STUDY DESIGNS

Genetic variation in metabolic phenotypes: study designs and applications

Karsten Suhre[1,2] and Christian Gieger[3]

Abstract | Many complex disorders are linked to metabolic phenotypes. Revealing genetic influences on metabolic phenotypes is key to a systems-wide understanding of their interactions with environmental and lifestyle factors in their aetiology, and we can now explore the genetics of large panels of metabolic traits by coupling genome-wide association studies and metabolomics. These genome-wide association studies are beginning to unravel the genetic contribution to human metabolic individuality and to demonstrate its relevance for biomedical and pharmaceutical research. Adopting the most appropriate study designs and analytical tools is paramount to further refining the genotype–phenotype map and eventually identifying the part played by genetic influences on metabolic phenotypes. We discuss such design considerations and applications in this Review.

NMR spectroscopy
An experimental technique that identifies molecules by the specific pattern in the chemical shift of specific atoms.

[1]*Department of Physiology and Biophysics, Weill Cornell Medical College in Qatar, Education City, Qatar Foundation, P.O. BOX 24144, Doha, Qatar.*
[2]*Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München, German Research Center for Environmental Health, Ingolstädter Landstraße 1, 85764 Neuherberg, Germany.*
[3]*Institute of Genetic Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Ingolstädter Landstraße 1, 85764 Neuherberg, Germany.*
e-mails: karsten@suhre.fr; christian.gieger@helmholtz-muenchen.de
doi:10.1038/nrg3314
Published online
3 October 2012

More than 100 years ago, Archibald Garrod realized that "inborn errors of metabolism" are "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"[1–3]. Disruptions in metabolic processes are associated with many common diseases, such as type 2 diabetes and cardiovascular disorders. Some disease-associated changes in metabolic phenotypes are causative and therefore constitute potential points of therapeutic intervention; other changes in metabolite levels are a consequence of the disease and thereby represent possible prognostic or diagnostic biomarkers of disease. Successful diagnosis, therapy and prevention of complex disorders thus requires a systems-wide understanding of the interactions between genetic, environmental and lifestyle factors in the resulting metabolic phenotype.

Modern bioanalytical techniques that have been built on recent advances in NMR spectroscopy, mass spectrometry and high-performance liquid-phase chromatography (HPLC) can now provide quantitative readouts for hundreds of small molecules that are detected in large sets of biological samples obtained from epidemiological population studies. At present, more than 4,200 compounds have been annotated in human metabolite databases[4]. Such a wide-ranging metabolic characterization of biological samples generates a wealth of phenotypic data that has never been accessible before and has given birth to the emerging field of metabolomics. Wide-ranging metabolic phenotypes can be analysed in association with genetic variance, disease-relevant phenotypes and lifestyle and environmental parameters, allowing dissection of the relative influences of these factors.

Now, genome-wide association studies (GWASs) can be carried out with broad panels of metabolite concentrations (TABLE 1). Using this largely hypothesis-free approach, common genetic variants in genes encoding enzymes and transporter proteins have been identified that can have substantial influences on human metabolic traits. These so-called genetically influenced metabotypes (GIMs) are starting to be combined with the increasing knowledge of disease-associated genetic loci to uncover new complex risk factors of common diseases and to provide functional insights into the pathophysiology of related disorders. Knowledge of the genetic basis of human metabolic individuality is a key ingredient of emerging gene-based personalized therapies, including pharmacogenomics[5] and nutrigenomics[6,7]. BOX 1 and TABLE 2 present emerging insights from the study strategy.

To gain the most from this emerging approach, it is necessary to use appropriate study designs and analytical tools. In this Review, we describe the metabolic phenotype, the experimental methods that are available

Table 1 | **Published genome-wide association studies with large panels of metabolic traits**

Sample type	Platform used	Metabolic panel	Number of study participants and study description	Number of traits	Number of reported loci	Refs
Serum	Targeted LC–MS/MS and FIA–MS/MS	Lipids, carbohydrates and amino acids	284 (KORA* study, only males, age >55 years)	363	4	21
Plasma	Trans-esterification of lipids followed by gas chromatography	Omega 3 and omega 6 fatty acids	1,075 (InCHIANTI‡ study); 1,076 (GOLDN§ study)	6	2	22
Plasma and serum	Targeted LC–MS/MS	Sphingolipids	4,400 (five European populations)	33	5	23
Serum	Targeted FIA–MS/MS	Lipidomics-oriented panel	1,809 (KORA study); 422 (TwinsUK‖ study)	163	9	16
Urine	Targeted ^{1}H NMR	Manual annotation against a library of known chemical shifts	862 (SHIP¶ study, males); 1,039 (SHIP study, females); 992 (KORA study)	59	5	14
Plasma and urine	Non-targeted ^{1}H NMR and targeted FIA–MS/MS	Unidentified chemical shifts (NMR); lipidomics-oriented panel (FIA–MS/MS)	142 (MolTWIN# study); 62 (MolOBB# study)	526 peaks (NMR); 163 (mass spectrometry)	3 (NMR, urine); 1 (NMR, plasma); 2 (mass spectrometry, plasma)	13
Serum	Non-targeted LC–MS/MS and GC–MS	Metabolome-wide coverage of 60 metabolic pathways	1,768 (KORA study); 1,052 (TwinsUK study)	276	37	15
Serum	NMR with automated metabolite annotation	Mainly serum lipid extracts, amino acids and some other metabolites	8,330 (Finnish population)	216	31	10
Plasma	Targeted LC–MS/MS	Phospholipids and sphingolipids	4,034 (five European populations)	115 phospholipids; 33 sphingolipids	25 (phospholipids); 10 (sphingolipids)	24

The 'Number of reported loci' corresponds to those associations that meet the individual studies' criteria of genome-wide significance. *Kooperative Gesundheitsforschung in der Region Augsburg, Germany. ‡A population-based epidemiological study in the older population living in the Chianti region of Tuscany, Italy. §Genetics of Lipid Lowering Drugs and Diet Network (United States). ‖An adult twin registry in the United Kingdom. ¶Study of Health in Pomerania, Germany. #Two cohorts from the MolPAGE programme in the United Kingdom. FIA–MS/MS, flow injection analysis coupled with tandem mass spectrometry; GC–MS, gas chromatography coupled with mass spectrometry; LC–MS/MS, liquid chromatography coupled with tandem mass spectrometry.

for high-throughput metabolic phenotyping and their application to larger human population studies. We then show how recently found genetic variants with metabolic traits provide new insights into the aetiology of complex diseases. We focus on the design considerations that need to be kept in mind in future studies.

What is a metabolic phenotype?

Evidence that the metabolome is at least in part genetic. The metabolic phenotype (or metabotype) of an individual can be viewed as the ensemble state of the concentrations of all endogenous small molecules (metabolic traits) in all body organs and bodily fluids. In relation to a disease, a metabolic trait may be a functional intermediate trait or merely a correlated biomarker. In contrast to the genotype of an individual, which remains almost identical over their lifespan, the metabotype substantially varies with time and is influenced by a wide range of environmental and lifestyle factors, including fasting and feeding states, time of day and menstrual cycle. A study that applied a wide range of physiological challenges to participants demonstrated that challenges increase the variability of certain metabolite profiles among volunteers with similar characteristics. Discrete metabotypes could thereby be identified that would not have been distinguishable under normal fasting conditions[8]. Thus, every metabolomic characterization of a biosample represents a snapshot of a part of that individual's present metabolic state at that particular time.

