Base* [24] we identified six GDMs as previously associated with toxicity or adverse reactions to medication. Noteworthy are polymorphisms in the *NAT2* and in *CYP4A* loci that associated with toxicities to docetaxel and thalidomide treatment [25], the *UGT1A* locus with irinotecan toxicity [26], *SLC2A9* with etoposide IC_{50} [27], *SLC22A1* with metformin pharmacokinetics [28,29], and *SLCO1B1* with statin-induced myopathy [30]. In all cases our associations with metabolic traits at these loci provide a possible novel biochemical basis for the genotype-dependant reaction to drug treatment, such as the association of *SLCO1B1* with a series of fatty acids, including tetradecanedioate and hexadecanedioate. This information can be used to support redesign of the respective drug molecules to avoid adverse reactions. Moreover, systematic inclusion of biochemically relevant GDMs as candidate SNPs during drug trials may permit early identification of

potentially adverse pharmacogenetic effects. Concretely this applies to *AKR1C*, which is a novel target of jasmonates in cancer cells [31]. We reported a GDM associated with *AKR1C* with a large effect size on androgen metabolism. Influence of SNP rs2518049 in *AKR1C* on the drug's efficiency and potential side effects should therefore be assessed in upcoming clinical trials.

DISCUSSION

Due to their large effect size and high explained variance, the 37 genetically determined metabotypes (GDMs) reported in this study indicate key genetic loci underpinning differences in human metabolism. Inclusion of these genetic variants in the statistical analysis of pre-clinical and clinical studies may facilitate identification of genotype-dependent outcomes, such as disease complications and adverse drug reactions. In two cases we could establish a direct functional link, supported by both our studies, between a genetic variant, an intermediate metabolic trait, and a disease relevant endpoint: *KLKB1*-bradykinin-hypertension and *NAT8*-N-acetylornithine-eGFR. We note that by discussing only associations that are supported by two independent studies at genome-wide significance we have chosen to take a very conservative approach. Based on QQ-plots and coarse assumptions, we estimate that over 500 loci with signals of association below that conservative threshold may be confirmed as GDMs in future, more highly powered studies. On a more technical note it is worthwhile mentioning that by using a single study to metabolically profile 2,820 individuals, based on only 100 micro-liters of blood serum, we replicated in this study a wide series of findings from previous large GWAS with quantitative traits, including serum fasting glucose [12], bilirubin [32,33], urate [34], and dehydroisoandrosterone sulfate [35] levels. Taken together, our study shows how GWAS with intermediate traits that are close to the underlying biological processes provide significant new functional insights into associations from GWAS with complex chronic disease endpoints and drug toxicity. Future GWAS that combine multiple Omics-technologies in a single study, including transcriptomics, proteomics, metabolomics and recent technologies for determining epigenetic modifications on a genome-wide scale are likely the next big step towards a full understanding of the interaction of genetic predispositions with environmental factors in complex chronic diseases and safe and efficient therapies, diagnosis and prevention.

ONLINE METHODS

1. STUDY DESIGN

Study populations—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 study design and standardized examinations of the survey (4,261 participants, response 67%) have been described in detail (ref. [39] and the references therein). A total of 3,080 subjects participated in a follow-up examination KORA F4 in 2006–2008 comprising individuals who, at that time, were aged 32–81 years. The TwinsUK cohort is an adult twin British registry in the age range 8-102 years and 84% are female. The samples used in this study are aged 23-85 (mean 48 years) and 97% female. 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 [40]. In both studies written informed consent has been given by all participants and the studies have been approved by the local ethics committees (Bayerische Landesärztekammer for KORA and Guy's and St. Thomas' Hospital Ethics Committee for TwinsUK).

Blood sampling—Blood samples for metabolic analysis and DNA extraction from KORA were collected between 2006 and 2008 as part of the KORA F4 follow-up. To avoid variation due to circadian rhythm, blood was drawn in the morning between 8:00 a.m. and 10:30 a.m. after a period of at least 10 hours 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. 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.

2. GENETIC AND METABOLOMICS DATA COLLECTION

Metabolomics measurements—Metabolon, an US- based commercial supplier of metabolic analyses, developed a platform that integrates the chemical analysis, including identification and relative quantification, data reduction, and quality assurance components of the process. The analytical platform incorporates two separate ultrahigh performance liquid chromatography/tandem mass spectrometry (UHPLC/MS/MS2) injections and one GC/MS injection per sample. The UHPLC injections are optimized for basic and acidic species. A total of 295 metabolites were measured, spanning several relevant classes (amino acids, acylcarnitines, sphingomyelins, glycerophospholipids, carbohydrates, vitamins, lipids, nucleotides, peptide, xenobiotics and steroids; a full list of metabolites is given in **Supplemental Table 1**). The detection of the entire panel was carried out with 24 min instrument analysis time (two injections at 12 min each), while maintaining low median process variability (<12% across all compounds). The resulting MS/MS^2 data were searched against a standard library generated by Metabolon that included retention time, molecular weight (m/z), preferred adducts, and in-source fragments as well as their associated MS/MS spectra for all molecules in the library. The library allowed for the identification of the experimentally detected molecules based on a multiparameter match without need for additional analyses. Metabolon has shown in a recent publication that their integrated platform enabled the high-throughput collection and relative quantitative analysis of analytical data and identified a large number and broad spectrum of molecules with a high degree of confidence [5]. The Metabolon platform has, among other studies, been successfully applied in the analysis of the adult human plasma metabolome [41] and the identification of sarcosine as a biomarker for prostate cancer [42].

Metabolomics data QC—For this study we measured the Metabolon panel in human blood from 1,768 individuals of the KORA cohort and in 1,052 individuals of the TwinsUK cohort. Quality control data (%RSD, upper and lower 95% confidence interval, minimum and maximum observed values in QC samples) are reported in **Supplemental Table 1**. In order to avoid spurious false positive associations due to small sample sizes, only metabolic traits with at least 300 non-missing values were included and data-points of metabolic traits that lay more than 3 standard deviations off the mean were excluded by setting them to missing in the analysis. 276 of 295 available metabolites and 37,179 metabolite ratios satisfied this criterion in KORA, resulting in a total of 37,455 metabolic traits. For the TwinsUK study, identical selection criteria for metabolic traits were used, resulting in 258 metabolites and 32,499 metabolite ratios, and a total of 32,757 metabolic traits.

