ARTICLE

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†, Panos Deloukas[4], Jeanette Erdmann[9], Elin Grundberg[4,10], Christopher J. Hammond[10], Martin Hrabé de Angelis[11,12], Gabi Kastenmüller[1], Anna Köttgen[13], Florian Kronenberg[14], Massimo Mangino[10], Christa Meisinger[15], Thomas Meitinger[16,17], Hans-Werner Mewes[1,8], Michael V. Milburn[6], Cornelia Prehn[11], Johannes Raffler[1,2], Janina S. Ried[5], Werner Römisch-Margl[1], Nilesh J. Samani[18], Kerrin S. Small[10], H.-Erich Wichmann[19,20,21], Guangju Zhai[10], Thomas Illig[22], Tim D. Spector[10], Jerzy Adamski[11,12], Nicole Soranzo[4]* & Christian Gieger[5]*

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 show effect sizes that are unusually high for GWAS and account for 10–60% differences in metabolite levels per allele copy. Our associations provide new functional insights for many disease-related associations that have been reported in previous studies, including those for cardiovascular and kidney disorders, type 2 diabetes, cancer, gout, venous thromboembolism and Crohn's disease. The 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 to the development of safe and efficient therapies, to diagnosis and to prevention. 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 (so-called genetically determined metabotypes (GDMs)) to understand the potential relevance of genetic variants for biomedical and pharmaceutical research[3,4]. Building on this previous work, we present here the most comprehensive evaluation of genetic variance in human metabolism so far, combining genetics and metabolomics for hypothesis generation in a GWAS. We used an extensive, non-targeted and metabolome-wide panel of small molecules, analysing >250 metabolites from 60 biochemical pathways in serum samples from 2,820 individuals from two large population-based European cohorts. We identified 37 genetic loci that were significant at a stringent genome-wide threshold. In contrast to most GWAS, these loci showed exceptionally large effect sizes of 10–60% per allele copy in 25 loci. In the majority of cases, a protein that is 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 for the functions of these loci. We have made a knowledge-base resource publically available via a web-server to aid future functional studies, and biological as well as clinical interpretation of GWAS findings. This study provides compelling evidence for novel associations of metabolic traits with a wide range of loci of biomedical and pharmaceutical interest, and indicates a powerful new paradigm for dissecting human metabolic and disease pathways.

Study design

Metabolic profiling was done on fasting serum from participants in the German KORA F4 study ($n = 1{,}768$) and the British TwinsUK study

[1]Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München, German Research Center for Environmental Health, Ingolstädter Landstraße 1, 85764 Neuherberg, Germany. [2]Faculty of Biology, Ludwig-Maximilians-Universität, Großhaderner Str. 2, 82152 Planegg-Martinsried, Germany. [3]Department of Physiology and Biophysics, Weill Cornell Medical College in Qatar, Education City, Qatar Foundation, PO Box 24144, Doha, State of Qatar. [4]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1HH, UK. [5]Institute of Genetic Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Ingolstädter Landstraße 1, 85764 Neuherberg, Germany. [6]Metabolon Inc., Durham, PO Box 110407, Research Triangle Park, North Carolina 27709, USA. [7]School of Life Sciences, Oxford Brookes University, Gipsy Lane, Headington, Oxford OX3 0BP, UK. [8]Department of Genome-oriented Bioinformatics, Life and Food Science Center Weihenstephan, Technische Universität München, Alte Akademie 1, 85354 Freising, Germany. [9]Universität zu Lübeck, Medizinische Klinik II, Ratzeburger Allee 160, 23538 Lübeck, Germany. [10]Department of Twin Research & Genetic Epidemiology, King's College London, St Thomas' Hospital Campus, 1st Floor South Wing Block, 4 Westminster Bridge Road, London SE1 7EH, UK. [11]Institute of Experimental Genetics, Genome Analysis Center, Helmholtz Zentrum München, German Research Center for Environmental Health, Ingolstädter Landstraße 1, 85764 Neuherberg, Germany. [12]Institute of Experimental Genetics, Life and Food Science Center Weihenstephan, Technische Universität München, Alte Akademie 1, 85354 Freising, Germany. [13]Renal Division, University Hospital Freiburg, Breisacherstrasse 66, 79106 Freiburg, Germany. [14]Division of Genetic Epidemiology, Department of Medical Genetics, Molecular and Clinical Pharmacology, Innsbruck Medical University, Christoph Probst Platz 1, 6020 Innsbruck, Austria. [15]Institute of Epidemiology II, Helmholtz Zentrum München, German Research Center for Environmental Health, Ingolstädter Landstraße 1, 85764 Neuherberg, Germany. [16]Institute of Human Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health, Ingolstädter Landstraße 1, 85764 Neuherberg, Germany. [17]Institute of Human Genetics, Klinikum rechts der Isar, Technische Universität München, Ismaninger Straße 22, 81675 München, Germany. [18]Department of Cardiovascular Sciences, University of Leicester and Leicester NIHR Biomedical Research Unit in Cardiovascular Disease, Glenfield Hospital, University Road, Leicester LE1 7RH, UK. [19]Institute of Epidemiology I, Helmholtz Zentrum München, German Research Center for Environmental Health, Ingolstädter Landstraße 1, 85764 Neuherberg, Germany. [20]Institute of Medical Informatics, Biometry and Epidemiology, Chair of Epidemiology, Ludwig-Maximilians-Universität, Geschwister-Scholl-Platz 1, 80539 München, Germany. [21]Klinikum Grosshadern, Marchioninistraße 15, 81377 München, Germany. [22]Research Unit of Molecular Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Ingolstädter Landstraße 1, 85764 Neuherberg, Germany.

*These authors contributed equally to this work.