Therefore, one may ask whether the concept of an individual metabolic phenotype is actually meaningful in the context of population-based studies. To assess this, it is useful to estimate how much of the population level variance is driven by genetic factors and how much is driven by environmental factors. A longitudinal study of plasma and urine samples from identical and non-identical twin pairs showed that the human metabolome is controlled by both genetic and environmental factors[9]. An analysis of Finnish twin pairs also found high heritability for certain metabolic phenotypes, measured on a different metabolomics platform[10]. What is important to note here is that every metabolite has specific properties: most of them are very sensitive to environmental influences, and their concentrations may vary over timescales of minutes, hours or days. Nevertheless, their biochemical processing is controlled by enzymes and transporters, and thus they are influenced by the genetic variation that affects the expression or function of these proteins.

Intermediate phenotypes. GWASs have identified many risk loci for complex disorders. The number of associations is increasing as more highly powered GWASs and meta-analyses are conducted. However, the effect sizes of genetic associations with complex disorders are

High-performance liquid-phase chromatography
(HPLC). A chromatographic technique used to separate a complex mixture of metabolites. Often used in combination with mass spectrometry.

Metabolomics
The field of identifying metabolites in a biological sample using techniques such as NMR spectrometry and liquid- or gas-phase chromatography coupled with mass spectroscopy. 'Metabonomics' is often synonymously used in connection with NMR-based experiments.

Metabolic traits
Quantitative measures of the concentrations of a specific metabolite.

Genetically influenced metabotypes
(GIMs). Associations between a genetic variant and a metabolic phenotype.

Box 1 | **Emerging insights from GWASs with metabolomics**

Several patterns are beginning to emerge from genome-wide association studies (GWASs) with different metabolic panels, experimental methods and sample types. See TABLE 1 for a summary of studies carried out so far and TABLE 2 for some examples of findings with biomedical relevance. The points summarized here demonstrate the use and potential of this experimental approach.

High allele frequencies and large effect sizes

The identified genetic variants are often frequent (>20%) and have exceptionally high effect sizes, explaining 10–20% of the observed variance[15]. These genetically influenced metabotypes (GIMs) do not result in the full loss-of-function of metabolism-related proteins (such as inborn errors of metabolism, in which metabolite concentrations can reach toxic levels in homozygous individuals) but still lead to substantial modifications in their efficiency. In most cases, the genetic variant is found in a gene that codes for an enzyme, a transporter or some other kind of metabolism-related gene.

Overlap with disease associations

Many disease end points most probably induce, or are induced by, a metabolic phenotype, which can be picked up in a GWAS with metabolomics (FIG. 1). For example, the *N*-acetyltransferase 8 (*NAT8*) locus encodes an *N*-acetylase protein and is a known risk locus for chronic kidney disease[36]. A study uncovered an association of the *NAT8* locus with serum levels of *N*-acetylornithine[15] and also showed that *N*-acetylornithine associates with estimated glomerular filtration rate (eGFR), providing new insights into the aetiology of chronic kidney disease.

Links to pharmacoogenomics

Similarly, many GIMs are associated with response to drug treatment. For example, the solute carrier organic anion transporter family, member 1B1 (*SLCO1B1*) locus associates with risk of statin-induced myopathy[5]. In a recent GWAS with metabolomics, it was found to associate with a series of fatty acids, including tetradecanedioate and hexadecanedioate[15]. This information can potentially be used to support the redesign of the respective drugs, for instance by using tetradecanedioate and hexadecanedioate as functional readouts in biochemical assays of drug side effect[37].

Replication of GWASs with individual metabolic traits

Associations from previous GWASs with clinically relevant traits, such as serum fasting glucose[38], bilirubin[39,40], urate[41] and dehydroisoandrosterone sulfate[42] levels can be replicated in a single GWAS with a large panel of metabolic traits[15].

Refinement of associations with bulk traits

Metabolic traits can also provide a more detailed representation of a 'bulk' trait. For instance, lipase, hepatic (*LIPC*) associates with blood triglyceride levels[43], which are bulk measures of a complex mixture of lipid traits. In a GWAS with metabolic traits, this locus was found to associate with a number of glycerophosphatidylethanolamines[21], therefore refining the association with the metabolic trait, which may ultimately be of clinical benefit.

Identification of true positives in GWASs with clinically relevant end points

A combination of a GWAS with metabolomics and data from previous GWASs can identify promising new candidate SNPs and provide new insights into the functional background of these associations. For example, although two early GWASs[44,45] reported an association of SNP rs174548 (near *FADS1*) with serum low-density lipoprotein (LDL) cholesterol, high-density lipoprotein (HDL) cholesterol and total cholesterol levels, these associations were not considered as potential candidates for replication in those studies. Gieger *et al.*[21] later identified an association of rs174548 with a number of glycerophospholipids in a GWAS with metabolomics, and on the basis of the fact that glycerophospholipids are major constituents of LDL and HDL particles, they argued that *FADS1* may be a true positive-lipid-associated locus, a prediction that was later confirmed in a large study with over 40,000 individuals[46].

generally small, and information on the underlying biological processes is often lacking. Therefore, the focus of GWASs is now shifting increasingly away from studying associations with disease end points and towards studying associations with intermediate traits that are known risk factors of disease (FIG. 1). Examples include GWASs for: blood triglyceride, cholesterol and bilirubin levels, which are risk factors for cardiovascular disease; fasting glucose levels and glucose levels after an oral glucose tolerance test, which are linked to diabetes; urate levels, which are linked to gout; and liver enzymes, which are indicators of liver disease. These studies have shown that genetic association with quantitative traits that are functional intermediates of complex disorders are often more highly powered, and furthermore they can provide information on the biological underpinning of the disease association.

However, by studying only known risk factors of disease, it is unlikely that any new biological processes or pathways will be discovered that may be involved or disrupted in the aetiology of the disease. Because metabolic phenotypes are important readouts of many biological processes, a largely hypothesis-free approach of GWASs with large panels of metabolic traits (metabolomics) may be used to respond to this challenge. One hundred and fifty years of biochemical research have created a wealth of knowledge on the biological properties of most metabolites and also on the pathways that link these metabolites in healthy or diseased individuals. The metabolic trait in a GWAS thus has the role of an intermediate phenotype that functionally links genetic variation to disease-predisposing factors and then to complex disease end points. The examples shown in BOX 1 and TABLE 2 demonstrate this potential.

Study design

Choice of metabolomics platform. FIGURE 2 presents the main steps of a high-throughput metabolomics experiment together with design considerations for genetics-oriented metabolomics studies. Robust and high-throughput measurement capabilities are required to carry out GWASs with metabolomics. The technologies that are most often used in metabolomics experiments are based on either mass spectrometry or NMR spectroscopy. Mass-spectrometry-based methods characterize a metabolite by its molecular mass, its specific fractionation pattern (tandem mass spectrometry) and its retention time when liquid-phase or gas-phase chromatography separation is used. The most widely implemented NMR-based method in metabolomics is ^{1}H NMR. A small molecule is identified here by a specific pattern (called the chemical shift) in the resonance spectrum of its protons when excited by an oscillating magnetic field.

The initial 'raw' quantitative readout of a metabolic feature is a specific pattern of peaks in a mass spectrum or an NMR spectrum and related information, such as the elution time, when using a chromatography method. Ascertaining the biochemical identity of the metabolites that are represented by these raw data is sometimes an issue. Comparison with reference spectra that are obtained from pure substances or spiking experiments can provide such information. Nevertheless, many of the experimentally observed metabolites (or metabolic features) are currently not biochemically identified. We thus distinguish between peak-based (or feature-based) metabolomics and metabolomics that uses annotated metabolite concentrations of known (and possibly also unknown) biochemical identity.

