An atlas of genetic influences on human blood metabolites

So-Youn Shin1,21,23, Eric B Fauman2,23, Ann-Kristin Petersen3,23, Jan Krumsiek4,23, Rita Santos5, Jie Huang1, Matthias Arnold6, Idil Erte7, Vincenzo Forgetta8, Tsun-Po Yang1, Klaudia Walter1, Cristina Menni7, Lu Chen1,9, Louella Vasquez1, Ana M Valdes7,10, Craig L Hyde11, Vicky Wang2, Daniel Ziemek2, Phoebe Roberts2,22, Li Xi2, Elin Grundberg8,12, The Multiple Tissue Human Expression Resource (MuTHER) Consortium13, Melanie Waldenberger14, J Brent Richards7,8,15, Robert P Mohney16, Michael V Milburn16, Sally L John17, Jeff Trimmer18,21, Fabian J Theis1,19, John P Overington5, Karsten Suhré6,20,24, M Julia Brosnan11,24, Christian Gieger3,24, Gabi Kastenmüller6,24, Tim D Spector7,24 & Nicole Soranzo1,9,24

Genome-wide association scans with high-throughput metabolic profiling provide unprecedented insights into how genetic variation influences metabolism and complex disease. Here we report the most comprehensive exploration of genetic loci influencing human metabolism thus far, comprising 7,824 adult individuals from 2 European population studies. We report genome-wide significant associations at 145 metabolic loci and their biochemical connectivity with more than 400 metabolites in human blood. We extensively characterize the resulting in vivo blueprint of metabolism in human blood by integrating it with information on gene expression, heritability and overlap with known loci for complex disorders, inborn errors of metabolism and pharmacological targets. We further developed a database and web-based resources for data mining and results visualization. Our findings provide new insights into the role of inherited variation in blood metabolic diversity and identify potential new opportunities for drug development and for understanding disease.

The discovery of mutations causing severe congenital metabolic disorders, or inborn errors of metabolism, revolutionized early understanding of how genes control biochemical reactions and metabolic pathways in the human body.1 Recent technological advances in metabolomics and genetics allowing the collection of high-dimensional data sets in large population samples suggest that inborn errors of metabolism are only extreme cases of a wide spectrum of genetic variation in human metabolism and that the loci involved often influence multifactorial traits and complex diseases. The so-called genetically influenced metabolotypes (GIMs) identified thus far have been shown to display larger effect sizes than most loci associated with complex multifactorial traits and complex diseases.2–12.

Although the biomedical and pharmaceutical relevance of these associations may become clearer as focused gene­by­gene investigations are conducted, little is known about their system­wide interconnectivity or how this knowledge can be translated into medical practice. A comprehensive blueprint of human metabolic pathways and the genes that regulate them would inform strategies for modifying deregulated metabolites in a rational and targeted manner, potentially using already existing drugs, as has been suggested for other genome­wide association study (GWAS) findings.13. In this context, genetic associations provide powerful tools to identify genes that could be targeted to modulate metabolite levels.

Here we present the most comprehensive investigation of genetic influences on human metabolism thus far, extending previous studies based on the same metabolomic platform.1,14,15. We applied powerful hypothesis­generating genome­wide scans to survey regions of the genome associated with a wide range of metabolic traits. The hundreds of associations and their metabolic context reported in this study (GWAS) findings.13. In this context, genetic associations provide powerful tools to identify genes that could be targeted to modulate metabolite levels.

The discovery of mutations causing severe congenital metabolic disorders, or inborn errors of metabolism, revolutionized early understanding of how genes control biochemical reactions and metabolic pathways in the human body.1 Recent technological advances in metabolomics and genetics allowing the collection of high-dimensional data sets in large population samples suggest that inborn errors of metabolism are only extreme cases of a wide spectrum of genetic variation in human metabolism and that the loci involved often influence multifactorial traits and complex diseases. The so-called genetically influenced metabolotypes (GIMs) identified thus far have been shown to display larger effect sizes than most loci associated with complex multifactorial traits and complex diseases.2–12.

Although the biomedical and pharmaceutical relevance of these associations may become clearer as focused gene­by­gene investigations are conducted, little is known about their system­wide interconnectivity or how this knowledge can be translated into medical practice. A comprehensive blueprint of human metabolic pathways and the genes that regulate them would inform strategies for modifying deregulated metabolites in a rational and targeted manner, potentially using already existing drugs, as has been suggested for other genome­wide association study (GWAS) findings.13. In this context, genetic associations provide powerful tools to identify genes that could be targeted to modulate metabolite levels.

Here we present the most comprehensive investigation of genetic influences on human metabolism thus far, extending previous studies based on the same metabolomic platform.1,14,15. We applied powerful hypothesis­generating genome­wide scans to survey regions of the genome associated with a wide range of metabolic traits. The hundreds of associations and their metabolic context reported in this study (GWAS) findings.13. In this context, genetic associations provide powerful tools to identify genes that could be targeted to modulate metabolite levels.
study identify a system-wide atlas of molecular readouts of activity for human genes measured in vivo. We not only provide information on hundreds of single genotype-metabolite associations, greatly expanding knowledge of metabolic intermediates of gene function and disease, but we also generate a network that includes the majority of metabolites reliably measurable in blood using the Metabolon platform, allowing future exploration of perturbations caused by individual genetic variants across hundreds of metabolites at once. This information is based on experimental in vivo data from a human population study, thus complementing classical metabolic pathway maps that have been obtained as composites of single in vitro biochemistry experiments. The newly discovered loci will empower future clinical and pharmacological research in a number of key areas, spanning from a better understanding of genetic predisposition to disease to the identification of potential new biomarkers, drug surveillance tools and drug targets and the causal evaluation of environmental and modifiable influences on human traits and disease. Moreover, to maximize the downstream usefulness of the data for the broader scientific community, we have made the atlas freely available through an extensive suite of web resources, including a database of detailed functional annotations and disease associations for each locus and a network view of the data with links to genetic and metabolic web resources (see URLs). Below we summarize the key results of our study.