†A list of authors and their affiliations appears in Supplementary Information.

($n = 1{,}052$), using ultrahigh-performance liquid-phase chromatography and gas-chromatography separation, coupled with tandem mass spectrometry[5–7]. We achieved highly efficient profiling (24 min per sample) with low median process variability (<12%) of more than 250 metabolites, covering more than 60 biochemical pathways of human metabolism (Supplementary Table 1). On the basis of 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-analysing more than 37,000 metabolites and ratios, we applied a staged approach for selection of promising association signals (Supplementary Fig. 1). In the initial screening stage, we assessed the association of approximately 600,000 genotyped single nucleotide polymorphisms (SNPs) with more than 37,000 metabolic traits (concentrations and their ratios) by fitting linear models separately in both cohorts to log-transformed metabolic traits and adjusting for age, gender and family structure (Supplementary Fig. 2 and Supplementary Table 2). Next, we selected all association signals showing suggestive evidence of 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 reassessed the 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 Methods for details). The combination of SNP and trait that yielded 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 $P < 2.0 \times 10^{-12}$.

Study results

We identified 37 independent loci that reached genome-wide significance in the meta-analysis (Table 1 and Supplementary Tables 3 and 4). Twenty-three 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 the locations of SNPs in genes, on known gene functions and on regional association plots (Supplementary Fig. 2) to prioritize plausible candidate genes within associated loci. In most cases, our annotation was further supported by a statistical analysis of the association of gene relationships in published literature[9] (Supplementary Table 5). Associations with additional metabolic traits at the 37 loci presented in Table 1 may capture further biochemical information, and are provided as Supplementary Table 6. At 30 loci, the sentinel SNP mapped to a protein that was biochemically linked to the associating metabolites, for instance because it was responsible for their synthesis, degradation or metabolism. Next, we searched literature and databases extensively (see web links in Methods) to identify which of these 37 loci were previously reported as being associated with a clinical endpoint, a medically relevant intermediate phenotype, or a pharmacogenetic effect. Associations of metabolites with disease loci can be used to gain novel information about possible metabolic changes associated with biological processes underlying that association (Fig. 1, Table 1 and Supplementary Table 7). In 15 cases, such a relationship could be identified on the basis of an association of the lead SNP or a proxy ($r^2 \geq 0.8$) with the disease-associated SNPs, including those for cardiovascular disease, kidney disease, Crohn's disease, gout, cancer, adverse reactions to drug therapy and predisposing risk factors for diabetes and cardiovascular disease. In all except three loci, the SNPs are common, with minor allele frequencies greater than 10%. In 25 cases, the effect size per allele copy is larger than 10%, and up to 60% in the case of the acyl-CoA dehydrogenase (*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 aetiology of these common diseases is complex and testable hypotheses are needed to develop new avenues for diagnosis and therapy. Associations of known disease-risk loci with metabolic traits allow the identification of new and potentially relevant biological processes and pathways. Here we report some examples from our study that illustrate this idea. The full association data set is freely available for further analysis and reference at http://www.gwas.eu.

Detoxification and kidney failure

N-acetylation is an important mechanism to detoxify nephrotoxic medications and environmental toxins. A reduced ability to detoxify such substances could lead to impaired kidney function. A key GDM is the N-acetyltransferase 8 (*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 investigated whether N-acetylornithine concentrations were associated with kidney function. In both our studies, we found a clear association with estimated glomerular filtration rate (eGFR), whereby higher levels of N-acetylornithine were correlated with lower eGFR ($P_{\mathrm{KORA}} = 7.6 \times 10^{-4}$, $P_{\mathrm{TwinsUK}} = 3.6 \times 10^{-8}$ after adjusting for age and gender). In accord with the genetic effect of the *NAT8* polymorphism in chronic kidney disease (CKD), the risk allele identified here was associated with higher N-acetylornithine concentrations. Although causality cannot be inferred from this kind of association study, the role of ornithine acetylation in the aetiology of CKD warrants further exploration.

Diabetes

Glucokinase (hexokinase 4) regulator (*GCKR*) is a major pleiotropic risk locus associated with diabetes- and cardiometabolic-related traits, such as fasting glucose and insulin levels[12], triglyceride levels[13] and CKD[11]. Here we identified a highly significant association of this locus with mannose:glucose ratios. The fasting level of mannose is lower in carriers of the risk allele, as opposed to glucose being higher. Notably, we also observed a 3.3% increase in lactate concentration 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*, indicate a need for further investigation of 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-α peptides can be detected. Fibrinogen has a role in the formation of blood clots. Its active form, the fibrinogen A-α chain ADSGEGDFXAEGGGVR, can be phosphorylated at serine 3 to ADpSGEGDFXAEGGGVR[16]. The ratio between the concentrations of these fibrinogen A-α peptides provides a measure for fibrinogen A-α phosphorylation (FAαP). Increased levels of FAαP have been observed under various physiological and pathophysiological conditions[17]. Here, three loci (*ABO*, *ALPL* and *FUT2*) associated with FAαP. Notably, these three genes are functionally linked: *ABO* (encoding ABO blood group (transferase A, α-1-3-N-acetylgalactosaminyltransferase; transferase B, α-1-3-galactosyltransferase)) and *FUT2* (encoding fucosyltransferase 2) are involved in determining the blood group, and the *ABO* locus is associated with blood levels of the alkaline phosphatase ALPL[18]. The association of *ALPL* with FAαP may be explained either by a genotype-dependent dephosphorylation of fibrinogen by ALPL, or by a genotype-dependent change in the phosphorus pool available for FAαP. Variants in the *ABO* gene are associated with many different outcomes, including venous thromboembolism (VTE)[19]. The association of *ABO* with FAαP, and thus with modified blood coagulation properties, provides a

