THEMATIC REVIEW

A description of large-scale metabolomics studies: increasing value by combining metabolomics with genome-wide SNP genotyping and transcriptional profiling

Georg Homuth, Alexander Teumer, Uwe Völker and **Matthias Nauck**[1]

Department of Functional Genomics, Interfaculty Institute for Genetics and Functional Genomics, University Medicine Greifswald, Friedrich-Ludwig-Jahn-Straße 15A, D-17487 Greifswald, Germany

[1]Institute for Clinical Chemistry and Laboratory Medicine, University Medicine Greifswald, Ferdinand-Sauerbruch-Straße, D-17475 Greifswald, Germany

(Correspondence should be addressed to G Homuth; Email: georg.homuth@uni-greifswald.de)

Abstract

The metabolome, defined as the reflection of metabolic dynamics derived from parameters measured primarily in easily accessible body fluids such as serum, plasma, and urine, can be considered as the omics data pool that is closest to the phenotype because it integrates genetic influences as well as nongenetic factors. Metabolic traits can be related to genetic polymorphisms in genome-wide association studies, enabling the identification of underlying genetic factors, as well as to specific phenotypes, resulting in the identification of metabolome signatures primarily caused by nongenetic factors. Similarly, correlation of metabolome data with transcriptional or/and proteome profiles of blood cells also produces valuable data, by revealing associations between metabolic changes and mRNA and protein levels. In the last years, the progress in correlating genetic variation and metabolome profiles was most impressive. This review will therefore try to summarize the most important of these studies and give an outlook on future developments.

Journal of Endocrinology (2012) **215,** 17–28

Toward a combination of genomics and metabolomics: a chronology

Technological prerequisites and general significance

Key prerequisites for the feasibility of genome-wide association studies (GWAS) were the availability of a first human genome sequence draft and subsequently the continuously increasing information about inter-individual differences in the human genome. Genetic polymorphisms underlying human phenotypic variability, namely single nucleotide polymorphisms (SNPs) and copy number variations (CNVs), were detected, mapped, and stored in large, publicly available databases, and their number is still growing as more and more human genomes are completely sequenced. Whereas some years ago the focus was on polymorphisms exhibiting minor alleles that were relatively frequent in the population (common polymorphisms), now the huge amount of available data also allows for the inclusion of rare polymorphisms in genome-wide analyses. On the other hand, advancing array technologies allowed genome-wide individual genotyping of thousands of polymorphisms at the same time, and with sinking costs it became affordable to apply this technology using large cohorts of thousands of individuals. This was the starting point of GWAS relating specific phenotypes to genetic polymorphisms and testing for statistically significant associations.

At present, SNPs are in the focus of GWAS, and the association of CNVs with specific phenotypes plays a clearly lesser role. This is most probably due to two reasons: first, the precise detection of CNVs represents a methodically more challenging undertaking compared to the differentiation of SNP alleles. Secondly, a key publication in 2010 described a GWAS of CNVs in 16 000 cases of eight common human diseases and 3000 shared controls (Wellcome Trust Case Control Consortium *et al.* 2010). The authors typed these 19 000 individuals into distinct copy-number classes at 3432 polymorphic CNVs, thereby including an estimated ~50% of all common CNVs with a size of larger than 500 bps. It turned out that most common CNVs are well tagged by defined SNPs and have therefore already been

This paper is one of three papers that form part of a thematic review section on Metabolomics. The Guest Editor for this section was Henri Wallasschofski, Ernst-Moritz-Arndt University, Greifswald, Germany.

DOI: 10.1530/JOE-12-0144
0022–0795/12/0215–017
Online version via http://www.endocrinology-journals.org

analyzed in SNP-based GWAS on the respective phenotypes using the established array platforms. Therefore, it was concluded that at least common CNVs are unlikely to contribute significantly to the genetics of common human diseases. However, time will tell whether rare CNVs may do so, which might also be the case for rare SNPs (see Keinan & Clark (2012)).