RESULTS
Eighty-four newly discovered metabolic loci
The study sample included a total of 529 metabolites profiled using liquid-phase chromatography and gas chromatography separation coupled with tandem mass spectrometry in either plasma or serum from 7,824 adult individuals from 2 European population studies (Supplementary Table 1). The entire KORA (Kooperative Gesundheitsforschung in der Region Augsburg) data set (n = 1,768) and a small proportion of the TwinsUK data set (1,052 individuals and 250 metabolites) have been described in 2 previous studies11,14, with 5,002 TwinsUK individuals newly profiled in this study. Over half of the 529 metabolites (n = 333; 63%) were chemically identified and could be assigned to 8 broad metabolic groups (amino acids, carbohydrates, cofactors and vitamins, energy, lipids, nucleotides, peptides and xenobiotic metabolism) as described in the KEGG (Kyoto Encyclopedia of Genes and Genomes) database16. These groups could be further subdivided into 63 distinct biochemical pathways. Another 196 metabolites (37%) were classified as ‘unknown’, indicating that their chemical identity was not yet determined at the time of analysis. Further information on the unknown metabolites, including measurement platform, retention time, m/z mass-to-charge ratio and fragmentation spectra, can be found in Supplementary Table 2. Analyte overlap between unknown and known metabolites was excluded using correlation analysis (Supplementary Table 3). After stringent quality controls, a subset of 486 metabolites was available for genetic analysis in both cohorts, including 309 known and 177 unknown metabolites.

Primary genome-wide discovery analysis was carried out on approximately 2.1 million SNPs, either directly genotyped or imputed from the HapMap 2 panel12 and passing stringent quality control metrics (Online Methods and Supplementary Fig. 1). In this initial discovery effort, 137 independent variants were significantly associated with metabolite concentrations at a stringent genome-wide P-value cutoff of 1.03 × 10−10 (≈ 5 × 10−8/486 metabolites; Supplementary Tables 4 and 5)18. Subsequent discovery analysis on 98,346 pairwise metabolite ratios identified 8 additional loci at a more stringent cutoff of 5.08 × 10−13 (≈ 5 × 10−8/98,346 ratios; see the Supplementary Note for discussion on the interpretation of ratios). Overall, our study identified 299 SNP-metabolite associations of genome-wide significance at 145 statistically independent SNPs (Fig. 1, Supplementary Fig. 2 and Supplementary Table 5). Of these, 84 loci have not been reported before, and 3 loci were reported while this work was in revision19. Another 61 loci were identified in previous studies, suggesting validation across different platforms2,10,11,14,20. A subset of the loci was previously reported in smaller-scale studies conducted in other tissues, suggesting that metabolite associations are reproducible across tissue types and that blood metabolite loci may for the most part be representative of associations in other tissues. For instance, all 5 genome-wide significant loci detected in a study of urine in a discovery sample of 862 individuals21 were also detected in blood in this study. In another example, the PYROX2 locus was previously reported to be associated with a metabolite in urine detected by 1H nuclear magnetic resonance spectroscopy22. The newly discovered loci identify a rich catalog of metabolic associations, allowing linkage of genetic and disease associations with underlying molecular mechanisms and greatly enriching understanding of the genetic control of human metabolism. We systematically evaluated evidence for the genes within 1-Mb windows centered on the sentinel SNPs to identify cases where the function of the gene matched the relevant metabolite (Online Methods). This effort identified a plausible or established biochemical link for 101 of the 145 loci, involving 94 unique genes (reviewed in Supplementary Table 6). We henceforth refer to these 94 genes as ‘predicted causal’. For the remaining loci, we annotated the gene nearest to the association peak.

Where a predicted causal gene was associated with a metabolite ratio and both metabolites had been identified, we could characterize the ratio in terms of the underlying biochemistry. In 11 cases, the metabolite ratio seemed to reflect flux through a particular metabolic reaction that was influenced by the SNP (annotated as ‘activity’ in Supplementary Table 6). For instance, a variant in GOT2 (encoding mitochondrial glutamic-oxaloacetic transaminase 2) was associated with the ratio between phenyllactate and phenylalanine. GOT2 catalyzes the conversion of phenylalanine to phenylpyruvate, which is then converted into phenyllactate23. Another such example was MBOAT7, associated with the ratio of arachidonate (20:4n6) to 1-arachidonoylglyceryl-phosphatidinositol. MBOAT7 encodes a lysophosphatidylinositol acyltransferase that has specificity for arachidonoyl-CoA as an acyl donor24. Arachidonate is readily converted to arachidonoyl-CoA.

In five cases, either both metabolites in a ratio were linked to a substrate or both were linked to a product (for instance at ACE, SULT2A1, AKR1C4, ABP1 and THEM4). In these cases, the effect of the genetic variant may be explained as causing one molecule to be consumed or acted on faster than the other (‘selectivity’).

In seven cases, the levels of one metabolite ‘normalizing’ the statistical signal for the other may explain the association with the ratio. For instance, PRODH encodes proline dehydrogenase, which catalyzes the first step in proline degradation25. However, the P value for the ratio of valine to proline was even stronger than the P value for proline itself, suggesting that valine may be normalizing the proline concentration against the overall amino acid pool.

A de novo atlas of human metabolic relationships in blood
We next generated a network view of genetic-metabolic interactions in the two cohorts by combining genetic and metabolite information. First, we connected metabolites with metabolites using Gaussian graphical models (GGMs), which are based on partial correlation coefficients. We have previously shown that GGMs connect
biochemically related metabolites,14,26 and can be applied to reconstruct metabolic pathways directly from metabolomics data. Second, we connected metabolites with genetic loci on the basis of our primary GWAS results, with one link for each genome-wide significant association. For a more detailed description of the network generation process, see the Supplementary Note. To verify the stability of partial correlation values in both cohorts, we performed a bootstrapping-based subsampling approach. We observed generally low standard deviation, especially for high partial correlation values, indicating high stability for the estimation. The resulting network (Fig. 2) recapitulates the relationships between 397 metabolites from 60 different pathways (249 known and 148 unknown) and 131 of the 145 genetic loci. To our knowledge, it provides the first comprehensive and high-resolution reference map of human metabolic relationships and their genetic influences measured in a single in vivo‘ experiment’ in blood.

This in vivo reference map complements existing knowledge of gene–metabolite relationships in specialized biochemical databases. It facilitates the visualization of genetic associations in the context of complex relationships between metabolites and SNPs, as illustrated by the subnetwork of dipeptide and oligopeptide metabolites and peptide-related genes (Fig. 2, box). To maximize the usefulness of the data for the scientific community, we have made the network available for download with extensive biochemical and biological annotations (see URLs). We have further made available a version of the network with basic annotation through a web browser, allowing rapid visualization and exploration of the data. Furthermore, an extensive database of genetic associations and their biological, medical and pharmacological annotations is available in Supplementary Table 6 and through the supporting online website (see URLs).