Table 2 | **Selected loci of genetic association with metabolic traits that also associate with end points of biomedical relevance**

Gene locus and SNP	Associated metabolic trait	*P* value	Function of gene product	Disease association
GCKR; rs780094	Glucose/mannose ratio	5.5×10^{-53}	Regulates glucokinase in liver and pancreatic islet cells	Type 2 diabetes related traits[38]; Crohn's disease[49]
ENPEP; rs2087160	Amino-terminal-cleaved fibrinogen A-alpha peptide levels	6.5×10^{-13}	Functions in the catabolic pathway of the renin–angiotensin system and regulates blood pressure	Blood pressure[50]
NAT2; rs1495743	1-methylxanthine/4-acetamidobutanoate ratio	1.7×10^{-40}	Participates in the detoxification of hydrazine and arylamine drugs	Coronary artery disease[43]; bladder cancer[51]
NAT8; rs13391552	*N*-acetylornithine levels	5.4×10^{-252}	*N*-acetyltransferase; *N*-acetylornithine associates with eGFR[15]	Chronic kidney disease[52]
SLC2A9; rs4481233	Urate levels	5.5×10^{-34}	Urate transporter	Gout[53–55]

In most cases, more than one metabolic trait associates with a genetic locus. Full association data are available from the GWAS-server. *P* values are taken from REF. 15. Overlaps with disease associations are reported when the lead SNPs are in high linkage disequilibrium with the metabolite-associated SNPs ($R^2 > 0.8$). eGFR, estimated glomerular filtration rate; *ENPEP*, glutamyl aminopeptidase (aminopeptidase A); *GCKR*, glucokinase (hexokinase 4) regulator; *NAT2*, *N*-acetyltransferase 2; *SLC2A9*, solute carrier family 2 (facilitated glucose transporter), member 9.

The most notable advantage of mass-spectrometry-based methods compared with NMR methods is their higher sensitivity. However, this advantage comes at the cost of more complex demands in terms of sample preparation and in carrying out the actual measurement, thereby providing many potential sources for experimental errors and uncontrolled-for variances in the resulting data sets. NMR-based measurements, however, do not require the extraction of metabolites and leave the samples intact for further analysis. Also, absolute quantification with mass-spectrometry-based methods requires external reference standards for most of the measured metabolites, whereas NMR-based methods provide quantification with one or two references. Furthermore, the reproducibility of NMR experiments is excellent, whereas batch effects are often observed when mass spectrometry experiments are conducted at different times. Both methods thus have their strengths and weaknesses. If resources permit, a combination of both would be optimal.

The measurement set-ups of these platforms are complex and can rarely be fully replicated by any independent laboratory. In a pilot study that determined 423 unique metabolite concentrations in blood samples from identical study participants using three different commercial platforms, 50 metabolites were quantified on more than one platform. The median correlation coefficient, *R*, between the platforms was 0.61. In three cases, no correlation was found, indicating that the different techniques may actually be measuring different metabolites in these cases. For other metabolites, a very strong correlation (up to $R = 0.95$) was observed[11]. Even if described in great detail, subtle differences in machine set-up and sample processing may have a great impact on certain metabolic readouts. It is therefore essential to compare and to harmonize measurements taken from identical samples across platforms and to ensure that the final metabolomics readouts are within a well-defined range of experimental error. When investigating the same genetic association using identical samples on different platforms, the differences in the strength of the resulting association signals are solely dependent on the experimental errors incurred by these platforms. Therefore, the data from the platform that displays the strongest association to the genetic variant is likely to be the most accurate.

At this point, the choice of the metabolomics provider should be considered: relying on in-house methods has the advantage of providing full control over the measurements, but this comes with the requirement of having to build up and to maintain such a platform. Using a commercial provider is an alternative that can bring metabolomics experiments within the reach of groups that do not have access to local metabolomics core facilities. Potential drawbacks of this approach are the generally rather limited access to details of the implemented methods and also fewer options available for tweaking the experimental set-up during the measurement process. Intermediate options are the use of commercial metabolomics kit technologies or out-licensing of proprietary know-how and software protocols on local platforms.

Choosing which metabolites and tissues to study. Targeted methods study specific (known) metabolites and thereby provide more precise measurements and are easy to replicate but are limited to analysing only a subset of preselected compounds. Non-targeted metabolomics offers a wider and largely hypothesis-free approach but also increases the need to manage multiple testing during analysis (see below). Additionally, targeted metabolomics methods are able to provide absolute quantification by comparison to isotope-labelled external standards, whereas non-targeted methods often only provide semi-quantitative traits, such as ion counts per sampling time, which may vary extensively between experiments. This is, in principle, not a problem in GWASs, in which the experiment identifier can be added as a covariate to the statistical model to correct for such batch effects. However, it could limit the usability of the metabolomics data in other (non-genetic) studies.

In the choice of the metabolites to study, there is generally a trade-off to be made between a wide and largely non-targeted panel, which often comes at the

Metabolic individuality
The metabolic capacities of an individual, as defined by the ensemble of all functional genetic variants (genetically influenced metabotypes) in their metabolism-related genes. Historically, Garrod introduced the term 'chemical individuality' to represent this concept.

Metabolome
The ensemble of all small molecules (metabolites) that are processed by the body's enzyme and transporter proteins.

Glycerophosphatidylethanolamines
Glycerophospholipids with ethanolamine head groups.