Genotyping and imputation

KORA: For all individuals profiled in this study, genome-wide SNP data were already available. GWAS data of KORA and TwinsUK have been used and described extensively in the past in the context of numerous genome-wide association studies and meta-analyses [3,34,43]. We therefore only summarize the essential details here. Genotyping of the KORA F4 population was carried out using the Affymetrix GeneChip array 6.0. Genotypes were determined using Birdseed2 clustering algorithm. For quality assurance we applied as filters for SNP quality: call rate > 95% and p(HWE) > 10^{-6}. 655,658 autosomal SNPs satisfied these criteria. These genotyped SNPs were used for genome-wide analysis of the metabolic traits. For selection of the best associated SNP in a meta-analysis of KORA and TwinsUK within a region we used genotyped as well dosages of imputed SNPs. In KORA F4 imputation was done using IMPUTE v0.4.2 [44] based on HapMap2 (see below).

TwinsUK: Genotyping of the TwinsUK dataset was done with a combination of Illumina arrays (HumanHap300, HumanHap610Q, 1M-Duo and 1.2MDuo 1M) [45,46]. We pooled the normalised intensity data for each of the three arrays separately (with 1M-Duo and 1.2MDuo 1M pooled together). For each dataset we used the Illluminus calling algorithm [47] to assign genotypes in the pooled data. No calls were assigned if an individual's most likely genotyped was called with less than a posterior probability threshold of 0.95. Validation of pooling was achieved via a visual inspection of 100 random, shared SNPs for overt batch effects. Finally, intensity cluster plots of significant SNPs were visually inspected for over-dispersion biased no calling, and/or erroneous genotype assignment. SNPs exhibiting any of these characteristics were discarded. We applied similar exclusion criteria to each of the three dataset separately. *Samples*: Exclusion criteria were: (i) sample call rate <98%, (ii) heterozygosity across all SNPs ≥2 s.d. from the sample mean; (iii) evidence of non-European ancestry as assessed by PCA comparison with HapMap3 populations; (iv) observed pairwise IBD probabilities suggestive of sample identity errors; (v). We corrected misclassified monozygotic and dizygotic twins based on IBD probabilities. *SNPs*. Exclusion criteria were (i) Hardy-Weinberg p-value<10^{-6}, assessed in a set of unrelated samples; (ii) MAF<1%, assessed in a set of unrelated samples; (iii) SNP call rate <97% (SNPs with MAF≥5%) or < 99% (for 1% ≤MAF < 5%). Alleles of all three datasets were aligned to HapMap2 or HapMap3 fwd strand alleles. Prior to merging, we performed pairwise comparison among the three datasets and further excluded SNPs and samples to avoid spurious genotyping effects, indentified as follows: (i) concordance at duplicate samples <1%; (ii) concordance at duplicate SNPs <1%; (iii) visual inspection of QQ plots for logistic regression applied to all pairwise dataset comparisons; (iv) Hardy-Weinberg p-value<10^{-6}, assessed in a set of unrelated samples; (v) observed pairwise IBD probabilities suggestive of sample identity errors. We then merged the three datasets, keeping individuals typed at the largest number of SNPs when an individual was typed at two different arrays. The merged dataset consists of 5,654 individuals (2,040 from the HumanHap300, 3,461 from the HumanHap610Q and 153 from the HumanHap1M and 1.M arrays) and up to 874,733 SNPs depending on the dataset (HumanHap300: 303,940, HumanHap610Q: 553,487, HumanHap1M and 1.M: 874,733). Imputation was performed using the IMPUTE software package (v2) [44] using two reference panels, P0 (HapMap2, rel 22, combined CEU+YRI +ASN panels) and P1 (610k+, including the combined HumanHap610k and 1M reduced to 610k SNP content). 534,665 autosomal SNPs were used for the analysis of this study (basically 610K SNPs extracted from the final merged data set).

3. DATA ANALYSIS

Statistical analyses—The primary association testing was carried out using linear regressions on all metabolite concentrations and all possible ratios of metabolite concentrations. This was motivated by our previous observation [4,8] that the use of ratios may

lead to a strong reduction in the overall trait variance. A test of normality showed that in 29,338 cases the log-transformed ratio distribution was significantly better represented by a normal distribution than when untransformed ratios were used. In 5,145 cases untransformed distribution was closer to a normal distribution. For concentrations 149 were closer to a lognormal distribution while 124 were better represented by a normal distribution. Based on this observation, and also for sake of simplicity, we decided to log-transform all metabolites and their ratios. We used the p-gain statistics [4,8] to quantify the decrease in p-value for the association with the ratio compared to the p-values of the two corresponding concentrations. A high p-gain (above 250) indicates that two metabolites are more likely to be functionally linked in a metabolic pathway that is impacted the associating genotype. KORA and TwinsUK are population-based studies. They comprise only individuals who are not displaying any severe clinical symptoms at the time of sampling. Therefore, disease state has not considered as a confounding factor in the statistical analysis. In KORA, the software PLINK (version 1.06)[48] and SNPTEST was used with age and gender as covariates. In order to account for the family structure in the TwinsUK study, we used variance components applied to a score test implemented in the software Merlin [49].

Correction for multiple testing—We applied a conservative Bonferroni correction to control for false positive error rates deriving from multiple testing. Using the KORA study as reference, we corrected for tests on 655,658 SNPs and 37,455 metabolic traits, thus obtaining a Bonferroni-adjusted p-value of $p = 2.04\times10^{-12}$. For ratios we required in addition that the increase in the strength of association, expressed as the change in p-value when using ratios compared to the larger of the two p-values when using two metabolite concentrations individually (p-gain), be larger than the number of tested metabolic traits (p-gain>250) [4,8]. This limit is considered as a Bonferroni-type conservative cut-off for identifying those metabolite concentration pairs for which the use of ratios strongly improves the strength of association. Others than the strongest associating metabolic trait often provide additional insight into the underlying biochemical processes. In such cases we consider a p-value of $p = 1.33\times10^{-6}$ to represent a conservative level of significance (Bonferroni correction for 37,455 tests at a nominal significance level of 5%).