Allelic architecture of metabolic loci

To fully quantify the extent to which metabolic loci capture variance in metabolite concentrations, we carried out exhaustive characterization of the allelic architecture of the newly discovered loci. Metabolite heritabilities were estimated using the classical twin (ACE) model and the twin structure of the TwinsUK cohort (Fig. 3 and Supplementary Note). The contribution of metabolic loci to variance in metabolite concentrations was high (median of 6.9%, range of 1–62%; Supplementary Table 7 and Supplementary Note), with variants explaining greater than 20% of heritability for approximately 10% of the metabolites and greater than 50% of the heritability in four cases. These findings support previous observations that variants explain on average greater proportions of trait variance for some metabolite classes than what is generally observed for complex traits2, confirming the value of these intermediate molecular traits for dissecting genetic contributions to complex traits with greater statistical power. We carried out local imputation with a denser haplotype reference map (1000 Genomes Project) for each of the 145 loci to explore the contribution of additional variants with lower minor allele frequency present in this panel and poorly represented by HapMap 2 imputation. With the exception of the CYP3A cluster, the two imputation panels yielded lead SNPs with highly correlated frequencies, \( P \) values and explained variance at most loci (Supplementary Fig. 3 and Supplementary Table 8).

We further explored the contribution of SNP-SNP interactions between the metabolic loci, defined as a departure from additive marginal effects (Online Methods). This analysis suggested that the effects of metabolic loci were predominantly additive, apart from the statistically significant interaction observed between NAT8 and PYROXD2 variants (Fig. 4, Supplementary Fig. 4 and Supplementary Table 9).
Finally, we systematically compared the metabolite-associated SNPs against public repositories of variants (cis expression quantitative trait loci, cis-eQTLs) affecting gene expression in liver27, fat, skin and lymphoblastoid cell lines (LCLs)28. For each lead metabolomic SNP, we first identified all SNPs with high linkage disequilibrium (r2 ≥ 0.8) in the 1000 Genomes Project pilot phase (CEU population of European ancestry). Each lead SNP and its proxies were then used as baits to search for cis-eQTL SNPs in these four tissues. A total of 57 lead SNPs identified cis-eQTLs in at least 1 of the 4 tissues searched under a nominal permutation P value of < 0.001, defining a total of 101 SNP-gene pairs and 97 different genes (Supplementary Table 10). Of the 97 genes, 38 were predicted to be causal on the basis of our annotations and 59 were predicted to be non-causal. When compared to non-causal genes, causal genes showed 3.25-fold enrichment in the liver eQTL data set (Fisher’s exact test P value = 0.023, two-tailed), possibly reflecting the greater contribution that liver metabolism makes to blood metabolite levels. Furthermore, causal genes were enriched by 1.6-fold in fat compared to non-causal genes (Fisher’s exact test P value = 0.038, two-tailed). No enrichment was seen in LCLs or skin.

One major challenge in interpreting associations from GWAS is formulating and testing hypotheses on the causal effects of a SNP on an associated trait. We argued earlier that molecular QTL studies on metabolite concentrations or gene expression have greater statistical power than with more complex traits such as, for instance, high-density lipoprotein (HDL) or low-density lipoprotein (LDL) cholesterol levels2. In this context, the parallel analysis of SNPs, metabolite concentrations and other molecular phenotypes (for instance, epigenetic profiles or transcript levels) provides new opportunities to investigate causal pathways at the molecular level (Supplementary Note). We applied mendelian randomization analysis to test the hypothesis that metabolite-associated SNPs affect metabolite concentrations through variation in the transcript levels of corresponding causal genes. We focused on a subset of 32 predicted causal genes that had a matching eQTL in at least one MuTHER (Multiple Tissue Human Expression Resource) tissue and used a subset of 484 individuals from the MuTHER data set for whom gene expression levels were measured at the time of metabolomic measurement (Online Methods). This analysis identified two loci that were significant at the Bonferroni-corrected P value of < 1.5 × 10−3, namely THEM4 and CYP3A5. In both cases, the allele associated with increased metabolite levels was associated with decreased gene expression in one or more tissues (Fig. 4 and Supplementary Table 11). This analysis provides support for the notion of underlying causal variants having regulatory consequences for these two loci.
Biological relevance
The newly discovered associations span the large majority of the metabolic pathways explored, indicating widespread genetic influences on the human metabolome. We could assign a known biochemical function to approximately two-thirds of the overall associations (101 of the 145 loci; Supplementary Table 6). One-third (n = 34) of the newly discovered loci involved amino acids, and a similar number (n = 33) were involved with intermediates of lipid metabolism, including sterols, carnitines and intermediates of inositol and fatty acid metabolism. The remaining newly discovered associations were across a wide range of metabolic classes and functions, notably pathways with a central role in cellular metabolism and energy, intermediates of purine and pyrimidine metabolism, glucose homeostasis, and vitamin and cofactor levels, among others. In cases where a metabolite was of unknown identity, the metabolic function of the associated gene provided a hypothesis on its identity, as described previously.14

The current characterization of hundreds of loci embedded in their metabolic context further allows exploration of complex systems to a greater depth and breadth than in previous studies.2,10,11,14,20 For instance, 12 newly discovered associations were within phenylalanine, tyrosine and tryptophan metabolism pathways, implicated in key brain functions through dopamine and serotonin biosynthesis. Among these associations, two common variants in TDO2 (encoding tryptophan 2,3-dioxygenase) and IDO1 (encoding indoleamine 2,3-dioxygenase 1) were associated with the concentrations of tryptophan and 4-hydroxytryptophan (X-12100), respectively, two intermediates in the

Figure 3 Heritability and variance explained. We used the ACE model to partition the variance for each metabolite into narrow-sense heritability and common and unique environmental components. The proportion of heritability explained by all SNPs associated with a given metabolite at the genome-wide level is shown in red. The corresponding numeric values for heritability estimates are given in Supplementary Table 7.
synthesis of the neurotransmitter serotonin. Several loci mapped to transporters, including association of plasma tyrosine and tryptophan levels with SLC16A10 (encoding the T-type amino acid transporter 1 (TAT1) for tryptophan, tyrosine and phenylalanine29), associations of plasma kynurenine levels with SLC7A5 (encoding LAT1), which mediates cellular exchange of tryptophan and the tryptophan metabolite kynurenine30, and many others. It will be important to characterize whether these associations provide reliable molecular readouts of the function of these genes in the brain to ascertain the value of accessible measurable levels and associations will be necessary to address this question.