Of course, from the beginning, disease-relevant phenotypes were of particular interest in GWAS. However, the individual susceptibility for a given disease is only partially determined by genetic factors, and environmental influences are also of great importance, as are the individual lifestyle integrating specific risk factors like smoking, lack of physical exercise, and alcohol misuse. Whereas associated genetic polymorphisms are detectable with the mapping arrays, the influences of nongenetic factors like the aforementioned are probably reflected in the transcriptome and/or proteome of circulating blood cells. Total RNA prepared from whole blood of participants of large epidemiological cohorts can be used for array-based genome-wide gene expression profiling, and the generated data can be related to specific phenotypes as well as to genetic polymorphisms. Of course, it has to be kept in mind that the whole-blood transcriptome cannot be regarded as fully representative for all organs and tissues of the human body and that it also does not necessarily reflect other tissue-specific effects on the environment. For instance, the hepatocyte transcriptome can be predicted to exhibit a quite different reaction pattern in response to a defined toxic substance compared with the whole-blood transcriptome. However, if these limitations are adequately considered, the latter nevertheless represents a valuable data source.

Corresponding analyses of the proteome of blood cells or plasma samples are also possible; however, these require considerable technical expenditure, especially when performed in large numbers. In principle, all the three 'omics' technologies – genomics, transcriptomics, and proteomics – are able to identify biomarkers that allow for risk estimation of developing specific diseases. However, genomics approaches provide only information about the genetic contribution, the latter two also mirror environmental influences (see Fig. 1).

The metabolome, defined as the reflection of metabolic dynamics derived from parameters primarily measured in easily accessible body fluids such as serum, plasma, and urine, can be considered as the omics data pool closest to the phenotype because it integrates genetic influences as well as nongenetic factors including their impact on protein activity. Therefore, the metabolome represents a highly suitable source for deriving putative biomarkers. On the one hand, metabolome data can be related to genetic polymorphisms in GWAS, allowing for the identification of genetic factors underlying specific metabolic traits. The specific advantages of metabolomics for GWAS were recently summarized in a review by Adamski (2012). On the other hand, metabolic traits can be related to specific phenotypes, resulting in the identification of metabolome signatures caused by lifestyle factors. Finally, analyzing the correlation of metabolome data

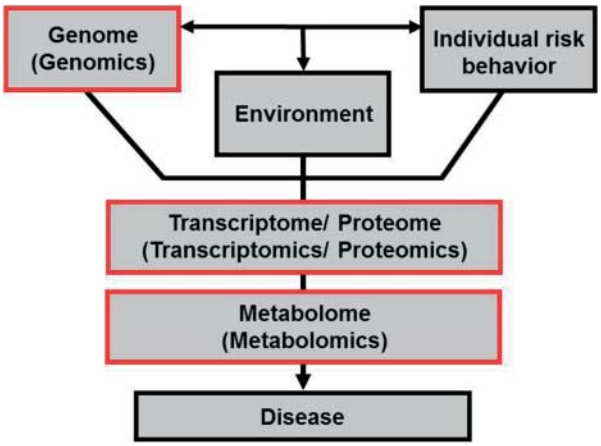

Figure 1 Interrelationship between the different 'omes', nongenetic factors, and their influence on disease development as well as the respective 'omics' technologies.

with transcriptional or/and proteome profiles of blood cells also produces valuable data, e.g. reveals the influence of changes in mRNA and protein levels as well as in their activity.

In recent years, the progress in correlation of genetic variation and metabolome profiles was most impressive, and therefore these developments will represent the main focus of this review.