Inflation—In most cases the assumption of a linear additive model was valid (see box plots in **Supplemental Figure 3**) and there was no inflation of summary statistics which could be indicative of population stratification (see QQ-plots in **Supplemental Figure 3**). Lambda values ranged from 0.965 to 1.024 (median=1.006) in KORA and from 0.940 to 1.013 (median=0.985) in TwinsUK.

Candidate gene selection and overlap with disease loci—Regional association plots (**Supplemental Figure 3)** were created using imputed and meta-analyzed data. Within this region the SNP with the strongest signal of association in the meta-analysis was retained as the final SNP to be reported. Association data for all metabolic traits at the 37 SNPs reported in Table 1 (for KORA, TwinsUK and meta-analysis), limited to associations with $p<1.33\times10^{-6}$ (Bonferroni correction for multiple testing of metabolic traits at a single locus) and p-gain>250 (for ratios) in the meta-analysis are reported in **Supplemental Table 4.** For the strongest associating trait box plots were plotted to visualize the actual quantitative dependence of the trait on genotype (**Supplemental Figure 3**). Based on association data alone, it is in most cases not possible to identify the implicated gene within a locus that causes the association. However, using knowledge on the function of genes within linkage disequilibrium of the reported SNP as well as the biochemical characteristic of the associating metabolite, it is in many cases possible to identify a single most likely candidate gene. These cases are tagged as ‘match between gene function and metabolic trait’ and are supported by arguments provided as supplemental text (e.g. association of a SNP in LD with

OPLAH (oxoprolinase) and oxoproline concentrations). At two loci (*CYP4A* and *UGT1A*) variants with alternative splice variants exist. We named these loci without attempting to specify the exact variant.

GWAS catalog—Using the catalogue of published genome-wide association studies (accessed 10 October 2010) [1] we identified for each entry the SNPs in the KORA and TwinsUK studies that correlate most strongly ($r^2 \geq 0.5$) and that was present in our association database ($p<10^{-3}$, p-gain>10). The resulting associations are available online on our GWAS server. New associations shall be included as the database of published GWAS is updated.

Enrichment analysis—We downloaded the actual version of the GWAS catalogue from NHGRI and deleted all records that correspond to our previous studies. As a sampling dataset, we chose the 655,658 SNP from the Affymetrix 6.0 array, which have been tested in the KORA part of this study. The 37 SNPs that we report are from this array and can thus be considered as representing one draw out of this set. We then drew 1,000,000 sets of 37 SNPs at random (with replacement) from this sampling dataset. To account for comparable MAF distributions between the reference and the random set we then rejected all draws where the mean or the variance of the MAF distributions were significantly different ($p<0.05$) between the random and the reference set. 330,775 random sets were hence retained. Using an LD criterion of $r^2>0.8$ (based on HapMap2 release #27, NCBI B36, CEU population), we then counted for every random set the overlap with the GWAS catalogue. The reference set was included as a technical positive control in the computations. For the 330,775 tested random sets, at most six overlapping SNPs were found (8 times), and in over half of the cases no overlapping SNPs were present in the sampled dataset (see Table below).

number of SNPs overlapping with the NHGRI GWAS catalogue	number of random occurrences
0	182,924
1	109,288
2	31,744
3	5,931
4	778
5	102
6	8
total	330,775

For our reported 37 metabolomics SNPs, we identified 14 overlapping SNPs (note that we report 15 overlapping loci Figure 1; the *ENPEP* locus was not yet included in the GWAS catalogue and was not used in this analysis). As we never found 14 overlapping loci by chance, the p-value of our observations being due to chance is below $p = 1/330,775 = 3\times10^{-6}$.

Functional characterization of SLC16A9—The *SLC16A9* (*MCT9*) clone (IMAGE ID 40146598) was purchased from Autogen Bioclear (Wiltshire, UK). Plasmid was linearised with SpeI restriction enzyme (New englan Biolabs, UK) and cRNA synthesised in vitro using the T7 mMachine in vitro transcription system (Ambion, Applied Biosystems, Warrington, UK). *MCT9* was expressed in *Xenopus laevis* oocytes as described previously (Meredith 2004). Briefly, stage V-VI oocytes were injected with 10ng of *MCT9* cRNA and incubated in modified Barth's solution for 3-4 days at 18°C with the medium changed daily. Control oocytes had either no injection (NI) or an injection of an equal volume (50nl) of

Europe PMC Funders Author Manuscripts

distilled H_2O (WI) and were incubated for the same length of time. Uptake and efflux experiments were performed similarly to those described previously [50] except the substrate was [^{3}H]-carnitine (specific activity 81Ci/mmol, GE Healthcare, UK).

Data access—This study generated millions of individual data points through the profiling of *n* metabolites and *n**(*n*-1)/2 ratios in ~3,000 individuals, and the subsequent associations with millions of genetic variants from GWAS. We created a web-based interface and visualization tools for the dissemination of results to the scientific community, with the aims of allowing rapid storage and retrieval of data as well as managing the integration of metabolomics summary statistics vis-a-vis published GWAS studies. The association data is freely available at through the server http://metabolomics.helmholtz-muenchen.de/gwa/ and mirror sites located at the Wellcome Trust Sanger Institute and King's College London sites.

Acknowledgments

We gratefully acknowledge the contributions of P. Lichtner, G. Eckstein, Guido Fischer, T. Strom and all other members of the Helmholtz Zentrum München genotyping staff in generating the SNP dataset, as well as the contribution of all members of field staffs who were involved in the planning and conduct of the MONICA/KORA Augsburg studies. The KORA group consists of H.E. Wichmann (speaker), A. Peters, C. Meisinger, T. Illig, R. Holle, J. John and their co-workers who are responsible for the design and conduct of the KORA studies. For TwinsUK we thank the staff from the Genotyping Facilities at the Wellcome Trust Sanger Institute for sample preparation, quality control and genotyping. Guido Fischer (KORA) and Gabriela Surdulescu (TwinsUK) selected the samples, sample handling and shipment was organized by Humberto Chavez (KORA) and Dylan Hodgkiss (TwinsUK), and Ulrike Goebel (Helmholtz) provided administrative support. Special thanks go to Daniel Garcia-West for his role in facilitating this study. We are grateful to the CARDIoGRAM investigators for access to their dataset. Finally, we wish to express our appreciation to all study participants of the KORA and the TwinsUK studies for donating their blood and time.