Disease and pharmacological relevance
Integration of metabolic associations with complex traits and disease empowers understanding of the molecular underpinnings of disease, as shown in the case of the bradykinin–kininogen–kinin system and cardiovascular disease (Supplementary Fig. 5 and Supplementary Note). We searched the National Human Genome Research Institute (NHGRI) GWAS Catalog (July 2013) for cases where the sentinel SNP either matched a variant associated with a complex trait or disease or was in high linkage disequilibrium with one (defined as $r^2 \geq 0.8$; Online Methods). A total of 41 metabolite-associated SNPs (Supplementary Table 4) and 14 of 84 newly discovered SNPs matched SNPs previously associated with complex disease or drug response endpoints (Fig. 5 and Supplementary Table 4). We further searched the 1-Mb intervals centered on the associated SNPs for inborn errors of metabolism or for being targets, metabolizing enzymes or transporters of inborn errors of metabolism or for being targets, metabolizing enzymes or transporters of FDA-approved drugs. Variants were annotated on the basis of their overlap with complex trait and disease loci. New associations reported for the first time in this study are highlighted in bold. The symbols “dev” and “b” identify genes associated with compounds in active stages of drug development (preclinical, phase I–III, preregistration, registration) and bioactive drug-like compounds, respectively. Full details on locus annotation are provided in Supplementary Tables 6, 13 and 14 and in the full version available on the supporting website (see URLs).
genes causative for inborn errors of metabolism, identifying significant enrichment (26 cases, hypergeometric test P value = 5.9 × 10⁻16; Supplementary Table 12). As recently postulated⁴, this suggests that several genes (including CPS1, UGT1A1, CBS, SLC22A4–SLC22A5 and others) harbor genetic variants with effects ranging from loss-of-function alleles in metabolic disorders all the way to common polymorphisms with moderate phenotypic consequences in multifactorial complex diseases and, finally, quantitative variation within the normal range of healthy individuals. One such example was the CPS1 locus, encoding a mitochondrial carbamoyl phosphate synthetase, an enzyme catalyzing the generation of carbamoyl phosphate from ammonia and bicarbonate. Rare mutations in CPS1 cause carbamoyl phosphate synthetase I deficiency, a severe autosomal recessive disorder characterized by congenital hyperammonemia and defective citrulline synthesis. A common variant in CPS1 was previously associated with increased risk of chronic kidney disease⁵,6. The same variant was associated with glycine levels in this and previous studies⁶,10,20,33. Glycine is interconverted to ammonia via the glycine cleavage complex⁷ and provides a molecular intermediate trait for detection of the disease.

The extension of clinical chemistry analysis to a richer set of metabolites in this study and the underlying linkage to genetic differences further present new opportunities to identify variants of possible pharmacogenomic relevance and new pharmacological targets. Of the 132 unique genes associated with the 145 metabolic loci, 56 (42%) identified genes of potential interest for pharmacological development. Of these, 24 (18%) included drug targets (10 genes), drug-metabolizing enzymes (11 genes) or transporters (3 genes) for US Food and Drug Administration (FDA) and/or European Medicine Agency (EMA)-approved drugs. For instance, 45 different FDA- approved drugs with a broad range of indications were linked with the 10 drug targets alone (Supplementary Table 13). A further 11 genes encoded drug targets for compounds in early to late stages of development (from preclinical to phase I, II and III trials through registration), and 24 genes were linked to drugs that are either suspended or discontinued or have no reported development activity (Supplementary Table 14). Thus, overall, 21 of the 132 reported genes (16%) were either established or promising drug targets. This finding indicates that druggable targets are highly enriched within metabolic loci (2.8-fold enrichment, χ² test P value = 1.19 × 10⁻⁶) compared to the estimated druggable fraction of human genes (1,089 of 19,258 human genes or 5.6%; ref. 14). Finally, 21 additional genes were targets, drug-metabolizing enzymes or transporters for bioactive drug-like compounds (Supplementary Table 6). In these and other cases, the metabolic associations and linkage to disease status identify possible new therapeutic targets concordant with biomarkers for the clear establishment of efficacy and may inform patient stratification based on genetic profiling. Furthermore, the associated metabolites identify potential biomarkers that are readily detectable in accessible samples, for example, urine or plasma, and that may improve the evaluation of disease or efficacy of new medicines while accounting for individual genetic background. Finally, this catalog of associations provides new opportunities for drug repositioning or for the identification of new indications for existing drugs, thus potentially unlocking some of the original investment.

**DISCUSSION**

In summary, we carried out the most comprehensive analysis of genetic influences on human blood metabolites thus far. Our observations suggest widespread genetic control over a large range of different pathways and functions and support the notion of human metabolism as a complex continuum governed by genetic effects of variable intensity, complex regulatory influences and non-genetic effects. Our results advance knowledge in a number of areas of biomedical and pharmacogenetic interest, generating nearly 100 new hypotheses of SNP-metabolite and disease correlates and identifying a large catalog of new potential biomarkers as well as associations to drug targets, transporters and metabolic enzymes. Lastly, the network provides a comprehensive in vivo reference map of genetic influences on blood metabolites in a healthy human population sample, linking genetic variants to hundreds of biochemical relationships and pathways. The data and results generated by this study are made freely available through a web-based database and a downloadable network to facilitate the necessary functional and clinical exploration of these new hypotheses.

**METHODS**

Methods and any associated references are available in the online version of the paper.

**ACKNOWLEDGMENTS**

For TwinsUK, we thank the Genotyping Facilities at the Wellcome Trust Sanger Institute and the Center for Inherited Disease Research (CIDR)/US National Institutes of Health (NIH) for SNP genotyping. The KORA Study Group consists of A. Peters (speaker), J. Heinrich, R. Holle, R. Leidl, C. Meisinger, K. Strauch and their coworkers, who are responsible for the design and implementation of the KORA studies. For KORA, we thank P. Lichtner, G. Eckstein, G. Fischer, T. Strom, the Helmholtz Zentrum München genotyping staff and the field staff of the MONICA/KORA Augsburg studies. We also thank G. Fischer (KORA) and G. Sürdulescu (TwinsUK) for sample handling and H. Chavez (KORA) and D. Hodgkiss (TwinsUK) for sample shipment. We are grateful to the MuTHER investigators for the transcriptomic data. Finally, we wish to express our appreciation to all study participants of the TwinsUK and KORA studies for donating their blood and time.