Table 1 | Thirty-seven loci that displayed genome-wide significance in the meta-analysis

Locus & SNP id	Metabolic trait	*P* value	Relationship between gene function and the associated metabolic traits	Biomedical and pharmaceutical interest
ACADS rs2066938	Butyrylcarnitine/propionylcarnitine	$< 4.4 \times 10^{-305}$	Butyrylcarnitine^{+} and propionylcarnitine^{+} are substrates/products of ACADS	ACADS is a key enzyme in mitochondrial fatty acid β-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-arachidonoylglycerophosphoethanolamine/ 1-linoleoylglycerophosphoethanolamine	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 and triglycerides**, **fasting glucose and homeostatic model assessment B (HOMA-B) Crohn's disease** and **resting heart rate**
UGT1A rs887829	Bilirubin (E,E)/oleoylcarnitine	2.9×10^{-74}	Bilirubin^{+} is a substrate of UGT1A1	Association with **hyperbilirubinaemia**; low serum concentrations of bilirubin associate with increased risk of CAD; 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 mitochondrial fatty acid β-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 catalyses the Δ-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 has 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**; **serum creatinine (eGFRcrea)**, **Crohn's disease** and **hypertriglyceridaemia**
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** and toxicities to docetaxel and thalidomide treatment
CYP3A4 rs17277546	Androsterone sulphate	8.7×10^{-40}	CYP3A cytochrome P450 proteins metabolise androsterone sulphate^{+}	Genetic variation 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 with fibrinogen peptide phosphorylation^{+}; additive effect on fibrinogen A-α phosphorylation together with FUT2 and ALPL	Association with **blood alkaline phosphatase level**; **pancreatic cancer**; **venous thromboembolism** and **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 ω-hydroxylases; 10-undecenoate (11:1n1)$^{+}$ is biochemically related to ω-hydroxylated C10 fatty acids	Possible role in the etiology of hepatic steatosis, in interaction with stearoyl-coenzyme A desaturase 1
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 indicates 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 indicates a role of LACTB in the butanoate/succinate^{+} pathway	*LACTB* transgenic 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 variations in the *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 an H antigen, and has an additive effect on fibrinogen A-α phosphorylation together with ABO and ALPL	Association with **vitamin B12 levels**, **total cholesterol** and **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

Table 1 | Continued

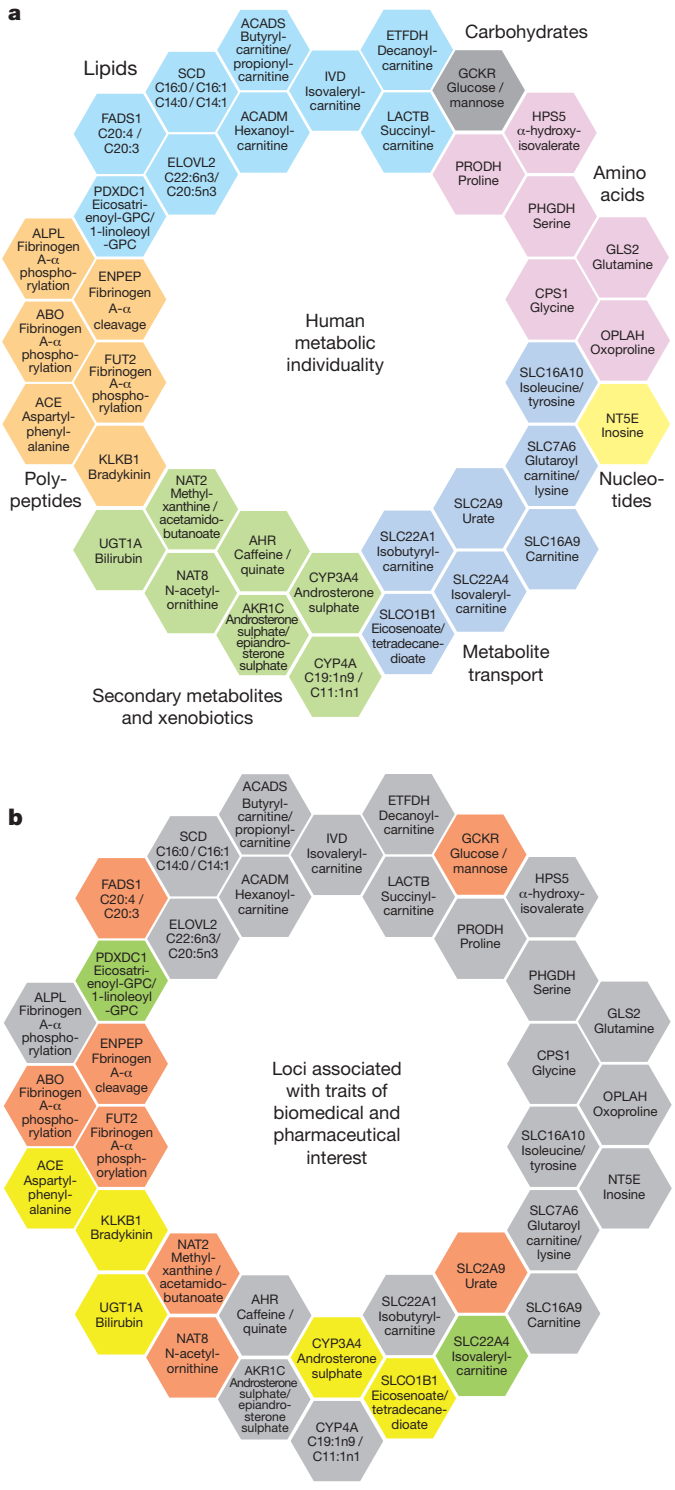
Locus & SNP id	Metabolic trait	*P* value	Relationship between gene function and the associated metabolic traits	Biomedical and pharmaceutical interest
PHGDH rs477992	Serine	2.6×10^{-14}	PHGDH catalyses the first and rate-limiting step in the phosphorylated pathway of serine^{+} biosynthesis	
ENPEP rs2087160	ADpSGEGDFXAEGGGVR/DSGEGDFXAEGGGVR	6.5×10^{-13}	ENPEP (APA, aminopeptidase A) is an amino-terminal amino peptidase; association with ratios between fibrinogen A-α peptide ADSGEGDFXAEGGGVR^{+} and its N-terminal cleaved form DSGEGDFXAEGGGVR^{+} indicate that fibrinogen is a substrate of ENPEP	ENPEP has 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 sulphate/epiandrosterone sulphate	6.7×10^{-13}	AKR1C isoforms have a role in androgen^{+} metabolism	AKR1C has a role in the etiology of cancers including prostate, brain, breast, bladder and leukaemia; 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 catalyses the first step in proline^{+} degradation	
HPS5 rs2403254	α-hydroxyisovalerate	1.0×10^{-20}	α-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-α fibrinogen phosphorylation^{+}; it has an additive effect on fibrinogen A-α 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 dehydrogenase 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 catalyses the hydrolysis of glutamine^{+}	
PDXDC1 rs7200543	1-eicosatrienoylglycerophosphocholine/ 1-linoleoylglycerophosphocholine	4.5×10^{-16}	Association with the 1-eicosadienoyl- to 1-eicosatrienoyl-glycerophosphocholine^{+} ratio indicates a 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 metabolises caffeine^{+}	
ETFDH rs8396	Decanoylcarnitine	5.5×10^{-15}	Decanoylcarnitine^{+} is used for energy production via β-oxidation to the electron transfer complex	ETFDH is a key enzyme in mitochondrial fatty acid β-oxidation
ELOVL2 rs9393903	Docosahexaenoate (DHA; 22:6n3)/eicosapentaenoate (EPA; 20:5n3)	1.7×10^{-14}	EPA (20:5n3)$^{+}$ is a substrate of ELOVL2; DHA (22:6n3)$^{+}$ is related to its product 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 of isovaleryl coenzyme A dehydrogenase (IVD)	IVD is a key enzyme in mitochondrial fatty acid β-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^{+}	