FUNDING The KORA research platform (KORA: *Kooperative Gesundheitsforschung in der Region Augsburg*) and the MONICA Augsburg studies (Monitoring trends and determinants on cardiovascular diseases) 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. This study was supported by a grant from the German Federal Ministry of Education and Research (BMBF) to the German Center for Diabetes Research (DZD e.V.). Part of this work was financed by the German National Genome Research Network (NGFNPlus: 01GS0823). 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) and the FP-5 GenomEUtwin Project (QLG2-CT-2002-01254). The study also receives support from the Dept of Health via the 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. TDS is an NIHR senior Investigator. The project also received support from a *Biotechnology and Biological Sciences Research Council* (BBSRC) project grant (G20234). Both studies received support from ENGAGE project grant agreement HEALTH-F4-2007-201413. NJS holds a British Heart Foundation Chair and is an NIHR Senior Investigator and is supported by the Leicester NIHR Biomedical Research Unit in Cardiovascular Disease. The authors acknowledge the funding and support of the National Eye Institute via an NIH/CIDR genotyping project (PI: Terri Young). Genotyping was also performed by CIDR as part of an NEI/NIH project grant. DM received support from the Early Career Researcher Scheme, Oxford Brookes University, UK. JR is supported by DFG Graduiertenkolleg "GRK 1563, Regulation and Evolution of Cellular Systems" (RECESS), EA by BMBF Grant no. 0315494A (project SysMBo), WRM by BMBF grant no. 03IS2061B (project Gani_Med), and BW by Era-Net grant no. 0315442A (project PathoGenoMics). AK is supported by the Emmy Noether Programme of the DFG (KO-3598/2-1) and FK by grants from the "Genomics of Lipid-associated Disorders – GOLD" of the "Austrian Genome Research Programme GEN-AU". NS is supported by the Wellcome Trust (Core Grant Number 091746/Z/10/Z).

REFERENCES

1. Hindorff LA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A. 2009; 106:9362–9367. [PubMed: 19474294]
2. Newgard CB, Attie AD. Getting biological about the genetics of diabetes. Nat Med. 2010; 16:388–391. [PubMed: 20376050]
3. Illig T, et al. A genome-wide perspective of genetic variation in human metabolism. Nat Genet. 2010; 42:137–141. [PubMed: 20037589]
4. Gieger C, et al. Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. PLoS Genet. 2008; 4:e1000282. [PubMed: 19043545]
5. Evans AM, DeHaven CD, Barrett T, Mitchell M, Milgram E. Integrated, Nontargeted Ultrahigh Performance Liquid Chromatography/Electrospray Ionization Tandem Mass Spectrometry Platform for the Identification and Relative Quantification of the Small-Molecule Complement of Biological Systems. Analytical Chemistry. 2009; 81:6656–6667. [PubMed: 19624122]
6. Ohta T, et al. Untargeted metabolomic profiling as an evaluative tool of fenofibrate-induced toxicology in Fischer 344 male rats. Toxicol Pathol. 2009; 37:521–535. [PubMed: 19458390]
7. Suhre K, et al. Metabolic footprint of diabetes: a multiplatform metabolomics study in an epidemiological setting. PLoS ONE. 2010; 5:e13953. [PubMed: 21085649]
8. Altmaier E, et al. Bioinformatics analysis of targeted metabolomics--uncovering old and new tales of diabetic mice under medication. Endocrinology. 2008; 149:3478–3489. [PubMed: 18372322]
9. Raychaudhuri S, et al. Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. PLoS Genet. 2009; 5:e1000534. [PubMed: 19557189]
10. Chambers JC, et al. Genetic loci influencing kidney function and chronic kidney disease. Nat Genet. 2010; 42:373–375. [PubMed: 20383145]
11. Kottgen A, et al. New loci associated with kidney function and chronic kidney disease. Nat Genet. 2010; 42:376–384. [PubMed: 20383146]
12. Dupuis J, et al. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. Nat Genet. 2010; 42:105–116. [PubMed: 20081858]
13. Aulchenko YS, et al. Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. Nat Genet. 2009; 41:47–55. [PubMed: 19060911]
14. Panneerselvam K, Freeze HH. Mannose enters mammalian cells using a specific transporter that is insensitive to glucose. J Biol Chem. 1996; 271:9417–9421. [PubMed: 8621609]
15. Taguchi T, et al. Hepatic glycogen breakdown is implicated in the maintenance of plasma mannose concentration. Am J Physiol Endocrinol Metab. 2005; 288:E534–540. [PubMed: 15536204]
16. Blombaeck B, Blombaeck M, Edman P, Hessel B. Amino-acid sequence and the occurrence of phosphorus in human fibrinopeptides. Nature. 1962; 193:833–834. [PubMed: 13870090]
17. Martin SC, Ekman P, Forsberg PO, Ersmark H. Increased phosphate content of fibrinogen in vivo correlates with alteration in fibrinogen behaviour. Thromb Res. 1992; 68:467–473. [PubMed: 1341057]
18. Yuan X, et al. Population-based genome-wide association studies reveal six loci influencing plasma levels of liver enzymes. Am J Hum Genet. 2008; 83:520–528. [PubMed: 18940312]
19. Tregouet DA, et al. Common susceptibility alleles are unlikely to contribute as strongly as the FV and ABO loci to VTE risk: results from a GWAS approach. Blood. 2009; 113:5298–5303. [PubMed: 19278955]
20. Teslovich TM, et al. Biological, clinical and population relevance of 95 loci for blood lipids. Nature. 2010; 466:707–713. [PubMed: 20686565]
21. Schunkert H, et al. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. Nat Genet. 2011; 43:333–338. [PubMed: 21378990]
22. Doring A, et al. SLC2A9 influences uric acid concentrations with pronounced sex-specific effects. Nat Genet. 2008; 40:430–436. [PubMed: 18327256]