Part of this work was funded by Pfizer Worldwide Research and Development. For TwinsUK, the study was funded by the Wellcome Trust European Community’s Seventh Framework Programme (FP7/2007-2013). The study also receives support from the National Institute for Health Research (NIHR) BioResource Clinical Research Facility and Biomedical Research Centre based at Guy’s and St Thomas’ National Health Service (NHS) Foundation Trust and King’s College London. T.D.S. is the holder of a European Research Council (ERC) Advanced Principal Investigator award. SNP genotyping was performed by the Wellcome Trust Sanger Institute and the National Eye Institute via NIH/CIDR. The KORA (Kooperative Gesundheitsforschung in der Region Augsburg) research
platform and the MONICA Augsburg studies were initiated and financed by the Helmholtz Zentrum München National Research Center for Environmental Health, which is funded by the German Federal Ministry of Education, Science, Research and Technology and by the state of Bavaria. This study was supported by a grant from the German Federal Ministry of Education and Research (BMBF) to the German Center for Diabetes Research (DDZ). The German National Genome Research Network financed part of this work (NGFNPlus 01GS0823). Computing resources have been made available by the Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities (HLRB project HLRB project HLRB project HLRB 2013). Part of this research was supported within the Munich Center of Health Sciences (MC Health) as part of LM-Unistativ. S. Y. S. is supported by a Post-Doctoral Research Fellowship from the Oak Foundation. F. J. T. is supported by an ERC starting grant (LatentCauses). J. K. is supported by the German Research Foundation (SPP 1395, InKoMBio) and by a grant from the German Helmholtz Postdoctoral Programme. K. S. is supported by Biomedical Research Program funds at Weill Cornell Medical College in Qatar, a program funded by the Qatar Foundation. C. G. is supported by the European Union’s Seventh Framework project MIMomics (FP7-Health-F5-2012-305280) and by the Russian Foundation for Basic Research (RFBR)-Helmholtz research program. N. S. is supported by the Wellcome Trust (grants WT098051 and WT091310) and by the European Commission (EUFP7 EPIGENESYS grant 257082 and BLUEPRINT grant HEALTH-F5-2011-282510). J. R. B. and V. F. are supported by the Canadian Institutes of Health Research, Fonds du Recherche du Science Québec and the Québec Consortium for Drug Discovery.

The Pfizer colleagues dedicate this manuscript to the memory of our friend Phoebe Roberts, whose passion for text mining, molecular biology and drug discovery contributed to the identification of causal genes in this research and to our work in general.

AUTHOR CONTRIBUTIONS
Study organization: M. J. B., K. S., G. K. and N. S. Manuscript preparation: S. Y. S., J. K., K. S., G. K. and N. S. Data collection: R. P. M. and M. V. M. Analysis of associations: S. Y. S., A. K.-P., J. H., G. K. and N. S. Locus bioinformatics annotation: E. B. F., D. Z., R. S., V. F., L. C., L. J. V., K. W., V. W., P. R., L. X., J. B. R., J. P. O. Manuscript preparation: M. J. B., K. S., G. K., E. B. F., M. W., C. G., J. T., I. E., A. M. V., C. L. H., T. P. Y., C. M., S. I. J. and E. G.