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 Supplementary Tables 3 & 5 and at http://www.gwas.eu. The loci are labelled by the gene that is considered most likely to carry the causative SNP. Where the metabolic trait is consistent with a nearby gene's function, details are provided in the column labelled 'Relationship between gene function and the associated metabolic traits'. Overlaps with associations from other GWAS studies are highlighted in bold ($R^2 > 0.8$, details are in Supplementary Table 6). $^{+}$, Metabolic traits that are associated with the SNP at the corresponding locus. Further information and full bibliographic references are presented in Supplementary Table 4.

functional explanation for the reported association of *ABO* with VTE risk. Moreover, if FAαP is at the basis of VTE, then *FUT2* and *ALPL* should also be investigated as VTE risk genes. This hypothesis can now be tested in the respective patient groups.

Coronary artery disease

We have shown previously[4] that strong associations with metabolic traits, derived from GWAS, can point to interesting associations with clinical endpoints that would not otherwise be considered relevant. A recent meta-analysis with lipid traits[20], using a similar strategy, identified several genetic loci that were found to affect the risk of coronary artery disease (CAD) in the CARDIoGRAM study[21]. Six of these loci are also reported here (*ABO*, *NAT2*, *CPS1*, *NAT8*, *ALPL* and *KLKB1*), although some of them showed only weak evidence for association with CAD ($P < 0.01$) in the CARDIoGRAM study (Supplementary Table 8). Although the links are not statistically strong, the



biochemical function of the associated metabolic traits identified here may support a possible role in heart disease. For example, *NAT8* may be linked to CKD via ornithine acetylation (see above). *KLKB1*, encoding kallikrein B plasma (Fletcher factor) 1, controls blood pressure via the bradykinin pathway. In this study, a genetic variant in *KLKB1* was associated with bradykinin concentrations and we also confirmed the expected directional association of bradykinin with hypertension in both our studies ($P_{KORA} = 1.7 \times 10^{-9}$, $P_{TwinsUK} = 0.0495$, with the covariates age and gender). *ABO* and *ALPL* associated with FAαP, and we therefore speculate that genetically determined differences in FAαP and resulting blood-coagulation properties may be the basis of these associations with CAD. Furthermore, our associations indicate that the role of FAαP as a biomarker for acute myocardial infarction, and the combined additive genetic effect of the *ABO*, *ALPL* and *FUT2* loci (Supplementary Fig. 4) on CAD risk, should be investigated in greater detail.

Figure 1 | Genetic basis of human metabolic individuality and its overlap with loci of biomedical and pharmaceutical interest. More than 100 years ago, it was realised 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) that we have 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. **a**, GDMs are shown colour-coded 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). **b**, GDMs colour-coded by overlap with associations in previous GWAS with disease (red), intermediate risk factors for disease (yellow) and other traits (green). Locus overlap is defined here by the lead SNP reported in the national human genome research institute (NHGRI) GWAS catalogue being highly correlated ($R^2 \geq 0.8$) with the most associated SNP in the metabolomics scan (see column 5 in Table 1 and Supplementary 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 ($P < 3 \times 10^{-6}$, see Methods).