23. Caulfield MJ, et al. SLC2A9 is a high-capacity urate transporter in humans. PLoS Med. 2008; 5:e197. [PubMed: 18842065]
24. Klein TE, et al. Integrating genotype and phenotype information: an overview of the PharmGKB project. Pharmacogenetics Research Network and Knowledge Base. Pharmacogenomics J. 2001; 1:167–170. [PubMed: 11908751]
25. Deeken JF, et al. A pharmacogenetic study of docetaxel and thalidomide in patients with castration-resistant prostate cancer using the DMET genotyping platform. Pharmacogenomics J. 2010; 10:191–199. [PubMed: 20038957]
26. Lankisch TO, et al. Gilbert's Syndrome and irinotecan toxicity: combination with UDP-glucuronosyltransferase 1A7 variants increases risk. Cancer Epidemiol Biomarkers Prev. 2008; 17:695–701. [PubMed: 18349289]
27. Huang RS, et al. A genome-wide approach to identify genetic variants that contribute to etoposide-induced cytotoxicity. Proc Natl Acad Sci U S A. 2007; 104:9758–9763. [PubMed: 17537913]
28. Chen Y, et al. Effect of genetic variation in the organic cation transporter 2 on the renal elimination of metformin. Pharmacogenet Genomics. 2009; 19:497–504. [PubMed: 19483665]
29. Shu Y, et al. Effect of genetic variation in the organic cation transporter 1, OCT1, on metformin pharmacokinetics. Clin Pharmacol Ther. 2008; 83:273–280. [PubMed: 17609683]
30. Link E, et al. SLCO1B1 variants and statin-induced myopathy--a genomewide study. N Engl J Med. 2008; 359:789–799. [PubMed: 18650507]
31. Davies NJ, et al. AKR1C isoforms represent a novel cellular target for jasmonates alongside their mitochondrial-mediated effects. Cancer Res. 2009; 69:4769–4775. [PubMed: 19487289]
32. Sanna S, et al. Common variants in the SLCO1B3 locus are associated with bilirubin levels and unconjugated hyperbilirubinemia. Hum Mol Genet. 2009; 18:2711–2718. [PubMed: 19419973]
33. Johnson AD, et al. Genome-wide association meta-analysis for total serum bilirubin levels. Hum Mol Genet. 2009; 18:2700–2710. [PubMed: 19414484]
34. Kolz M, et al. Meta-analysis of 28,141 individuals identifies common variants within five new loci that influence uric acid concentrations. PLoS Genet. 2009; 5:e1000504. [PubMed: 19503597]
35. Zhai G, et al. Eight Common Genetic Variants Associated with Serum DHEAS Levels Suggest a Key Role in Ageing Mechanisms. PLoS Genet. 2011; 7:e1002025. [PubMed: 21533175]
36. Mootha VK, Hirschhorn JN. Inborn variation in metabolism. Nat Genet. 2010; 42:97–98. [PubMed: 20104246]
37. Meredith D, Christian HC. The SLC16 monocaboxylate transporter family. Xenobiotica. 2008; 38:1072–1106. [PubMed: 18668440]
38. Koepsell H, Endou H. The SLC22 drug transporter family. Pflugers Arch. 2004; 447:666–676. [PubMed: 12883891]
39. Wichmann HE, Gieger C, Illig T. KORA-gen--resource for population genetics, controls and a broad spectrum of disease phenotypes. Gesundheitswesen. 2005; 67(Suppl 1):S26–30. [PubMed: 16032514]
40. Andrew T, et al. Are twins and singletons comparable? A study of disease-related and lifestyle characteristics in adult women. Twin Res. 2001; 4:464–477. [PubMed: 11780939]
41. Lawton KA, et al. Analysis of the adult human plasma metabolome. Pharmacogenomics. 2008; 9:383–397. [PubMed: 18384253]
42. Sreekumar A, et al. Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. Nature. 2009; 457:910–914. [PubMed: 19212411]
43. Soranzo N, et al. A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. Nat Genet. 2009; 41:1182–1190. [PubMed: 19820697]
44. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 2009; 5:e1000529. [PubMed: 19543373]
45. Richards JB, et al. Bone mineral density, osteoporosis, and osteoporotic fractures: a genome-wide association study. Lancet. 2008; 371:1505–1512. [PubMed: 18455228]

46. Soranzo N, et al. Meta-analysis of genome-wide scans for human adult stature identifies novel Loci and associations with measures of skeletal frame size. PLoS Genet. 2009; 5:e1000445. [PubMed: 19343178]
47. Teo YY, et al. A genotype calling algorithm for the Illumina BeadArray platform. Bioinformatics. 2007; 23:2741–2746. [PubMed: 17846035]
48. Purcell S, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007; 81:559–575. [PubMed: 17701901]
49. Abecasis GR, Cherny SS, Cookson WO, Cardon LR. Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet. 2002; 30:97–101. [PubMed: 11731797]
50. Meredith D. Site-directed mutation of arginine 282 to glutamate uncouples the movement of peptides and protons by the rabbit proton-peptide cotransporter PepT1. J Biol Chem. 2004; 279:15795–15798. [PubMed: 14715671]