COMPETING FINANCIAL INTERESTS
The authors declare competing financial interests: details are available in the online version of the paper.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Garrod, A. E. Inborn Factors in Disease (Oxford University Press, Oxford, 1931).
2. Kettunen, J. et al. Genome-wide association study identifies multiple loci influencing human serum metabolite levels. Nat. Genet. 44, 269–276 (2012).
3. Sabatine, M. S. et al. Metabolic identification of novel biomarkers of myocardial ischemia. Circulation 112, 3868–3875 (2005).
4. Holmes, E. et al. Human metabolic phenotype diversity and its association with diet and blood pressure. Nature 453, 396–400 (2008).
5. Sreekumar, A. et al. Metabolicomic profiles delineate potential role for sarcosine in prostate cancer progression. Nature 457, 910–914 (2009).
6. Bichtash, M. et al. Opening up the “Black Box”: metabolic phenotyping and metabolome-wide association studies in epidemiology. J. Clin. Epidemiol. 63, 970–979 (2010).
7. Backshall, A., Sharma, R., Clarke, S. J. & Keun, H. C. Pharmacometabonomics profiling as a predictor of toxicity in patients with inoperable colorectal cancer treated with capecitabine. Clin. Cancer Res. 17, 3619–3628 (2011).
8. Wang, T. J. et al. Metabolite profiles and the risk of developing diabetes. Nat. Med. 17, 448–453 (2011).
9. Suhre, K. & Gieger, C. Genetic variation in metabolic phenotypes: study designs and applications. Nat. Rev. Genet. 13, 759–769 (2012).
10. Gieger, C. et al. Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. PLoS Genet. 4, e1000282 (2008).
11. Suhre, K. et al. Human metabolic individuality in biomedical and pharmaceutical research. Nature 477, 54–60 (2011).
12. Nicholson, G. et al. Human metabolic profiles are stably controlled by genetic and environmental variation. Mol. Syst. Biol. 7, 525 (2011).
13. Sansseau, P. et al. Use of genome-wide association studies for drug repositioning. Nat. Biotechnol. 30, 317–320 (2012).
14. Krumsieck, J. et al. Mining the unknown: a systems approach to metabolite identification combining genetic and metabolic information. PLoS Genet. 8, e1003005 (2012).
15. Lawlor, D. A., Harbord, R. M., Sterne, J. A., Timpson, N. & Davey Smith, G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. Stat. Med. 27, 1133–1163 (2008).
16. Kanemisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. 40, D109–D114 (2012).
17. Frazer, K. A. et al. A second generation human haplotype map of over 3.1 million SNPs. Nature 449, 851–861 (2007).
18. Pe‘er, I., Yelensky, R., Altshuler, D. & Daly, M. J. Estimation of the multiple testing burden for genewide association studies of nearly all common variants. Genet. Epidemiol. 32, 381–385 (2008).
19. Rhee, E. P. et al. A genome-wide association study of the human metabolome in a community-based cohort. Cell Metab. 18, 130–143 (2013).
20. Illig, T. et al. A genome-wide perspective of genetic variation in human metabolism. Nat. Genet. 42, 137–141 (2010).
21. Suhre, K. et al. A genome-wide association study of metabolic traits in human urine. Nat. Genet. 43, 565–569 (2011).
22. Nicholson, G. et al. A genome-wide metabolic QTL analysis in Europeans implicates two loci shaped by recent positive selection. PLoS Genet. 7, e1002270 (2011).
23. Shawrder, E. & Martinez-Carrion, M. Evidence of phenylalanine transaminase activity in the isoenzymes of asparagine transaminase. J. Biol. Chem. 242, 2486–2492 (1972).
24. Lee, H. D. et al. Caenorhabditis elegans mboat-7, a member of the MBOAT family, is required for selective incorporation of polyunsaturated fatty acids into phosphatidylinositol. Mol. Biol. Cell 19, 1174–1184 (2008).
25. Hu, C. A. et al. Overexpression of proline oxidase induces proline-dependent and mitochondrial-mediated apoptosis. Mol. Cell. Biochem. 295, 85–92 (2007).
26. Krumsieck, J., Suhre, K., Illig, T., Adams, J. & Theis, F. J. Gaussian graphical modeling reconstructor pathway reactions from high-throuput metabolomics data. BMC Syst. Biol. 5, 21 (2011).
27. Schadt, E. E. et al. Mapping the genetic architecture of gene expression in human liver. PLoS Biol. 6, e107 (2008).
28. Grundberg, E. et al. Mapping cis- and trans-regulatory effects across multiple tissues in twins. Nat. Genet. 44, 1084–1089 (2012).
29. Kim, D. K. et al. The human T-type amino acid transporter-1: characterization, gene organization, and chromosomal location. Genomics 79, 95–103 (2002).
30. Kaper, T. et al. Nanosensor detection of an immunoregulatory tryptophan influx/kyurenine efflux cycle. PLoS Biol. 5, e257 (2007).
31. Mohotha, V. K. & Hirschhorn, J. N. Inborn variation in metabolism. Nat. Genet. 42, 97–98 (2010).
32. Kottgen, A. et al. New loci associated with kidney function and chronic kidney disease. Nat. Genet. 42, 376–384 (2010).
33. Xie, W. et al. Genetic variants associated with glycine metabolism and their role in insulin sensitivity and type 2 diabetes. Diabetes 62, 2141–2150 (2013).
34. Kikuchi, G., Motokawa, Y., Yoshida, T. & Hiraga, K. Glycine cleavage system: reaction mechanism, physiological significance, and hyperglycinemia. Proc. Jpn. Acad., Ser. B, Phys. Biol. Sci. 84, 246–263 (2008).
ONLINE METHODS

Study samples. The TwinsUK cohort is an adult UK twin registry composed of mostly women aged 18 to 102 years. These twins were recruited from the general UK population through national media campaigns and were shown to have similar disease-related and lifestyle characteristics to population-based singletons in the same age group. The samples used in this study were 93% female in the age range of 17 to 85 years (mean of 53 years). The Cooperative Health Research in the Region of Augsburg (KORA) study is a series of independent population-based epidemiological surveys and follow-up studies of participants living in the region of Augsburg in southern Germany. The present study includes data from the follow-up study KORA F4 (2006–2008) of the KORA S4 survey (1999/2000). Individuals were aged between 32 and 77 years (mean of 61 years), and there were equal numbers of males and females (Supplementary Table 1). All participants in both TwinsUK and KORA have given written informed consent, and local ethics committees, the Guy's and St. Thomas' Hospital Ethics Committee for TwinsUK and Bayerische Landesärztekammer for KORA, approved the studies.

Primary genome-wide association analysis for metabolites. Primary association testing was carried out at each SNP (in the HapMap 2–based imputed genotype data set) for all 486 metabolite concentrations present in both the TwinsUK and KORA data sets after quality control steps. Linear regression models (assuming an additive genetic model) were used. Age and sex were included as covariates in both cohorts. Additionally, batch effect (see Supplementation) was included as a covariate in both cohorts. Furthermore, genome-wide association models were processed in a single batch. In TwinsUK, associations were carried out using Merlin software. Briefly, Merlin is based on the variance-component regression model and provides two family-based association tests under multivariate normality assumptions: a likelihood ratio test and a score test. In contrast to recently developed linear mixed model–based programs that estimate individual relationships from genotypes, Merlin takes pedigree information from direct input of family pedigrees and reported twin status (monozygotic or dizygotic). Association tests were run on best-guess genotypes (where genotypes were called if they had a posterior probability of >0.9) using the computationally more efficient score test. In KORA, linear models were fit for unrelated individuals using QUICKTEST software. Briefly, QUICKTEST is based on maximum-likelihood estimation and assesses association tests for unrelated individuals under a linear model framework with the assumption of a normal mixture model for error distribution. QUICKTEST can handle uncertain genotypes (for example, imputed genotypes with uncertain scores) and non-normal traits (for example, traits with skewness and heavy tails). Association tests were carried out on allelic dosages (0–2), accounting for uncertainty in imputed genotypes. For the 145 most associated SNP–metabolite pairs, associations were recalculated in TwinsUK using allelic dosage to ensure that association results were not affected by genotype modeling. The resulting association P values were virtually identical to those obtained from Merlin ($R^2 = 0.99$), demonstrating that there was no bias associated with the use of best-guess genotypes in TwinsUK and allelic dosages in KORA.

Meta-analysis. Association results in TwinsUK and KORA were combined using inverse variance meta-analysis based on effect size estimates and standard errors, adjusting for genomic control. Heterogeneity in each association between data sets was tested using Cochran's Q test (which is equivalent to the Mc Nemar test here as the number of data sets is two). All above analyses were carried out using Metal software. To control for false positive error rates resulting from the large number of SNPs tested, a conservative Bonferroni-adjusted P value of $1.03 \times 10^{-12}$ ($5 \times 10^{-9}/486$) was applied in declaring genome-wide significance for the SNP–metabolite associations. SNPs with low imputation quality (info < 0.4), low minor allele frequency ($<0.01$) or significant heterogeneity of effects between the two cohorts (defined by heterogeneity $P < 0.001$ and a $P$ value in either cohort of $>0.001 = 0.05/486$ and SNPs present in only one cohort were removed after meta-analysis. A lead SNP at each locus was selected as the SNP with the lowest $P$ value for association with any trait at that locus. For each metabolite, all associations passing this cutoff were assigned to independent loci by iteratively allocating the lead SNP with the lowest $P$ value and SNPs within 500 kb of it to the same locus. These assignments were further revised as described below to estimate the effective number of independent loci.