New biological and functional insights

Genome-wide association studies merely uncover statistically significant associations, and can therefore only generate biological hypotheses. Although providing experimental validation of all associations is beyond the scope of a single study, we nevertheless attempted to show that, in principle, validation is possible. The association of SNP rs7094971 in solute carrier family 16, member 9

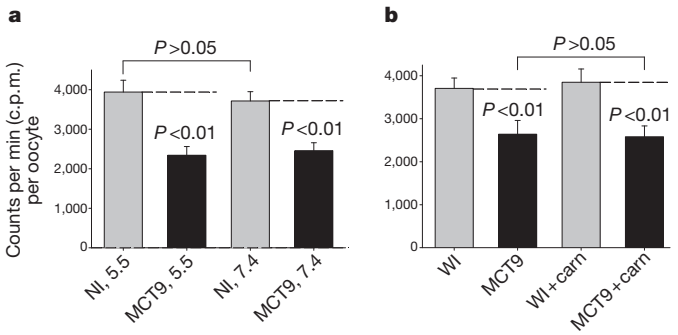


Figure 2 | Experimental evidence for SLC16A9 (MCT9) as a carnitine efflux transporter. **a**, **b**, When 4.6 nl of [^{3}H]-carnitine was injected into *Xenopus* oocytes, followed by incubation in medium for 90 min, efflux was significantly higher in oocytes expressing *MCT9* than in the non-injected (NI, **a**) or water-injected (WI, **b**) controls. By contrast, when oocytes were incubated in medium containing [^{3}H]-carnitine (4 μCi ml^{-1}), there was no significant uptake, indicating that MCT9 does not mediate carnitine uptake (data not shown). Because some previously characterized monocarboxylic acid transporters are proton-coupled[37], the experiments were conducted at both pH_{out} 7.4 and pH_{out} 5.5, but no significant difference was observed (**a** and data not shown). In agreement with this, 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 5 mM carnitine (MCT9 versus MCT9 + carn, **b**). Data are means ± s.e.m. of 6–10 oocytes per data point from 2 oocyte preparations. The *y*-axes represent remaining [^{3}H]-carnitine levels (c.p.m. per oocyte). Statistical significance was determined by the Student's *t*-test. These results are consistent with MCT9 acting as a unidirectional carnitine efflux system when expressed in *Xenopus* oocytes. 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, after absorption via the well-characterized SLC22 (also known as OCTN) family of apical epithelial proton-coupled carnitine transporters[38].

(*SLC16A9*, also known as *MCT9*) with carnitine indicated 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 Fig. 2, our data show that SLC16A9 is a 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 association between *SLC2A9* variants and urate levels was also observed here. Although it was 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 should be motivated and guided by our association data. Among the 37 GDMs reported here, we suggest that the associations with coarsely-characterized genes for enzymes and transporters that are known disease-risk loci may warrant further experimental investigation, for instance in experiments using isotope-labelled derivatives of the associated metabolites reported here as putative target substrates. We deem *NAT8* to be a prime candidate for such a study.

Pharmacogenomics

Using the pharmacogenomics knowledge base[24], we identified six GDMs as being previously associated with toxicity or adverse reactions to medication. Noteworthy are polymorphisms in the *NAT2* and *CYP4A* loci that associated with toxicities to docetaxel and thalidomide treatment[25]; the *UGT1A* locus with irinotecan toxicity[26], *SLC2A9* with the IC_{50} of etoposide[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-dependent 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 the 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. This applies specifically to *AKR1C*, which is a novel target of jasmonates in cancer cells[31]. We reported a GDM associated with *AKR1C* which has a large effect-size on androgen metabolism. The influence of SNP rs2518049 in *AKR1C* on the efficiency and potential side effects of jasmonates should therefore be assessed in future clinical trials.

Discussion

Owing to their large effect-size and high explained variance, the 37 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* with bradykinin and hypertension, and *NAT8* with N-acetylornithine and 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. On the basis of QQ-plots and coarse assumptions, we estimate that more than 500 loci with signals of association below that conservative threshold may be confirmed as GDMs in more highly powered studies in the future. Technically, it is of note that by using a single study to profile 2,820 individuals metabolically, using only 100 µl of blood serum, we replicated in this study a wide series of findings from previous large GWAS with quantitative traits, including serum levels of fasting glucose[12], bilirubin[32,33], urate[34] and dehydroisoandrosterone sulphate[35]. Our study shows how GWAS with intermediate traits that are close to the underlying biological processes can provide new functional insights into associations from GWAS with the endpoints of complex chronic disease 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 to be the next big step towards a full understanding of the interaction between genetic predispositions and environmental factors in the development of complex chronic diseases, their diagnosis, prevention and safe and efficient therapy.

METHODS SUMMARY

The full Methods section provides information about study design, genetic and metabolic data collection, and data analysis.

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

Received 22 November 2010; accepted 30 June 2011.

1. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA* **106,** 9362–9367 (2009).
2. Newgard, C. B. & Attie, A. D. Getting biological about the genetics of diabetes. *Nature Med.* **16,** 388–391 (2010).
3. Illig, T. *et al.* A genome-wide perspective of genetic variation in human metabolism. *Nature Genet.* **42,** 137–141 (2010).
4. Gieger, C. *et al.* Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. *PLoS Genet.* **4,** e1000282 (2008).
5. Evans, A. M., DeHaven, C. D., 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. *Anal. Chem.* **81,** 6656–6667 (2009).
6. Ohta, T. *et al.* Untargeted metabolomic profiling as an evaluative tool of fenofibrate-induced toxicology in Fischer 344 male rats. *Toxicol. Pathol.* **37,** 521–535 (2009).
7. Suhre, K. *et al.* Metabolic footprint of diabetes: a multiplatform metabolomics study in an epidemiological setting. *PLoS ONE* **5,** e13953 (2010).
8. Altmaier, E. *et al.* Bioinformatics analysis of targeted metabolomics—uncovering old and new tales of diabetic mice under medication. *Endocrinology* **149,** 3478–3489 (2008).
9. Raychaudhuri, S. *et al.* Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet.* **5,** e1000534 (2009).
10. Chambers, J. C. *et al.* Genetic loci influencing kidney function and chronic kidney disease. *Nature Genet.* **42,** 373–375 (2010).
11. Köttgen, A. *et al.* New loci associated with kidney function and chronic kidney disease. *Nature Genet.* **42,** 376–384 (2010).
12. Dupuis, J. *et al.* New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nature Genet.* **42,** 105–116 (2010).
13. Aulchenko, Y. S. *et al.* Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. *Nature Genet.* **41,** 47–55 (2009).
14. Panneerselvam, K. & Freeze, H. H. Mannose enters mammalian cells using a specific transporter that is insensitive to glucose. *J. Biol. Chem.* **271,** 9417–9421 (1996).
15. Taguchi, T. *et al.* Hepatic glycogen breakdown is implicated in the maintenance of plasma mannose concentration. *Am. J. Physiol. Endocrinol. Metab.* **288,** E534–E540 (2005).
16. Blombaeck, B., Blombaeck, M., Edman, P. & Hessel, B. Amino-acid sequence and the occurrence of phosphorus in human fibrinopeptides. *Nature* **193,** 833–834 (1962).
17. Martin, S. C., Ekman, P., Forsberg, P. O. & Ersmark, H. Increased phosphate content of fibrinogen *in vivo* correlates with alteration in fibrinogen behaviour. *Thromb. Res.* **68,** 467–473 (1992).
18. Yuan, X. *et al.* Population-based genome-wide association studies reveal six loci influencing plasma levels of liver enzymes. *Am. J. Hum. Genet.* **83,** 520–528 (2008).
19. Tregouet, D. A. *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* **113,** 5298–5303 (2009).
20. Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466,** 707–713 (2010).
21. Schunkert, H. *et al.* Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nature Genet.* **43,** 333–338 (2011).
22. Döring, A. *et al.* SLC2A9 influences uric acid concentrations with pronounced sex-specific effects. *Nature Genet.* **40,** 430–436 (2008).
23. Caulfield, M. J. *et al.* SLC2A9 is a high-capacity urate transporter in humans. *PLoS Med.* **5,** e197 (2008).
24. Klein, T. E. *et al.* Integrating genotype and phenotype information: an overview of the PharmGKB project. *Pharmacogenomics J.* **1,** 167–170 (2001).

25. Deeken, J. F. *et al.* A pharmacogenetic study of docetaxel and thalidomide in patients with castration-resistant prostate cancer using the DMET genotyping platform. *Pharmacogenomics J.* **10,** 191–199 (2010).
26. Lankisch, T. O. *et al.* Gilbert's Syndrome and irinotecan toxicity: combination with UDP-glucuronosyltransferase 1A7 variants increases risk. *Cancer Epidemiol. Biomarkers Prev.* **17,** 695–701 (2008).
27. Huang, R. S. *et al.* A genome-wide approach to identify genetic variants that contribute to etoposide-induced cytotoxicity. *Proc. Natl Acad. Sci. USA* **104,** 9758–9763 (2007).
28. Chen, Y. *et al.* Effect of genetic variation in the organic cation transporter 2 on the renal elimination of metformin. *Pharmacogenet. Genomics* **19,** 497–504 (2009).
29. Shu, Y. *et al.* Effect of genetic variation in the organic cation transporter 1, OCT1, on metformin pharmacokinetics. *Clin. Pharmacol. Ther.* **83,** 273–280 (2008).
30. The SEARCH Collaborative Group. *SLCO1B1* variants and statin-induced myopathy—a genomewide study. *N. Engl. J. Med.* **359,** 789–799 (2008).
31. Davies, N. J. *et al.* AKR1C isoforms represent a novel cellular target for jasmonates alongside their mitochondrial-mediated effects. *Cancer Res.* **69,** 4769–4775 (2009).
32. Sanna, S. *et al.* Common variants in the *SLCO1B3* locus are associated with bilirubin levels and unconjugated hyperbilirubinemia. *Hum. Mol. Genet.* **18,** 2711–2718 (2009).
33. Johnson, A. D. *et al.* Genome-wide association meta-analysis for total serum bilirubin levels. *Hum. Mol. Genet.* **18,** 2700–2710 (2009).
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.* **5,** e1000504 (2009).
35. Zhai, G. *et al.* Eight common genetic variants associated with serum DHEAS levels suggest a key role in ageing mechanisms. *PLoS Genet.* **7,** e1002025 (2011).
36. Mootha, V. K. & Hirschhorn, J. N. Inborn variation in metabolism. *Nature Genet.* **42,** 97–98 (2010).
37. Meredith, D. & Christian, H. C. The SLC16 monocaboxylate transporter family. *Xenobiotica* **38,** 1072–1106 (2008).
38. Koepsell, H. & Endou, H. The SLC22 drug transporter family. *Pflugers Arch.* **447,** 666–676 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements 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, as well as all members of field staff who were involved in the planning and conduct of the MONICA (Monitoring trends and determinants on cardiovascular diseases) and KORA (Kooperative Gesundheitsforschung in der Region Augsburg) studies. The KORA group consists of H. E. Wichmann (speaker), A. Peters, R. Holle, J. John, C.M., T.I. 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. G. Fischer (KORA) and G. Surdulescu (TwinsUK) selected the samples; sample handling and shipment was organized by H. Chavez (KORA) and D. Hodgkiss (TwinsUK); and U. Goebel (Helmholtz) provided administrative support. Special thanks go to D. Garcia-West for his role in facilitating this study. We are grateful to the CARDIoGRAM investigators for access to their data set. Finally, we thank all study participants of the KORA and the TwinsUK studies for donating their blood and time. The KORA research platform and the MONICA studies were initiated and financed by the Helmholtz Zentrum München, National Research Center for Environmental Health, 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 were 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 MeMGenA). 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; the 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 Department 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. T.D.S. 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. N.J.S. holds a British Heart Foundation Chair, 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: T. Young). Genotyping was also performed by CIDR as part of an NEI/NIH project grant. D.M. received support from the Early Career Researcher Scheme at Oxford Brookes University. J.R. is supported by DFG Graduiertenkolleg 'GRK 1563, Regulation and Evolution of Cellular Systems' (RECESS); E.A., by BMBF grant 0315494A (project SysMBo); W.R.-M., by BMBF grant 03IS2061B (project Gani_Med); and B.W., by Era-Net grant 0315442A (project PathoGenoMics). A.K. is supported by the Emmy Noether Programme of the German Research Foundation (DFG grant KO-3598/2-1) and F.K., by grants from the 'Genomics of Lipid-associated Disorders (GOLD)' of the Austrian Genome Research Programme (GEN-AU). N.S. is supported by the Wellcome Trust (core grant number 091746/Z/10/Z).