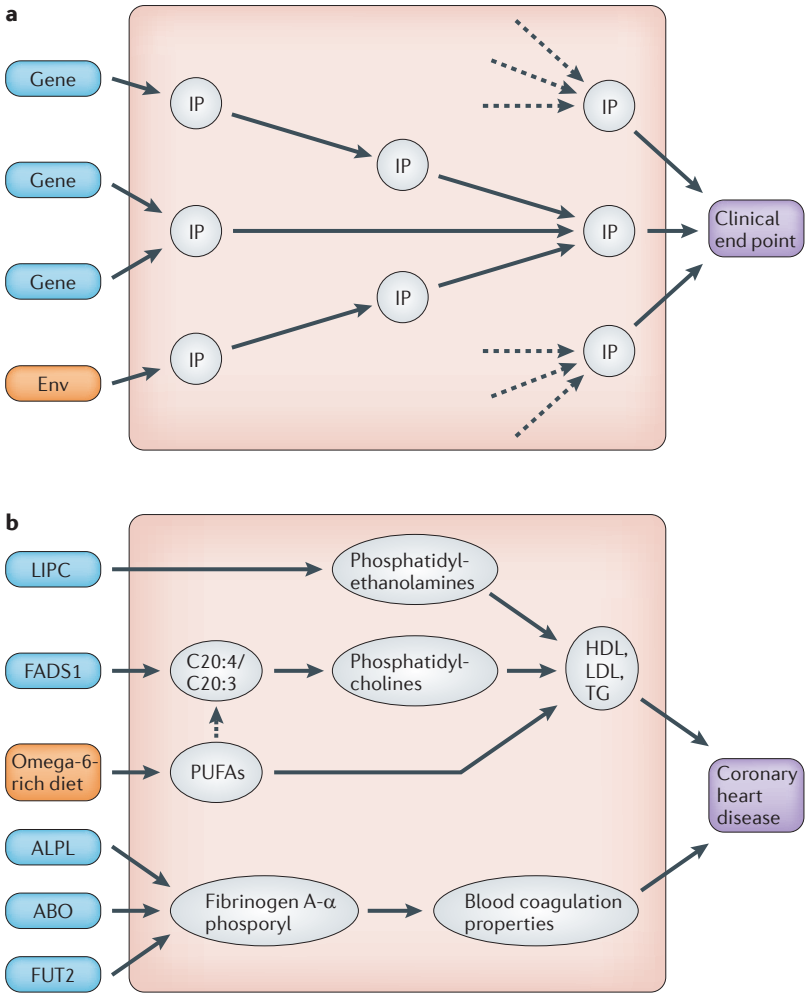

Figure 1 | **The metabolic trait as an intermediate phenotype.** The general concept (**a**) and an example using information from actual genome-wide association studies (GWASs) with metabolic traits[15,21] (**b**). The association of a genetic variant is strongest with its closest intermediate phenotype (IP; for example, the association of fatty acid desaturase 1 (FADS1) with its product–substrate pair; also see BOX 2 and REF. 47), although the association with the clinical end point may not even be detectable at a level of genome-wide significance ($P = 0.021$ for *FADS1* with coronary heart disease[48]). The ensemble of all genetic associations with metabolic traits defines our metabolic individuality and thereby our predisposition to disease[3]. ABO, ABO blood group; ALPL, alkaline phosphatase, liver/bone/kidney; C20:3, dihomolinolenate; C20:4, arachidonic acid; Env, environmental factor; FUT1 fucosyltransferase 1; HDL, high-density lipoprotein; LDL, low-density lipoprotein; LIPC, lipase, hepatic; PUFA, polyunsaturated fatty acid; TG, thyroglobulin.

cost of lower data quality, and a narrower targeted panel, which comes at the cost of missing potentially interesting metabolites. The decision of which method to use should factor in how much and which additional phenotype information is available on the individual samples and whether this information can be enriched by a specific targeted metabolomics panel. Bearing in mind that no single technique allows the measurement of all metabolites in one go, a non-targeted approach is currently more promising as it may allow the discovery of new associations with hitherto uncharacterized metabolites.

Q–Q plots
A graphical method for comparing probability distributions. In genome-wide association studies, it is used to verify whether the *P* values are normally distributed; an over-representation of low *P* values indicates possible true positive associations.

Study population and size. Most GWASs with metabolomics have so far been conducted in the general population, with participants mostly of Caucasian origin. It is therefore likely that many genetic effects that are specific to different ethnicities have not yet been discovered, calling for extended studies in other populations. It should be noted that some genetic variants in metabolism-related genes depict sexual dimorphisms[12] and need to be considered in study design and interpretation. Using samples from family-based studies and twin studies may allow for the familial component of variation in metabolite levels to be measured in addition to the heritability contribution[13]. If longitudinal data are available, the associations can be checked to verify that the genetic contribution to the metabolic phenotype of the individuals remains stable over a longer time period[14].

Most of the large-scale studies with metabolic traits conducted so far originated from epidemiological studies that had previously collected and stored sample aliquots. This strategy of collecting samples for future analysis in large national cohorts and bio-banks, without the knowledge of the precise analysis techniques to be applied on them, made possible many of the present GWASs with metabolomics. The collection of such samples needs to be continued and extended by, for example, collecting the most extensive variety of samples, as it is not clear today on which, and on how many, different platforms these samples shall eventually be analysed. Aliquot numbers should be high, and individual volumes should be small to avoid thawing cycles. Harmonization of standard operation protocols (SOPs) for sample collection across centres is needed, and the impact of laboratory-specific differences, such as variation in centrifugation time and speed, needs to be assessed.

Another source of valuable study material for GWASs with metabolomics is clinical case–control studies. Including individuals with disease in such studies allows the investigation of potentially extreme metabolic phenotypes and the discovery of genetic associations that are only revealed under disease conditions. However, ensuring SOPs are followed in a clinical setting can be more challenging than in an epidemiological study. For example, whereas blood and urine samples taken under standardized conditions are generally available from epidemiological population studies, such conditions are more difficult to meet in a clinical setting. In particular, samples from cases and controls need to be treated identically as certain metabolites may be very susceptible to slight deviations from standard protocol. Strict SOPs need to be implemented, with a strong focus on homogeneous sample treatment, including sample storage at −80 °C and sample aliquoting at collection time to avoid any thawing of the samples between storage and measurement.

A recent study of 2,820 individuals that used non-targeted liquid chromatography coupled with tandem mass spectrometry (LC–MS/MS) and gas chromatography coupled with mass spectrometry (GC–MS) identified 37 genetic loci with genome-wide significance[15]. Q–Q plots from that study suggest that more than 500

Workflow	Considerations	Choices
Study design	Population	For example, Caucasian, Asian, African
	Study type	For example, population-based, twins study, clinical studies
	Sample type	For example, blood, urine, saliva
Sample collection	Standard operating protocols	Compatibility between study centres
	Fasting state	For example, fasting, non-fasting, controlled nutritional challenges
	Sample quantities	Serum, plasma, small volumes to avoid thawing
Sample storage	Temperature	–80°C, liquid nitrogen
	Aliquoting	200 µl for mass spectrometry, 1 ml for NMR, avoid thawing cycles
	Biobanking	Manual, automated
Sample preparation	Metabolite extraction	For example, polar, charged
	Derivatization	Changing biochemical properties for better measurement
Sample analysis	Method	^{1}H NMR, LC–MS/MS, GC–MS/MS
	Identification	Targeted, non-targeted, quantitative
	Provider	Proprietary, core facility, fee-for-service
Data analysis	Covariates	Age, gender, body mass index, medication, lifestyle
	Statistical analysis	For example, linear model, using ratios, advanced statistics
	Initial data processing	Log-normal scaling, principal-component transformation
Data interpretation	Functional	For example, GRAIL, overlay with eQTL data
	Biochemical	KEGG, HMDB
	Medical	GWAS catalogue, pharmacogenomics database

Figure 2 | **Workflow and considerations.** The basic steps of a high-throughput metabolomics experiment and design considerations for a genome-wide association study (GWAS) with metabolic traits. eQTL, expression quantitative trait locus; GC–MS/MS, gas chromatography–tandem mass spectrometry; GRAIL, Gene Relationships Across Implicated Loci; HMDB, Human Metabolome Database; KEGG, Kyoto Encyclopedia of Genes and Genomes; LC–MS/MS, liquid chromatography–tandem mass spectrometry.

loci show signals of association and may be confirmed as GIMs in more highly powered studies in the future. Nicholson *et al.*[13] present a power analysis for studies of a similar design (see figure 5 in that paper), estimating that associations with effect sizes down to $R^2 = 0.01$ are detectable with 80% power using samples sizes of around $N = 6{,}000$. However, the current experimental coverage of the metabolome in GWASs is still incomplete, and metabolic characterization of various sample types (other than blood and urine) and metabolic states (other than overnight fasting) is scarce. Even small studies conducted under conditions that have not previously been studied can therefore be expected to provide new associations of high biomedical interest.

Data analysis

Here we describe the main steps in data analysis. TABLE 3 highlights some important analytical challenges.