Genome-wide analysis of metabolite/metabolite ratios. A second discovery step was carried out by testing genome-wide associations on all pairwise ratios of metabolite concentrations present in both cohorts after quality control steps, following the principle described previously. Xenobiotic metabolites were excluded owing to the low average call rates per individual. Because of the large high of traits and computation time and costs, for metabolite ratios, genome-wide analyses were carried out in the TwinsUK study alone, followed by replication of significant hits (Bonferroni-corrected $P < 5.08 \times 10^{-13} = 5 \times 10^{-9}/98,346$) in KORA. SNPs with low minor allele frequency ($<1$%) and info (<0.4) were removed from analysis, as in the association analysis for single metabolites. In the analysis in TwinsUK, a total of 430 loci survived Bonferroni correction, and only the top association for each locus (the metabolite ratio and the SNP pair with the lowest $P$ value) was carried forward for replication in KORA. Meta-analysis was performed on both discovery and replication results using the inverse variance model, and the combined result was filtered again by Bonferroni-adjusted $P$ value < $5.08 \times 10^{-13}$ and heterogeneity $P \geq 0.001$. Further details on the interpretation and reporting of ratios are provided in the Supplementary Note.

Refinement of independence at metabolite loci. A series of additional analyses were carried out to further refine the identity of single metabolite– and metabolite ratio–associated loci. Regional association plots were visually inspected to identify possible independent signals within each locus or to detect possible signals spanning two consecutive loci. For loci where putative independent SNPs were identified (defined as two SNPs having $r^2 \leq 0.5$ in the 1000 Genomes Project CEU population), conditional analyses were used to assess the statistical independence of the two loci. Specifically, we reciprocally tested association with each SNP while adding the second SNP as a term in the linear model. Bonferroni correction for the number of pairs tested was applied to declare significance for the conditional test $P$ value. In the case of non-independence, the best SNP–metabolite pair with the lowest association $P$ value was reported at each locus.

Total heritability estimates. Genetic and environmental influences on the concentrations of 503 single metabolites present in TwinsUK after quality control steps were inferred under the ACE model (which models trait variance as a function of additive genetics, common environment and unique environment and/or error effects). Narrow-sense heritabilities were inferred from the proportion of the total variance explained by estimated additive genetic effect. To estimate parameters of the ACE model, a maximum-likelihood method was applied under multivariate normality assumptions using OpenMx software, adjusting for age, sex and batch effect (Supplementary Table 7). Metabolites measured in fewer than 30 twin pairs (either monozygotic or dizygotic) were excluded from calculations for heritability estimation.

Known heritability estimates under additive models. Next, known heritability was estimated for 170 single metabolites associated with genetic variants in our GWAS. The explained genetic variance of each metabolite was estimated by multiple regression analysis including all associated SNPs under additive genetic models, after adjusting for covariates (age and sex in KORA; age, sex and experimental batch effect in TwinsUK). Only unrelated twin individuals were used for the analysis to avoid biases due to familial correlation. Known heritability was defined as the ratio of total variance in the metabolite to the variance explained by the multiple regression models including all SNPs significantly associated with the metabolite. This known heritability was then compared with the total heritability inferred above (Supplementary Table 7).

Imputation using denser haplotype maps from the 1000 Genomes Project. Genetic variants associated with metabolite concentrations below the genome-wide significance cutoff might contain true signals contributing to metabolite heritability. To more comprehensively survey genetic variation at the 145 metabolic loci, associations at each locus were recalculated following imputation using a denser reference set (1000 Genomes Project).

© 2014 Nature America, Inc. All rights reserved.
Genome-wide genotypes were imputed using the 1000 Genomes Project multi-population panel (March 2012 release for TwinsUK; June 2011 release for KORA), and associations with metabolite concentrations were tested as described within a 1-Mb window centered on each lead SNP. Analyses were carried out in TwinsUK and KORA separately, and the results were combined by inverse variance meta-analysis, following the procedure used in the HapMap 2-based GWAS. Minor allele frequency, association P value and the variance explained by the two models ($\gamma = \alpha + \beta_1 \times \text{SNP}_1 + \beta_2 \times \text{SNP}_2 + \beta \times \text{SNP}_{12} + \epsilon$) were then compared for the most strongly associated SNPs identified in HapMap 2- and 1000 Genomes Project-based imputation (respectively, $\text{SNP}_{12}$ and $\text{SNP}_{x12}$). Finally, the explained genetic variance of each metabolite was estimated by multiple regression analysis including all associated $\text{SNP}_{12}$ markers under an additive genetic model and compared with total heritability and with known heritability estimated by multiple regression analysis including all originally associated $\text{SNP}_{12}$ markers under an additive genetic model. Overall, at 80% of the loci, 1000 Genomes Project–based imputation yielded a different variant than the HapMap 2 imputation used for discovery (Supplementary Table 8). However, the 1000 Genomes Project–based lead SNPs had similar minor allele frequencies ($R^2 = 0.85$), association P values ($R^2 = 0.99$) and explained variance ($R^2 = 0.96$) to the HapMap 2-based SNPs (Supplementary Fig. 3a–c), with only one locus (CTP3A4-CTP3A5) showing a significant improvement in association in the 1000 Genomes Project data set (Supplementary Fig. 3d,e).