Europe PMC Funders Author Manuscripts

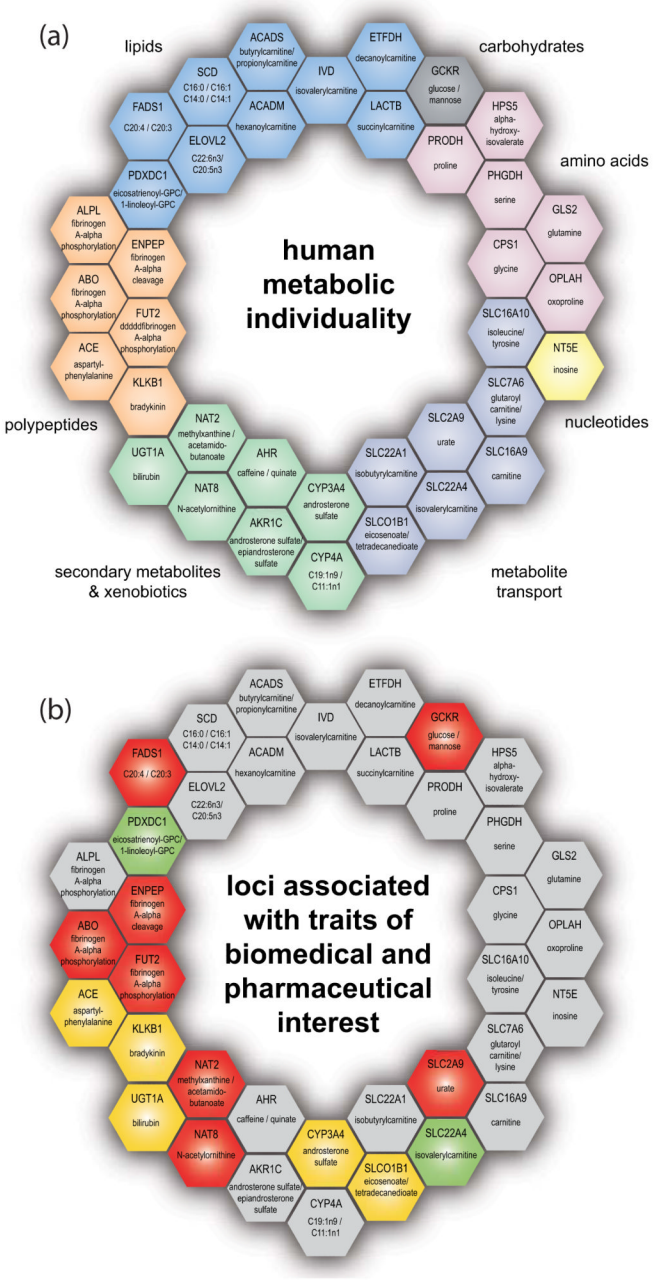

Figure 1. Genetic basis of human metabolic individuality and its overlap with loci of biomedical and pharmaceutical interest

Over 100 years ago Archibald Garrod realized that inborn errors in human metabolism were *'merely extreme examples of variations of chemical behaviour which are probably everywhere present in minor degrees'* and that this *'chemical individuality [confers] predisposition to and immunities from the various mishaps which are spoken of as diseases'* [36]. The 37 genetically determined metabotypes (GDMs) we reported here explain a highly relevant amount of the total variation in the studied population and therefore contribute substantially to the genetic part of human metabolic individuality. GDMs are presented here color-coded **(a)** by general metabolic pathways together with selected

associated metabolic traits, highlighting the relationship between gene function and the associated metabolic trait (see column 4 in Table 1), and **(b)** by overlap with associations in previous GWAS with disease [red], intermediate disease risk factors [yellow], and other traits [green]. Locus overlap is defined here by the lead SNP reported in the NHGRI GWAS catalogue being highly correlated (R^2>=0.8) with the most associated SNP in the metabolomics scan (see column 5 in Table 1 and **Suppl. Table 7**). Note that the overlap between the metabolomics loci and the loci reported by the NHGRI GWAS catalogue is highly significant when compared to a draw of 37 randomly selected SNPs with similar properties ($p<3\times10^{-6}$, see **Methods**).

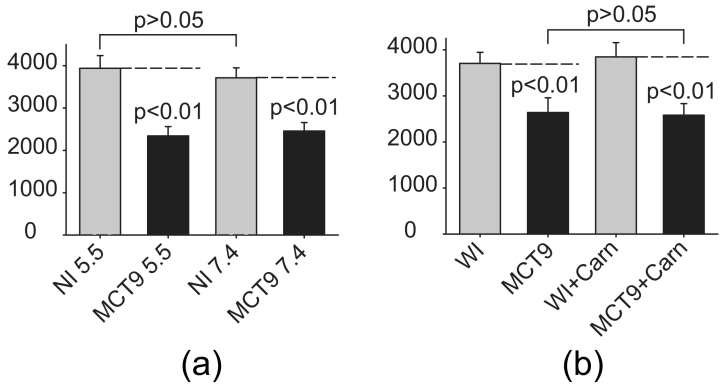

Figure 2. Experimental evidence for SLC16A9 (MCT9) as a carnitine efflux transporter

When incubated in uptake medium containing [^{3}H]-carnitine (4μCi/ml) there was no significant uptake indicating that MCT9 does not mediate carnitine uptake. As some of the previously characterised MCTs are proton-coupled [37], uptake was measured at both pH_{out} 7.4 and 5.5, but no significant difference was observed (data not shown). However, when 4.6nl of [^{3}H]-carnitine was injected into the oocyte followed by incubation in medium for 90 minutes, efflux was significantly higher in oocytes expressing MCT9 than in the non-injected (NI, **Figure a**) or water-injected (WI, **Figure b**) controls, while again changing the pH_{out} had no effect (**Figure a**). In agreement with the lack of uptake of [^{3}H]-carnitine, external unlabelled carnitine was unable to *trans*-stimulate [^{3}H]-carnitine efflux with no significant difference in efflux between MCT9-expressing oocytes in the absence or presence of 5mM carnitine (MCT9 vs. MCT9+carn, respectively, **Figure b**). Data are means ± SEM of 6-10 oocytes per data point from 2 oocyte preparations. Y-axis on plots represents remaining [^{3}H]-Carnitine (cpm/oocytes). Statistical significance was determined by the Student's t test. Taken together, these results are consistent with MCT9 acting as a unidirectional carnitine efflux system when expressed in *Xenopus* oocytes. Note that additional experiments are required to establish the full substrate specificity of MCT9. If future studies show an appropriate cellular distribution, MCT9 could be responsible for carnitine efflux across the basolateral membrane of absorptive epithelial cells following absorption via the well-characterized apical epithelial proton-coupled carnitine transporters OCTNs / SLC22 family [38].