Initial data processing. Initial data processing should include investigation of any hidden internal structure in the data, such as dependence on measurement run day, and validation of the metabolite data against related traits that have been measured by independent methods, such as blood triglyceride and glucose levels or urine creatinine[14]. Extreme outliers should be removed in a general GWAS to avoid spurious associations with rare genetic variants. Note, however, that these outliers may correspond to diseased states and can be useful in more focused studies. As larger metabolite panels are used in more highly powered GWASs that can test gene variants with lower minor allele frequencies, hitherto unknown inborn errors of metabolism may be discovered. For such applications, it is then necessary not to remove outliers to identify associations between extreme metabotypes and rare genetic variants. However, false-positive rates are then expected

Table 3 | **Analytical challenges**

Challenge	Description and examples	Perspectives and needs
How to deal with a high correlation between metabolic traits	Bonferroni correction is often too strict; Benjamin–Hochberg may not apply as traits correlate in complex ways	Data transformations, including the use of partial correlations and principle component analysis to de-convolute the data; however, associations with transformed variables may become more difficult to interpret biochemically
How to analyse metabolomics data using multivariate methods	Analysis of ratios between metabolite pairs is a powerful bivariate method that provides a direct biochemical interpretation	Use of dedicated multivariate methods — for example, as described by Ferreira and Purcell[56] — possibly including biochemical pathway information and machine learning
How to process raw metabolomics data and how to handle extreme outliers or missing data points	Distribution of metabolomics data is not always normal; missing values may be due to failed detection or true absence; extreme outliers may be genuine but may also generate many false associations	Log-normal transformation appears to be reasonable in many cases; consider more complex Box–Cox and principal-components analysis space transformation; use of parameter-free tests that are independent of an assumed distribution
How to analyse related but not identical traits jointly that originate from different experimental techniques in a meta-analysis	Glucose can be measured by NMR, and some mass spectrometry methods provide the sum of hexoses; individual lipid species can be measured by dedicated lipidomics methods, whereas others determine bulk parameters, such as the sum of all carbon atoms and degree of desaturation in the lipid side chain (or chains)	Dedicated data analysis methods that account for specific biochemical properties need to be developed
How to analyse metabolomics data from different bio-samples jointly in a meta-analysis and replication	Some metabolites are differently preserved in blood plasma and serum; platforms may differ in extraction protocols	Dedicated statistical methods that account for specific differences between biosamples
How to identify true positives below the genome-wide significance threshold	An association between a metabolite and a SNP in an enzyme that metabolizes that metabolite is more likely to be a true positive	Bayesian-style reasoning based on prior information[57]
How to identify causative genes and functional variants	This is a general problem in genome-wide association studies, but it may be facilitated in the case of metabolic traits by using biochemical connections between the metabolic trait and biochemically related genes at a locus	Extend approaches based on ideas implemented in GRAIL (Gene Relationships Across Implicated Loci)[58]
How to derive causality between genetic variants and disease end points	A metabolic phenotype may be a functional intermediate trait or a correlated biomarker	Mendelian randomization, in which effect sizes are large enough
How to overlay metabolomics data with data from other 'omics' experiments	Chromatin immunoprecipitation- and sequencing-based high-throughput technologies, such as genome-wide gene expression, DNA methylation and microRNA data should be integrated into the chain of intermediate traits that lead from genotype to disease end point	Gaussian graphical modelling[59]; other network-based methods; more advanced systems-biology methods

to be high, and replication must be done carefully. If longitudinal data are available, the persistence of an extreme metabolic phenotype can be verified in individuals over time, thus indicating whether it constitutes a true extreme value (for an example, see the association of SNP rs37369 with 3-aminoisobutyrate concentrations in REF. 14).

We recommend log-scaling of the data, as it has been observed that metabolite concentrations are more often close to log-normal distributions than to normal distributions[15,16]. This is also coherent with testing metabolite ratios (see below). More sophisticated methods of initial data processing could include methods based on principal components analysis (PCA) and are briefly mentioned in TABLE 3. By transforming the data to PCA space, it may be possible to reduce its dimensionality and thus to reduce the multiple-testing burden. Also, outliers may be spotted more readily on PCA plots. For this purpose, specialized Web servers can be used[17]. However, biological interpretation of the association results also becomes more challenging in PCA space.

Testing for associations between SNPs and metabolite data. Testing for genetic association with metabolic traits is basically done as for any other quantitative trait. This is typically achieved by fitting an additive linear model with the covariates age and gender to the metabolite data and correcting for population stratification and family structure using software such as PLINK[18], Merlin[19], SNPTEST[20] or in-house R-scripts (see The R Project website). Because some metabolic traits strongly vary with parameters such as body mass index (BMI) and fasting state, these should be added to the model as well. Ideally, all measured metabolic traits would be tested for association with all available phenotypic parameters, and then all significantly associating parameters would be included as covariates into the model. Also, when measurements are done in batches or when samples come from distinct study centres, this information should be included[19].

GWASs with metabolomics may return massive amounts of information (far greater than GWASs with a single or few traits), and this represents a computational challenge. Selection of loci for further investigation

Additive linear model
A mathematical model used in statistical association analysis; here, it assumes a linear additive effect of the minor alleles on the metabolite concentrations.

requires implementation of clearly defined algorithms that pick SNPs for follow-up on the basis of objective criteria. However, manual curation of these loci is also necessary to detect overlapping or independent signals. This step should include viewing of all association data, not just the top associating metabolite. Putative causative genes can often be spotted by examining the function of the genes that lie in the linkage disequilibrium (LD) block around the lead SNP while bearing in mind the characteristics of these metabolites. Criteria for reporting associations should not be different from other GWASs, in that associations should meet genome-wide significance after Bonferroni correction for all tested loci and all traits, with replication in independent studies. Nevertheless, owing to the high correlation between many metabolic traits, it is likely that this is a very conservative approach. Associations below the genome- and metabolome-wide significance level should therefore be made publically available for inclusion in future meta-analyses studies (TABLE 3).

Because many of the individual metabolite measures are highly correlated, it may appear to be difficult to determine what is really driving both the association and the clinical risk changes. However, often the gene that hosts the causative variant can be readily identified owing to a match between the function of the gene and the associated metabolic trait. In these cases, the most parsimonious hypothesis is then that of a causal relationship in the direction gene → metabolite → disease phenotype. Other metabolites that also associate at the same locus are then most often identified as lying on the same pathway as the leading metabolic trait. High proportions of matches between gene function and associating metabolite (or metabolites) have been detected in a number of studies[10,13–16,21–24], and most associations were found to involve SNPs that are near enzyme, transporter or other metabolism-related genes.

Replication across studies. A lack of replication between studies of a few metabolic traits is sometimes observed. This may in part be attributed to differences in sample treatment, such as storage of blood samples at different temperatures, but many other factors may come into play that require investigation on a case-by-case basis. For instance, 15 loci of genome-wide significance were identified in a GWAS that involved 1,809 individuals and used a metabolomics kit that quantifies 163 metabolic traits, many of which are lipid-related species[16]. However, only 9 out of the 15 associations could be replicated in an independent population of 422 individuals (with a significance level of 0.05 adjusted for 15 tests)[16]. Nevertheless, a subsequent study that used the same metabolomics platform on an independent population replicated 12 of the 15 loci (with a significance level of 0.05 adjusted for 15 tests)[13]. The remaining three loci have been reported in association with a similar trait by GWASs using different metabolomics platforms[15,23]. This indicates that the genome-wide significance cut-off, accounting for all tested SNPs and all tested metabolic traits, is indeed a conservative threshold, as may be expected in the case of traits that are in part highly correlated.