Author Contributions Designed the study: J.A., C.G., T.I., D.M., N.S. and K.S. Conducted the experiments: D.M., M.V.M. and R.P.M. Analysed the data: J.A., E.A., C.G., G.K., A.K., F.K., C.M., D.M., A.-K.P., C.P., J.R., J.S.R., W.R.-M., S.-Y.S., K.S. and B.W. Provided material, data or analysis tools: the CARDIoGRAM consortium, P.D., J.E., E.G., C.J.H., M.H.d.A., T.I., M.M., T.M., H.-W.M., N.J.S., K.S.S., T.D.S., H.-E.W. and G.Z. Wrote the paper: C.G., N.S. and K.S. All authors read the paper and contributed to its final form.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to K.S. (karsten@suhre.fr) or N.S. (ns6@sanger.ac.uk).

METHODS

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 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 a British adult-twin registry in the age range 8–102 years and 84% are female. The samples used in this study are aged 23–85 (mean age 48 years) and 97% are female. These unselected twins were recruited from the general population through national media campaigns 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 08:00 and 10:30 after a period of at least 10 h overnight fasting. Material was drawn into serum gel tubes, gently inverted twice 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 frozen at −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 of 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 at −45 °C until sampling.

Metabolomics measurements. Metabolon, a 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 gas chromatography/mass spectrometry (GC/MS) injection per sample. The UHPLC injections were 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, peptides, xenobiotics and steroids; a full list of metabolites is given in Supplementary Table 1). The detection of the entire panel was carried out with 24 min of 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 mass to charge ratio (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 on the basis of a multiparameter match without the 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].

Quality control of metabolomics data. 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 (relative standard deviation, upper and lower 95% confidence interval and minimum and maximum observed values in quality control samples) are reported in Supplementary Table 1. 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 three 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. For all individuals profiled from the KORA 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 GWAS and meta-analyses[3,34,43]. We therefore summarize only the essential details here. Genotyping of the KORA F4 population was carried out using the Affymetrix GeneChip array 6.0. Genotypes were determined using the Birdseed2 clustering algorithm. For quality assurance, we applied the criteria of call rate > 95% and P(Hardy-Weinberg) > 10^{-6} as filters for SNP quality: 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 within a region in a meta-analysis of KORA and TwinsUK, we used genotyped SNPs as well as dosages of imputed SNPs. In KORA F4, imputation was done using IMPUTE v0.4.2 (ref. 44) based on HapMap2 (see below).

Genotyping of the TwinsUK data set was done with a combination of Illumina arrays (HumanHap300, HumanHap610Q, 1M-Duo and 1.2MDuo 1M)[45,46]. We pooled the normalized intensity data for each of the three arrays separately (with 1M-Duo and 1.2MDuo 1M pooled together). For each data set, we used the Illluminus calling algorithm[47] to assign genotypes in the pooled data. No calls were assigned if an individual's most likely genotype was called with 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 showing any of these characteristics were discarded.

We applied similar exclusion criteria to each of the three data sets separately. Exclusion criteria for samples were: (1) sample call rate <98%; (2) heterozygosity across all SNPs ≥ 2 s.d. from the sample mean; (3) evidence of non-European ancestry as assessed by principle-component-analysis comparison with HapMap3 populations; (4) observed pairwise identity-by-descent (IBD) probabilities indicative of sample identity errors. We corrected misclassified monozygotic and dizygotic twins on the basis of IBD probabilities. Exclusion criteria for SNPs were: (1) Hardy–Weinberg P value $<10^{-6}$, assessed in a set of unrelated samples; (2) minor allele frequency (MAF) <1%, assessed in a set of unrelated samples; (3) SNP call rate <97% (SNPs with MAF ≥ 5%) or <99% (for 1% ≤ MAF < 5%).

Alleles of all three data sets were aligned to HapMap2 or HapMap3 forward-strand alleles. Before merging, we performed pairwise comparison among the three data sets and further excluded SNPs and samples to avoid spurious genotyping effects, identified as follows: (1) concordance at duplicate samples <1%; (2) concordance at duplicate SNPs <1%; (3) visual inspection of QQ-plots for logistic regression applied to all pairwise data-set comparisons; (4) Hardy–Weinberg P value $<10^{-6}$, assessed in a set of unrelated samples; (5) observed pairwise IBD probabilities indicative of sample identity errors. We then merged the three data sets, keeping individuals typed at the largest number of SNPs when an individual was typed at two different arrays. The merged data set 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 data set (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). The analysis of this study used 534,665 autosomal SNPs (basically 610K SNPs extracted from the final merged data set).