Epistatic interactions in known heritability estimates. Frequent statistical interaction (epistasis) between genetic variants has been proposed to inflate heritability estimates by creating so-called phantom heritability. Although exhaustively testing interactions between SNPs across the genome for all traits would be statistically intractable, it is nevertheless of interest to explore possible epistatic effects between genome-wide significant SNPs, particularly given their co-occurrence within tight metabolic networks. To test for epistasis, we focused on all pairs of SNPs associated with the same metabolite at genome-wide significance (106 pairs of SNPs associated with 51 metabolites; Supplementary Table 9). SNP-SNP interaction was defined as a departure from additive marginal effects and was verified by an ANOVA F test comparing the following two models with and without an interaction term, $\gamma \times \text{SNP}_1 \times \text{SNP}_2$:

$$
Y = \alpha + \beta_1 \times \text{SNP}_1 + \beta_2 \times \text{SNP}_2 + \epsilon
$$

$$
Y = \alpha + \beta_1 \times \text{SNP}_1 + \beta_2 \times \text{SNP}_2 + \gamma \times \text{SNP}_1 \times \text{SNP}_2 + \epsilon
$$

The same co-variates used in the primary discovery effort (age and sex in KORA; age, sex and batch effect in TwinsUK) were included in each model. For simplicity, in TwinsUK, only unrelated individuals were included in the analysis. A Bonferroni-corrected P value of 0.00047 ($\approx 0.05/106$) was applied to declare significant epistasis. These tests identified no evidence for interaction among the associated loci ($P \leq 4.7 \times 10^{-10}$) in both 1000 Genomes Project pilot phase (CEU population). Each lead SNP and its proxies were then used as baits to search the MuTHER Project expression database and a published liver eQTL data set. All significant cis-eQTLs within a 1-Mb window centered on the lead SNP were retrieved from these data sets, and the best eQTL P value in each tissue was noted. A total of 57 lead SNPs identified cis-eQTLs in at least 1 of the 3 tissues searched under nominal or permutation $P < 0.001$ (Supplementary Table 10). For additional information, see the Supplementary Note.

Mendelian randomization analysis on causal metabolite pathways. Relative normalized gene expression values for all eQTLs identified were retrieved for 484 unrelated TwinsUK participants with available genotype data and with gene expression and metabolite data obtained at the same time point by the MuTHER Project in 3 tissues (fat, skin and LCLs). Mendelian randomization analysis was used to test the hypothesis that changes in gene expression levels would cause changes in metabolite levels, using the genotype of the lead SNP as an instrumental variable. The causal effect of gene expression levels on metabolite levels was estimated by the Wald ratio method in the ratio of SNP effects (instrumental variable effects) on gene expression levels (GE) and metabolite levels (MET) as follows:

$$
\hat{\beta} = \hat{\beta}_{\text{SNP-MET}} / \hat{\beta}_{\text{SNP-GE}}
$$

where $\hat{\beta}_{\text{SNP-MET}}$ is the coefficient for the linear regression model of metabolite levels on SNPs and $\hat{\beta}_{\text{SNP-GE}}$ is the coefficient for the linear regression model of gene expression levels on SNPs. Covariate adjustments (age, sex and batch effect for metabolite levels; age, sex and probe batch for gene expression levels) were made before analysis. To test the causal effect of 32 lead SNPs from the loci where annotated causal genes overlapped with MuTHER eQTLs, a Bonferroni-corrected 99.85% confidence interval ($\approx 0.05/32$) for the causal effect of gene expression levels on metabolite levels was obtained from 10,000 permutations, and the null hypothesis of no causation was rejected if the confidence interval did not cross zero (Supplementary Table 11).

Drug associations. A list of drugs approved by the FDA and/or the EMA was retrieved from the ChEMBL database. A total of 132 unique genes reported in...
Supplementary Table 4, including either predicted causal genes (94 unique genes) or genes nearest to the associated SNPs, were considered. Genes were classified as drug efficacy targets, metabolizing enzymes and/or transporters according to their role in clinical pharmacology. For genes classified as drug targets, therapeutic annotation was extracted from the current version of the approved prescribing information, the summary of product characteristics (SPC) and the primary literature. For genes classified as metabolizing enzymes and transporters, annotation was compiled from the information present in reference sources. The mapping of UniProt IDs to gene loci was retrieved from Ensembl BioMart (release 70), and the mapping of UniProt IDs to ChEMBL IDs from ChEMBL14.

Further mapping of the 132 genes to drugs was also conducted using the Citeline Pharmaprojects Pipeline (see URLs; accessed 1 July 2013) to annotate targets for drugs in other stages of development. For each metabolic locus, the pipeline was searched using the gene symbol for the metabolic locus as the search criterion in the “biological target” field. Information was gathered from the search results, including the drug’s name and the global status of its development. Entries for drug targets corresponding to launched drugs were checked for consistency against the previous set of FDA- and/or EMA-approved drugs, and information was retained from the initial set. Only information for drug targets corresponding to the following categories was retained from the latter data set (excluding ‘launched’): (i) in development (preclinical, phase I, phase II, phase III, preregistration, registration) and (ii) no development (suspended, withdrawn, no development reported).

35. Moayyeri, A., Hammond, C.J., Hart, D.J. & Spector, T.D. The UK adult twin registry (TwinsUK Resource). Twin Res. Hum. Genet. 16, 144–149 (2013).
36. Wichmann, H.E., Gieger, C. & Illig, T. KORA-gen—resource for population genetics, controls and a broad spectrum of disease phenotypes. Gesundheitswesen 67 (suppl. 1), S26–S30 (2005).
37. Abecasis, G.R., Cherny, S.S., Cookson, W.O. & Cardon, L.R. Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. Nat. Genet. 30, 97–101 (2002).
38. Kutalik, Z., Whittaker, J., Waterworth, D., Beckmann, J.S. & Bergmann, S. Novel method to estimate the phenotypic variation explained by genome-wide association studies reveals large fraction of the missing heritability. Genet. Epidemiol. 35, 341–349 (2011).
39. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics 26, 2190–2191 (2010).
40. Boher, S. et al. OpenMx: an open source extended structural equation modeling framework. Psychometrika 76, 306–317 (2011).
41. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature 467, 1061–1073 (2010).
42. Zuk, O., Hechtter, E., Sunyaev, S.R. & Lander, E.S. The mystery of missing heritability: genetic interactions create phantom heritability. Proc. Natl. Acad. Sci. USA 109, 1193–1198 (2012).
43. Hao, T., Ma, H.W., Zhao, X.M. & Goryanin, I. Compartmentalization of the Edinburgh human metabolic network. BMC Bioinformatics 11, 393 (2010).
44. Schomburg, I. et al. BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA. Nucleic Acids Res. 41, D764–D772 (2013).
45. Smith, G.D. & Ebrahim, S. ‘Mendelian randomization’: can genetic epidemiology contribute to understanding environmental determinants of disease? Int. J. Epidemiol. 32, 1–22 (2003).