Despite the technical issues that may affect reproducibility, most associations can be replicated well, even across metabolic traits that are different but related. For example, the fatty acid desaturase 1 (*FADS1*) locus was associated in a number of different studies with various species of glycerophospholipids (including phosphocholines, phosphoethanolamines and phosphoinositoles[15,16,21]) and sphingolipids[23], and with omega 3 and omega 6 fatty acids[22]; all of these species are related to arachidonic acid (C20:4). Also, some loci are found to associate with related traits in urine and blood, such as *N*-acetyltransferase 2 (*NAT2*)[14,15] and *NAT8* (REFS 13,15). As the number of GWASs with metabolic traits increases, it will become increasingly challenging (and informative) to combine such related, but not identical, association data in a meta-analysis (TABLE 3).

Integration with biochemical information. The beauty of the metabolic phenotype is that there is a rich knowledge base regarding many endogenous human metabolic pathways. In addition, more than 2,200 enzyme-coding genes are annotated in the human genome. This allows the corroboration of candidate associations with biological and functional arguments. Therefore, it is possible to analyse the association data from the point of view of a biochemist. Genes that are related to enzymatic and transport activities, and that are located in regions in LD with the lead SNP, are prime candidates for harbouring the causative variant. If such genes are present, researchers can then verify a biochemical link between these genes and the metabolic traits, using databases such as the Human Metabolome Database (HMDB)[25] and the Kyoto Encyclopedia of Genes and Genomes (KEGG)[26]. Currently, this is mostly done manually; dedicated and automated network analysis methods with statistical evaluation tools need to be developed for this task.

Association of ratios. Many metabolites are, by nature, highly correlated as they are interlinked by biochemical reactions in complex metabolic networks. A genetic variant is, in most cases, in association with several, biochemically related traits rather than a single metabolite concentration. This 'feature' of the metabolic phenotype can be merely a consequence of the biochemical correlation between the metabolites without any link to a genetic cause. However, we found that testing all possible ratios between metabolite concentrations in a GWAS is a very powerful method for identifying those correlations between metabolite pairs that have a genetic underpinning. This hypothesis-free testing of all metabolite ratios for association was first applied by our group in a metabolome-wide association study on diabetic mice[27] and subsequently used in our first GWAS[21]. If two metabolites constitute a product–substrate pair of an enzymatic reaction, then ratios between their concentrations are potential proxies of the corresponding reaction rate. In our pilot study, we found that the association of ratios between two metabolites that are related to product–substrate pairs of the FADS1 reaction was many orders of magnitude stronger than that of the concentrations of the two individual

Linkage disequilibrium
(LD). A nonrandom association between neighbouring gene variants; it is used to describe a region of high correlation between SNPs.

Glycerophospholipids
Glycerol-based phospholipids are major constituents of the membrane bi-layers and are found in association with low-density lipoprotein (LDL) and high-density lipoprotein (HDL) particles.

Sphingolipids
A class of lipids that contain a backbone of sphingoid bases

Box 2 | **Ratios between product and substrate concentrations**

The ratio between the concentration of a product and the concentration of its substrate approximates the biochemical reaction rate under idealized steady-state assumptions, and therefore the product/substrate ratio can be viewed as a proxy of the reaction rate. It has been observed that using ratios as quantitative traits in a genome-wide association study (GWAS) reduces the variance in the data set and increases the power of the GWAS by several orders of magnitude[21]. As a measure of this increase in the strength of association, the '*P* gain' was introduced. It is defined as the change in *P* value when using ratios compared to the smaller of the two *P* values when using two metabolite concentrations individually[28]. In many cases, this *P* gain is much larger than the loss in statistical power incurred by the increased number of hypotheses tests. If computational resources permit, we therefore recommend testing all possible ratios between metabolites for association.

Here is an example. Fatty acid desaturase 1 (*FADS1*) encodes a key enzyme in the metabolism of long-chain polyunsaturated omega 3 and omega 6 fatty acids. The minor allele variant of the rs174547 SNP associates with a reduced efficiency of the fatty acid delta-5 desaturase reaction[15]. The *P* value of the association with the product of the FADS1 reaction, arachidonic acid (C20:4), is 1.7×10^{-30}, and with the substrate, dihomolinolenate (C20:3), the *P* value is 3.3×10^{-9}. However, a test for association with the ratio between the product and the substrate, C20:4/C20:3, results in a strengthening of the association by 70 orders of magnitude and a *P* value of 3.6×10^{-101} (*P* gain = 4.8×10^{70}).

This can be explained by looking at the idealized reaction pathway for C20:4. Assuming that other sources and sinks of C20:4 can be neglected when compared to the FADS1 desaturation and the ELOVL2 elongation reactions, it reads:

$$C20:3 \xrightarrow{k_{FADS1}} C20:4 \xrightarrow{k_{ELOVL2}} C22:4$$

The thereof derived differential equation is:

$$\frac{d}{dt}[C20:4] = k_{FADS1}[C20:3] - k_{ELOVL2}[C20:4]$$

Its solution under steady-state assumption is:

$$\frac{[C20:4]}{[C20:3]} = \frac{k_{FADS1}}{k_{ELOVL2}}$$

With the additional assumption that the elongation reaction does not depend on genotype, it follows that the rate of the FADS1 reaction is proportional to the ratio between the concentrations of its product and substrate.

metabolites (BOX 2). This association was highly significant at a genome-wide level, even after Bonferroni correction for the additional tests induced by the use of ratios. Numerous subsequent studies confirmed that using metabolite concentration ratios as proxies for enzymatic reaction rates reduces the variance and yields robust statistical associations[28]. In the case of FADS1, if the molecular function of that enzyme was not already known, the association between the SNP and the associated metabolites may have allowed the deduction of its enzymatic activity of inserting a fourth double bond into long-chain fatty acids. In a way, it is the genetics that tells us which associations make sense biochemically.

Therefore, even if computationally expensive, we recommend testing all ratios between metabolite pairs for association. The ratios should be log-scaled owing to the relationship log(a/b) = –log(b/a), which means that the association of a ratio and of its inverse then yields identical results and thereby halves the multiple-testing burden. Large increases in the strength of association are a signal for biochemically informative associations. Petersen *et al.*[28] showed that an increase in the strength of association by a factor of ten times the number of tested ratios is a conservative upper bound for significance. Note, however, that a lack of such an increase does not prove the contrary: an important metabolite may be missing from the panel for technical reasons and thus will not show up in a ratio.

Further interpretation

Integration with other GWASs. Systematically overlaying GIMs with associations from GWASs with disease or disease-related end points allows the identification of potentially true positives in the list of associations that did not attain genome-wide significance. When a genetic locus has been proved to harbour a functional genetic variant by displaying a strong and replicated association with a metabolic trait, then the likelihood that a marginal association of that same variant (or a variant in strong LD) with a clinical end point is due to chance is much lower than it is for a variant that does not show any other signal associations. Although a solid statistical foundation of such a Bayesian-style argumentation is lacking at present, a first approach could be to combine the statistical association data from both GWASs using classical meta-analysis methods[29]. More research on how to combine association data from related but non-identical traits is needed. Such methods would also be needed for the meta-analysis of association data from GWASs that use different metabolomics techniques.

Functional genomics and metabolomics

GIMs also have the potential to inform basic science. The field of functional genomics aims to identify the function of all genes in the human genome. To this end, an association of a poorly characterized enzyme or transporter gene with a metabolic trait may generate testable hypotheses on their substrate specificities. Following up on the predicted function of solute carrier family 16 member 9 (SLC16A9; also known as MCT9) as a carnitine transporter, on the basis of its association with serum carnitine concentrations, experiments using radio-labelled carnitine and SLC16A9-expressing *Xenopus laevis* oocytes showed that this transporter is indeed a carnitine efflux pump[15]. This concept can also be inversed. For around one-third of all measured metabolites, their biochemical identity is at present unknown. Association of a well-characterized enzyme or transporter gene with a metabolite of unknown identity may be used to infer its biochemical nature. We have recently applied this approach to predict and experimentally validate the identity of a number of unknown metabolites, such as dipeptides, on the basis of their association with the dipeptidase angiotensin-converting enzyme (ACE)[30].