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, the untransformed distribution was closer to a normal distribution. For concentrations, 149 were closer to a log-normal distribution and 124 were better represented by a normal distribution. On the basis of this observation, and also for sake of simplicity, we decided to log-transform all metabolites and their ratios. We used 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 (more than 250) indicates that two metabolites are likely to be functionally linked in a metabolic pathway that has an impact on 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 was 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. 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 a reference, we corrected for tests on 655,658 SNPs and 37,455 metabolic traits, thus obtaining a Bonferroni-adjusted P value of $P = 2.04 \times 10^{-12}$. For ratios, we also required 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), should be larger than the number of tested metabolic traits (P-gain > 250)[4,8]. This limit is considered a Bonferroni-type conservative cutoff for identifying those metabolite concentration pairs for which the use of ratios strongly improves the strength of association. In addition to the strongest associating metabolic trait, others can often provide additional insight into the underlying biochemical processes. In such cases, we consider a P value of 1.33×10^{-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 Supplementary Fig. 3) and there was no inflation of summary statistics, which could be indicative of population stratification (see QQ-plots in Supplementary Fig. 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 (Supplementary Fig. 3) were created using imputed and meta-analysed 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 $P < 1.33 \times 10^{-6}$ (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 Supplementary Table 4. For the strongest associating trait, box plots were plotted to visualize the actual quantitative dependence of the trait on genotype (Supplementary Fig. 3). On the basis of association data alone, it is not possible in most cases to identify the gene within a locus that causes the association. However, using knowledge of the function of genes in LD with the reported SNP, as well as the biochemical characteristics of the associating metabolite, it is possible to identify a single most likely candidate gene in many cases. These cases are tagged as 'match between gene function and metabolic trait' and are supported by arguments provided as Supplementary Text (for example, the association between a SNP in LD with *OPLAH* (oxoprolinase) and oxoproline concentrations). At two loci (*CYP4A* and *UGT1A*), alternative splice variants exist. We named these loci without attempting to specify the exact variant.

GWAS catalogue. Using the catalogue of published GWAS (accessed 10 October 2010)[1], we identified for each entry the SNPs in the KORA and TwinsUK studies that correlate most strongly with a previously reported SNP ($r^2 \geq 0.5$) and that were present in our association database ($P < 10^{-3}$, P-gain > 10). The resulting associations are available online on our GWAS server. New associations will 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 data set, we chose the 655,658 SNPs 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 therefore be considered to represent one draw-out of this set. We then drew 1,000,000 sets of 37 SNPs at random (with replacement) from this sampling data set. To account for comparable MAF distributions between the reference and the random set, we then rejected all draws in which 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 the overlap with the GWAS catalogue for every random set. 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 more than half of the cases, no overlapping SNPs were present in the sampled data set (see Supplementary Table 9). For our reported 37 metabolomics SNPs, we identified 14 overlapping SNPs (note that we report 15 overlapping loci in Fig. 1; the *ENPEP* locus was not yet included in the GWAS catalogue and was not used in this analysis). Because we never found 14 overlapping loci by chance, the P value of our observations being due to chance is less than $1/330{,}775 \approx 3 \times 10^{-6}$.

Functional characterization of SLC16A9. The *SLC16A9* (*MCT9*) clone (IMAGE ID 40146598) was purchased from Autogen Bioclear. Plasmid was linearized with SpeI restriction enzyme (New England Biolabs) and complementary RNA was synthesized *in vitro* using the T7 mMachine *in vitro* transcription system (Ambion). *MCT9* was expressed in *Xenopus laevis* oocytes as described previously[50]. Briefly, oocytes at stage V–VI were injected with 10 ng 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 or an injection of an equal volume (50 nl) of distilled water, and were incubated for the same length of time. Uptake and efflux experiments were performed similarly to those described previously[50] except that the substrate was [^{3}H]-carnitine (specific activity 81 Ci mmol^{-1}, GE Healthcare).

Data access. This study generated millions of individual data points through the profiling of n metabolites ($n = 250$) and $n(n-1)/2$ ratios in about 2,800 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 aim of allowing rapid storage and retrieval of data as well as managing the integration of metabolomics summary statistics vis-à-vis published GWAS studies. The association data are freely available at http://www.gwas.eu and at mirror sites located at the Wellcome Trust Sanger Institute and King's College London sites.

Web links. GWAS server: http://www.gwas.eu, SNAP: http://www.broadinstitute.org/mpg/snap/, NHGRI catalogue 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

39. 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), 26–30 (2005).
40. 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).
41. Lawton, K. A. *et al.* Analysis of the adult human plasma metabolome. *Pharmacogenomics* **9,** 383–397 (2008).
42. Sreekumar, A. *et al.* Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. *Nature* **457,** 910–914 (2009).
43. Soranzo, N. *et al.* A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. *Nature Genet.* **41,** 1182–1190 (2009).
44. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5,** e1000529 (2009).
45. Richards, J. B. *et al.* Bone mineral density, osteoporosis, and osteoporotic fractures: a genome-wide association study. *Lancet* **371,** 1505–1512 (2008).
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.* **5,** e1000445 (2009).
47. Teo, Y. Y. *et al.* A genotype calling algorithm for the Illumina BeadArray platform. *Bioinformatics* **23,** 2741–2746 (2007).
48. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).
49. Abecasis, G. R., Cherny, S. S., Cookson, W. O. & Cardon, L. R. Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genet.* **30,** 97–101 (2002).
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.* **279,** 15795–15798 (2004).