Table 1
37 loci that displayed genome-wide significance in the meta-analysis

The metabolic trait with the strongest association at the discovery stage in both studies is reported together with the SNP identifier and the p-value of association from the meta-analysis. Full association data are available in **Supplemental Tables 3 & 5** and via a web-server (http://www.gwas.eu). The loci are labeled by the gene that is considered most likely to carry the causative SNP. Overlap with associations from other GWAS studies are highlighted in bold (R^2>0.8, details are in **Supplemental Table 6**). Where the metabolic trait is consistent with a nearby gene's function, details are provided in the column labeled 'Relationship between gene function and the associated metabolic traits'. Metabolic traits that are associated with the SNP at the corresponding locus are marked with a superscript '+'. Further information and full bibliographic references are presented in **Supplemental Table 4**.

Locus & SNP id	Metabolic trait	p-value	Relationship between gene function and the associated metabolic traits	Biomedical and pharmaceutical interest
ACADS rs2066938	butyrylcarnitine / propionylcarnitine	$<4.4\times10^{-305}$	Butyrylcarnitine$^+$ and propionylcarnitine$^+$ are substrates/products of ACADS	ACADS is a key enzyme in the mitochondrial fatty acid beta-oxidation
NAT8 rs13391552	N-acetylornithine	5.4×10^{-252}	N-acetyltransferase function of NAT8 matches the associating metabolite N-acetylornithine$^+$	Association with glomerular filtration and CKD; association of N-acetylornithine$^+$ with eGFR in this study
FADS1 rs174547	1-arachidonoylglycero-phosphoethanolamine/ 1-linoleoylglycerophospho-ethanolamine	8.5×10^{-116}	FADS1 substrate/product pair ratio arachidonate (20:4n6)$^+$ / dihomo-linolenate (20:3n3 or n6)$^+$ is among the top associations	Association with LDL cholesterol, HDL cholesterol & triglycerides, fasting glucose & HOMA-B, Crohn's disease, resting heart rate
UGT1A rs887829	bilirubin (E,E) / oleoylcarnitine	2.9×10^{-74}	Bilirubin$^+$ is a substrate of UGT1A1	Association with hyperbilirubinemia; low serum concentration of bilirubin associate with increased risk of coronary artery disease; a SNP in *UGT1A1* is a pharmacogenetic risk factor for irinotecan toxicity
ACADM rs211718	hexanoylcarnitine / oleate (18:1n9)	2.2×10^{-71}	Hexanoylcarnitine$^+$ is a substrate of ACADM	ACADM is a key enzyme in the mitochondrial fatty acid beta-oxidation
OPLAH rs6558295	5-oxoproline	1.5×10^{-59}	5-oxoproline$^+$ is a substrate of 5-oxoprolinase OPLAH	
SCD rs603424	myristate (14:0) / myristoleate (14:1n5)	2.9×10^{-57}	SCD catalyzes the delta-9-desaturation of fatty acids, such as myristate (14:0)$^+$ to myristoleate (14:1n5)$^+$ and palmitate (16:0)$^+$ to palmitoleate (16:1n7)$^+$	Palmitoleate (16:1n7) is a lipokine linking adipose tissue to systemic metabolism
GCKR rs780094	glucose / mannose	5.5×10^{-53}	GCKR plays a role in glucose homeostasis, strong association with mannose$^+$ to glucose$^+$ ratios matches the gene's function	Association with type 2 diabetes, fasting glucose, fasting insulin; serum uric acid; triglyceride levels; C-reactive protein; eGFRcrea; Crohn's disease; hypertriglyceridemia

Locus & SNP id	Metabolic trait	p-value	Relationship between gene function and the associated metabolic traits	Biomedical and pharmaceutical interest
NAT2 rs1495743	1-methylxanthine / 4-acetamidobutanoate	1.7×10^{-40}	4-acetamidobutanoate^{+}, 1-methylxanthine^{+}, and 1-methylurate^{+} are linked to NAT2 in the xenobiotics pathways	Association with triglyceride levels and CAD; bladder cancer; toxicities to docetaxel and thalidomide treatment
CYP3A4 rs17277546	androsterone sulfate	8.7×10^{-40}	CYP3A cytochrome P450 proteins metabolize androsterone sulfate^{+}	Genetic variance in androsterone metabolism is linked to the incidence of prostate cancer
ABO rs612169	ADpSGEGDFXAEGGGVR / ADSGEGDFXAEGGGVR	9.1×10^{-40}	Polymorphisms in ABO determine the blood group; association to fibrinogen peptide phosphorylation^{+}; additive effect on fibrinogen A-alpha phosphorylation together with *FUT2* and *ALPL*	Association with blood ALP level; pancreatic cancer; venous thromboembolism; phytosterol levels
SLC2A9 rs4481233	urate	5.5×10^{-34}	SLC2A9 (GLUT9) transports uric acid^{+}	Association with gout; several SNPs in *SLC2A9* associate with etoposide IC_{50}
CYP4A rs9332998	10-nonadecenoate (19:1n9) / 10-undecenoate (11:1n1)	5.1×10^{-32}	Cytochrome P450, family 4, subfamily A, are fatty acid omega-hydroxylases; 10-undecenoate (11:1n1)$^{+}$ is biochemically related to omega-hydroxylated C10 fatty acids	Possible role in the etiology of hepatic steatosis in interaction with *SCD*
CPS1 rs2216405	glycine	1.6×10^{-27}	Association with glycine^{+} and creatine^{+}; creatine is produced from glycine; glycine is metabolically related to carbamoyl phosphate, which is the product of CPS1 and the entry point of ammonia into the urea cycle	Metabolomics data suggests that this association is related to a perturbed ammonia metabolism
LACTB rs2652822	succinylcarnitine	7.2×10^{-27}	Association with succinylcarnitine^{+}; perturbed hepatic gene expression in transgenic *LACTB* mice suggests a role of LACTB in butanoate/succinate^{+} pathway	*LACTB*^{tg} mice are obese
SLC22A1 rs662138	isobutyrylcarnitine	7.3×10^{-25}	SLC22A1 (OCT1) translocates a broad array of organic cations, possibly also isobutyrylcarnitine^{+} or related metabolites	Genetic variation in SLC22A1/SLC22A2 region are determinants of metformin pharmacokinetics
SLCO1B1 rs4149081	eicosenoate (20:1n9 or 11) / tetradecanedioate	2.8×10^{-22}	SLCO1B1 (OATP2, OATP-C) is an organic anion transporter	Common variants in *SLCO1B1* are strongly associated with an increased risk of statin-induced myopathy
FUT2 rs503279	ADpSGEGDFXAEGGGVR / ADSGEGDFXAEGGGVR	4.3×10^{-20}	FUT2 is involved in the creation of a precursor of a H antigen, additive effect on fibrinogen A-alpha phosphorylation together with *ABO* and *ALPL*.	Association with vitamin B12 levels, total cholesterol, Crohn's disease; vitamin B12 deficiency is associated with cognitive decline, cancer and CAD
ACE rs4329	aspartylphenylalanine	8.2×10^{-20}	Angiotensin I converting enzyme (peptidyl-dipeptidase A) 1 is associated with the dipeptide aspartylphenylalanine^{+}	Association with angiotensin-converting enzyme activity, potential genetic interaction with *KLKB1* locus
PHGDH rs477992	serine	2.6×10^{-14}	PHGDH catalyses the first and rate-limiting step in the	