Gene–envrionment interactions

In addition to biomedical and pharmacological applications, co-association of genetic variation with metabolic traits and with certain lifestyles can also provide new insights into the functional basis of gene–environment

interactions. For instance, the aryl hydrocarbon receptor (*AHR*) locus was found to associate with coffee consumption habits[31] in a large population study. The strong association of this locus with serum caffeine concentrations[15] in a much smaller study suggests that genetic differences in caffeine metabolism are likely to be at the basis of this genetically influenced lifestyle choice.

At present, most interactions of genetic variance with environmental and lifestyle factors cannot be detected on a genome-wide scale owing to limitations in the statistical power of such studies. This problem may be overcome by limiting the tests of statistical interaction to genetic variants that are known to have a strong impact on the processes of human metabolism and that are linked to the relevant environmental or lifestyle factors. For instance, genetic variation in *FADS1* strongly modifies polyunsaturated fatty acid (PUFA) metabolism. Two independent studies showed that dietary intake of PUFAs modulates the association between genetic variation in *FADS1* and serum lipid levels[6,7] and thereby potentially also modifies the risk of cardiovascular disease. This statistical interaction between genetic variance and nutritional habits could only be identified because these studies focused on a known GIM locus. It would not have been statistically significant if it had been searched for on a genome-wide scale.

Current challenges and future directions

Knowledge of the full set of genetic variation in human metabolism will have a wide range of biomedical and pharmaceutical applications. The GIMs identified in GWASs can be used in clinical studies for association with response to drug treatment or with the development of particular complications during the course of a disease or treatment. Follow-up investigation of the GIMs in their biochemical context is likely to provide a better understanding of the pathogenesis of common diseases. Furthermore, it can be expected that knowledge of the genetic basis of human metabolic individuality will allow the separation of genetic and environmental factors in complex gene–environment interactions and will provide a rational starting point for personalized and gene-based health care and nutrition strategies. However, there remain many analytical challenges for GWASs with metabolomics, including those summarized in TABLE 3.

Eventually, the epidemiological approach of a wide range of patient phenotyping and sample collection, using strict SOPs, needs to be translated to clinical studies, as studies that implement physiological challenges may provide access to perturbed systems[8]. The most useful approach for understanding the causal roles of the metabolites (on the pathways from genetic variants to intermediate traits to disease end points) would be to use prospective cohorts that allow for future disease risks to be evaluated on the basis of both genetic and metabolic information.

Metabolic profiling of other biological samples, including saliva, cerebrospinal fluid, synovial fluid, semen and tissue homogenates, should be investigated in the future but have so far not been used in high-throughput population-based studies[32]. Studies in stool samples may be particularly challenging. Here, the effect of the gut microbiome on human metabolism needs to be taken into account, and this requires the additional characterization of the bacterial communities in the samples. Technical studies investigating, for example, the impact of differences between using serum and plasma[33] or of storing samples at different temperatures and for different times on the metabolite concentrations also need to be extended, ideally across platforms and institutions.

To date, GWASs have mostly focused on common variants from chip-based genotyping arrays. However, with better coverage of low-frequency variants through sequencing or dense imputation reference panels, more associations with metabolites will most probably be uncovered. In cases in which a metabolite can be identified as being functionally relevant and an intermediate trait on a pathway to a complex disorder, its genetic association can be used to fine-map the underlying disease risk locus to identify the disease-causing gene variant. For example, Tukiainen *et al.*[34] were able to fine-map known lipid loci using a dense marker set and detailed metabolite profiles.

The future resides in the combination of data from multiple 'omics' technologies. Inouye *et al.*[35] presented the first study of that kind by combining metabolomic, transcriptomic and genomic variation in a large, population-based cohort. One of the big challenges here is to combine all of these data in what may be termed a genome-wide systems-biology approach.

1. Garrod, A. E. The incidence of alkaptonuria a study in chemical individuality. *Lancet* **2**, 1616–1620 (1902).
2. Mootha, V. K. & Hirschhorn, J. N. Inborn variation in metabolism. *Nature Genet.* **42**, 97–98 (2010).
This comment provides an independant view on the potential of genetic studies with metabolomics.
3. Garrod, A. E. *Inborn Factors in Disease* (Oxford Univ. Press, 1931).
Archibald Garrod noted more than 80 years ago that "diathesis is nothing else but chemical individuality".
4. Psychogios, N. *et al.* The human serum metabolome. *PLoS ONE* **6**, e16957 (2011).
5. Link, E. *et al. SLCO1B1* variants and statin-induced myopathy—a genomewide study. *N. Engl. J. Med.* **359**, 789–799 (2008).
6. Dumont, J. *et al. FADS1* genetic variability interacts with dietary α-linolenic acid intake to affect serum non-HDL-cholesterol concentrations in European adolescents. *J. Nutr.* **141**, 1247–1253 (2011).
7. Lu, Y. *et al.* Dietary n-3 and n-6 polyunsaturated fatty acid intake interacts with FADS1 genetic variation to affect total and HDL-cholesterol concentrations in the Doetinchem Cohort Study. *Am. J. Clin. Nutr.* **92**, 258–265 (2010).
8. Krug, S. *et al.* The dynamic range of the human metabolome revealed by challenges. *FASEB J.* **26**, 2607–2619 (2012).
This paper reports a series of controlled physiological challenges that may be used in future GWASs with metabolic traits.
9. Nicholson, G. *et al.* Human metabolic profiles are stably controlled by genetic and environmental variation. *Mol. Syst. Biol.* **7**, 525 (2011).
This paper addresses essential questions about the heritability of metabolic traits.
10. Kettunen, J. *et al.* Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nature Genet.* **44**, 269–276 (2012).
This paper reports a large GWAS with NMR-derived metabolic traits.
11. Suhre, K. *et al.* Metabolic footprint of diabetes: a multiplatform metabolomics study in an epidemiological setting. *PLoS ONE* **5**, e13953 (2010).
This paper reports a pilot study on three different metabolomics platforms and provides practical insights into the possibilities and pitfalls of high-throughput metabolomics experiments.
12. Mittelstrass, K. *et al.* Discovery of sexual dimorphisms in metabolic and genetic biomarkers. *PLoS Genet.* **7**, e1002215 (2011).
13. 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).
14. Suhre, K. *et al.* A genome-wide association study of metabolic traits in human urine. *Nature Genet.* **43**, 565–569 (2011).

15. Suhre, K. *et al.* Human metabolic individuality in biomedical and pharmaceutical research. *Nature* **477**, 54–60 (2011).
This paper reports 37 loci of human metabolic individuality and provides examples for a wide range of biomedical and pharmaceutical applications.
16. Illig, T. *et al.* A genome-wide perspective of genetic variation in human metabolism. *Nature Genet.* **42**, 137–141 (2010).
17. Kastenmuller, G. Romisch-Margl, W., Wagele, B., Altmaier, E. & Suhre, K. metaP-server: a web-based metabolomics data analysis tool. *J. Biomed. Biotechnol.* **2011**, 839862 (2011).
18. 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).
19. 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).
20. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nature Rev. Genet.* **11**, 499–511 (2010).
21. Gieger, C. *et al.* Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. *PLoS Genet.* **4**, e1000282 (2008).
This paper reports the first GWAS with metabolic traits and with ratios between metabolite concentrations.
22. Tanaka, T. *et al.* Genome-wide association study of plasma polyunsaturated fatty acids in the InCHIANTI study. *PLoS Genet.* **5**, e1000338 (2009).
23. Hicks, A. A. *et al.* Genetic determinants of circulating sphingolipid concentrations in European populations. *PLoS Genet.* **5**, e1000672 (2009).
24. Demirkan, A. *et al.* Genome-wide association study identifies novel loci associated with circulating phospho- and sphingolipid concentrations. *PLoS Genet.* **8**, e1002490 (2012).
25. Wishart, D. S. *et al.* HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Res.* **37**, D603–D610 (2009).
26. Kanehisa, 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).
27. Altmaier, E. *et al.* Bioinformatics analysis of targeted metabolomics--uncovering old and new tales of diabetic mice under medication. *Endocrinology* **149**, 3478–3489 (2008).
28. Petersen, A. K. *et al.* On the hypothesis-free testing of metabolite ratios in genome-wide and metabolome-wide association studies. *BMC Bioinformatics* **13**, 120 (2012).
This paper provides a statistical underpinning to using ratios between metabolite concentrations in association studies.
29. Borenstein, M., Hedges, L. V., Higgins, J. P. T. & Rothstein, H. R. *Introduction to Meta-Analysis* (Wiley, 2009).
30. Krumsiek, J. *et al.* Mining the unknown: A systems approach to metabolite identification combining genetic and metabolic information. *PLoS Genet.* (in the press).
31. Cornelis, M. C. *et al.* Genome-wide meta-analysis identifies regions on 7p21 (AHR) and 15q24 (CYP1A2) as determinants of habitual caffeine consumption. *PLoS Genet.* **7**, e1002033 (2011).
32. Zhang, A., Sun, H., Wang, P., Han, Y. & Wang, X. Recent and potential developments of biofluid analyses in metabolomics. *J. Proteomics* **75**, 1079–1088 (2011).
33. Yu, Z. *et al.* Differences between human plasma and serum metabolite profiles. *PLoS ONE* **6**, e21230 (2011).
34. Tukiainen, T. *et al.* Detailed metabolic and genetic characterization reveals new associations for 30 known lipid loci. *Hum. Mol. Genet.* **21**, 1444–1455 (2012).
35. Inouye, M. *et al.* Metabonomic, transcriptomic, and genomic variation of a population cohort. *Mol. Syst. Biol.* **6**, 441 (2010).
36. Kottgen, A. *et al.* New loci associated with kidney function and chronic kidney disease. *Nature Genet.* **42**, 376–384 (2010).
37. Suhre, K. *et al.* Identification of a potential biomarker for FABP4 inhibition: the power of lipidomics in preclinical drug testing. *J. Biomol. Screen* **16**, 467–475 (2011).
38. 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).
39. 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).
40. Johnson, A. D. *et al.* Genome-wide association meta-analysis for total serum bilirubin levels. *Hum. Mol. Genet.* **18**, 2700–2710 (2009).
41. 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).
42. 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).
43. Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).
44. Kathiresan, S. *et al.* Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nature Genet.* **40**, 189–197 (2008).
45. Willer, C. J. *et al.* Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nature Genet.* **40**, 161–169 (2008).
46. Kathiresan, S. *et al.* Common variants at 30 loci contribute to polygenic dyslipidemia. *Nature Genet.* **41**, 56–65 (2009).
47. Kronenberg, F. in *Genetics Meets Metabolomics: from Experiment to Systems Biology* (ed. Suhre, K.) 255–264 (Springer, 2012).
48. The Wellcome Trust Case Control Conortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
49. Franke, A. *et al.* Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature Genet.* **42**, 1118–1125 (2010).
50. Kato, N. *et al.* Meta-analysis of genome-wide association studies identifies common variants associated with blood pressure variation in east Asians. *Nature Genet.* **43**, 531–538 (2011).
51. Rothman, N. *et al.* A multi-stage genome-wide association study of bladder cancer identifies multiple susceptibility loci. *Nature Genet.* **42**, 978–984 (2010).
52. Chambers, J. C. *et al.* Genetic loci influencing kidney function and chronic kidney disease. *Nature Genet.* **42**, 373–375 (2010).
53. Wallace, C. *et al.* Genome-wide association study identifies genes for biomarkers of cardiovascular disease: serum urate and dyslipidemia. *Am. J. Hum. Genet.* **82**, 139–149 (2008).
54. Li, S. *et al.* The *GLUT9* gene is associated with serum uric acid levels in Sardinia and Chianti cohorts. *PLoS Genet.* **3**, e194 (2007).
55. Doring, A. *et al.* SLC2A9 influences uric acid concentrations with pronounced sex-specific effects. *Nature Genet.* **40**, 430–436 (2008).
56. Ferreira, M. A. & Purcell, S. M. A multivariate test of association. *Bioinformatics* **25**, 132–133 (2009).
57. Ried, J. S. *et al.* PSEA: phenotype set enrichment analysis—a new method for analysis of multiple phenotypes. *Genet. Epidemiol.* **36**, 244–252 (2012).
58. Raychaudhuri, S. *et al.* Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet.* **5**, e1000534 (2009).
59. Krumsiek, J., Suhre, K., Illig, T., Adamski, J. & Theis, F. J. Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC Syst. Biol.* **5**, 21 (2011).
This paper introduces partial correlation networks to high-throughput metabolomics studies that may be used in a systems biology approach to GWAS with metabolic traits.

Acknowledgements

K.S. is supported by 'Biomedical Research Program' funds at Weill Cornell Medical College in Qatar, a program funded by the Qatar Foundation. The statements made herein are solely the responsibility of the authors. The authors thank G. Kastenmüller, A.-K. Petersen and J. Adamski for critical reading of the manuscript. We thank our reviewers for suggestions that led to the improvement of the manuscript.

Competing interests statement

The authors declare no competing financial interests.

FURTHER INFORMATION

Karsten Suhre's homepage: http://www.suhre.fr
GeneCards: http://www.genecards.org
GRAIL: http://www.broadinstitute.org/mpg/grail
GWAS server: http://www.gwas.eu
HapMap: http://hapmap.ncbi.nlm.nih.gov
Helmholtz Zentrum München (German Research Center for Environmental Health): http://www.helmholtz-muenchen.de/en
HMDB Serum Metabolome: http://www.serummetabolome.ca
Human Metabolome Database: http://www.hmdb.ca
Ingenuity Systems Pathway Analysis: http://www.ingenuity.com
KEGG: http://www.genome.jp/kegg
LIPID MAPS Lipidomics Gateway: http://www.lipidmaps.org
MassBank: http://www.massbank.jp
metaP-server: http://metabolomics.helmholtz-muenchen.de/metap2
***Nature Reviews Genetics* Series on Study designs:** http://www.nature.com/nrg/series/studydesigns/index.html
NHGRI Catalog of Published Genome-Wide Association Studies: http://www.genome.gov/gwastudies
Online Mendelian Inheritance in Man (OMIM): http://www.ncbi.nlm.nih.gov/omim
The Pharmacogenomics Knowledge Base (PharmGKB): http://www.pharmgkb.org
Software — Department of Epidemiology and Biostatistics — Karolinska Institutet: http://ki.se/ki/jsp/polopoly.jsp?d=26072&l=en
Weill Cornell Medical College in Qatar: http://qatar-weill.cornell.edu

ALL LINKS ARE ACTIVE IN THE ONLINE PDF