Locus & SNP id	Metabolic trait	p-value	Relationship between gene function and the associated metabolic traits	Biomedical and pharmaceutical interest
			phosphorylated pathway of serine[+] biosynthesis	
ENPEP rs2087160	ADpSGEGDFXAEGGGVR / DSGEGDFXAEGGGVR	6.5×10^{-13}	ENPEP (APA, Aminopeptidase A) is an N-terminal amino peptidase; association with ratios between fibrinogen A-alpha peptide ADSGEGDFXAEGGGVR[+] and its N-terminal cleaved form DSGEGDFXAEGGGVR[+] suggests that fibrinogen is a substrate of ENPEP	ENPEP plays a role in the catabolic pathway of the renin-angiotensin system and regulates blood pressure, association with blood pressure in Asian population
AKR1C rs2518049	androsterone sulfate / epiandrosterone sulfate	6.7×10^{-13}	AKR1C isoforms play a role in androgen+ metabolism	AKR1C plays a role in the etiology of different cancers, including prostate, brain, breast, bladder and leukemia; potential target of jasmonates in cancer cells
NT5E rs494562	inosine	7.4×10^{-13}	Inosine[+] is a substrate of the 5′-nucleotidase NT5E	NT5E is involved in purine salvage
PRODH rs2023634	proline	2.0×10^{-22}	PRODH catalyzes the first step in proline[+] degradation	
HPS5 rs2403254	alpha-hydroxyisovalerate	1.0×10^{-20}	Alpha-hydroxyisovalerate[+] is found in urine of patients with phenylketonuria, phenylalanine is required for melatonin biosynthesis	Melatonin homeostasis is deranged in patients with loss of *HPS* genes (albinism)
ALPL rs10799701	ADpSGEGDFXAEGGGVR / DSGEGDFXAEGGGVR	2.9×10^{-20}	ALPL is a phosphatase and associates with A-alpha fibrinogen phosphorylation[+]; additive effect on fibrinogen A-alpha phosphorylation together with *ABO* and of *FUT2*.	
SLC7A6 rs6499165	glutaroyl carnitine / lysine	9.8×10^{-19}	Glutaryl-CoA[+] is an intermediate in the metabolism of lysine[+] and tryptophan;	Deficiencies in glutaryl-CoA DH are linked to metabolic disorders
KLKB1 rs4253252	bradykinin, des-arg(9)	6.6×10^{-18}	Kallikrein B, plasma (Fletcher factor) 1; kallikrein-kininogen complex binds to cell surface receptors leading to the targeted action of bradykinin[+]	Association of bradykinin[+] with hypertension confirmed in this study; potential genetic interaction with ACE locus
GLS2 rs2657879	glutamine	3.1×10^{-17}	GLS2 catalyzes the hydrolysis of glutamine[+]	
PDXDC1 rs7200543	1-eicosatrienoylglycerophosphocholine / 1-linoleoylglycerophosphocholine	4.5×10^{-16}	Association with 1-eicosadienoyl- to 1-eicosatrienoyl-glycerophosphocholine[+] ratio suggests role of PDXDC1 in the metabolism of C20:2 and C20:3 fatty acids	Association with body height
SLC22A4 rs272889	isovalerylcarnitine	7.4×10^{-16}	SLC22A4 (OCTN1) transports isovalerylcarnitine[+]	Association with body height
AHR rs12670403	caffeine / quinate	4.8×10^{-15}	AHR is a transcription factor for CYP1A1, which metabolizes caffeine[+]	
ETFDH rs8396	decanoylcarnitine	5.5×10^{-15}	Decanoylcarnitine[+] used for energy production via beta oxidation to electron transfer complex	ETFDH is a key enzyme in the mitochondrial fatty acid beta-oxidation
ELOVL2 rs9393903	docosahexaenoate (DHA; 22:6n3) / eicosapentaenoate	1.7×10^{-14}	EPA (20:5n3)[+] is a substrate of ELOVL2, DHA (22:6n3)[+] is related to	

Europe PMC Funders Author Manuscripts

Locus & SNP id	Metabolic trait	p-value	Relationship between gene function and the associated metabolic traits	Biomedical and pharmaceutical interest
	(EPA; 20:5n3)		its product through by a single desaturation reaction	
SLC16A9 rs7094971	carnitine	3.4×10^{-14}	SLC16A9 (MCT9) transports free carnitine^{+} (shown in this paper)	
IVD rs10518693	3-(4-hydroxyphenyl)lactate / isovalerylcarnitine	1.1×10^{-13}	Isovalerylcarnitine^{+} is a transport form of isovalerate, which is the substrate isovaleryl coenzyme A dehydrogenase (IVD)	IVD is a key enzyme in the mitochondrial fatty acid beta-oxidation
SLC16A10 rs7760535	isoleucine / tyrosine	1.4×10^{-12}	*SLC16A10* encodes the T-type amino acid transporter-1 (TAT1); this transporter transports tyrosine^{+} and phenylalanine^{+}	