GWAS of serum metabolites: introduction of the 'genetically determined metabotype' concept

One of the first key publications that described the successful combination of the GWAS approach and metabolome analyses was that of Gieger *et al.* (2008) published in November 2008. Within this study, the authors performed GWAS on metabolite levels in sera of 284 male individuals between 55 and 79 years from the population-based epidemiological KORA (Cooperative Health Research in the Region of Augsburg) F3 study. A targeted quantitative metabolomics platform based on electrospray ionization tandem mass spectrometry (ESI-MS/MS) was chosen to measure the fasting serum concentrations of up to 363 endogenous metabolites, including nine sugar molecules, seven biogenic amines, seven prostaglandins, 29 acylcarnitines, 18 amino acids, 85 sphingolipids, and 208 glycerophospholipids. MS data for 201 of these metabolites were obtained for more than 95% of the samples. Genome-wide individual SNP data of these individuals were produced with the Affymetrix GeneChip Human Mapping 500 K Array Set. In order to avoid false-positive effects from associations based on small numbers, the GWAS was limited to SNPs in which at least 5% of the population was homozygous for the minor allele. The corresponding minor allele frequencies in the analyzed data set were $>18{\cdot}2\%$, which means that only quite frequent SNPs, all representing 'common SNPs' with a minor allele frequency $\geq 5\%$, were included.

The identified associations between these SNPs and defined metabolite concentrations explained up to 12% of the observed variance in the human body's metabolic homeostasis. When ratios of certain metabolite concentrations were used as proxies for enzyme activities, even 28% of the variance was explained, where highly significant *P* values between 10^{-10} and 10^{-21} could be determined. In the case of four associated SNPs within the genes *FADS1*, *LIPC*, *SCAD* (*ACADS*), and *MCAD* (*ACADM*) encoding fatty acid desaturase 1, hepatic lipase, C-2 to C-3 short-chain acyl-CoA dehydrogenase, and C-4 to C-12 straight-chain acyl-CoA dehydrogenase, respectively, the corresponding metabolic phenotypes or 'metabotypes' clearly matched biochemical pathways of the lipid metabolism in which these enzymes are involved. Individuals with different alleles of these SNPs exhibited significantly different metabolic capacities concerning the synthesis of some polyunsaturated fatty acids (PUFAs), the β-oxidation of short- and medium-chain fatty acids, and the breakdown of triglycerides.

As one example, the associated SNP in *FADS1*, namely rs174548, strongly influenced the serum glycerophospholipid homeostasis. As these molecules play major roles in cholesterol metabolism, Gieger *et al.* (2008) hypothesized that rs174548 might affect corresponding serum parameters in a sufficiently large population. Indeed, two earlier GWAS with up to 18 000 participants (Kathiresan *et al.* 2008, Willer *et al.* 2008) reported *P* values of association for rs174548 with serum LDL cholesterol, HDL cholesterol, and total cholesterol levels between $1{\cdot}89\times10^{-4}$ and $6{\cdot}07\times10^{-5}$. Due to the fact that these *P* values were not sufficiently low to pass the generally accepted threshold for genome-wide significance of $5{\cdot}0\times10^{-8}$ (Pe'er *et al.* 2008), the associations were not considered for replication in the two original GWAS mentioned earlier. The newly detected association of rs174548 with different glycerophospholipids could be now viewed as an indirect replication of the association of *FADS1* with HDL, LDL, and total cholesterol levels in an independent population sample. In addition, it was now possible to hypothesize that the observed change in cholesterol levels caused by this SNP is functionally linked to the availability of polyunsaturated long-chain fatty acids with four and more double bonds and its impact on the homeostasis of different glycerophospholipids.

These results convincingly demonstrated that SNPs cause major differences in the individual metabolic basic equipment within the human population and pointed toward a novel approach to personalized health care by combining genotyping and metabolic characterization. It became clear that genetically determined metabotypes (GDMs) as intermediate phenotypes may strongly influence the individual susceptibility for diseases as well as responses to drug treatments or nutritional interventions and environmental challenges.

Characterization of the smoker's serum metabolome

Shortly after the pioneering work of Gieger *et al.* (2008), a further study demonstrating the remarkable opportunities of metabolomics in biomedical research even without the integration of genetic data was published, with a partially overlapping list of authors and using almost the same study cohort. This study by Wang-Sattler *et al.* (2008) dealt with a topic of even more medical relevance: the metabolic consequences of cigarette smoking. Among the 283 randomly selected male participants of the population-based KORA cohort that were included in this study (one fewer than in the work of Gieger *et al.* (2008)), 28 were current smokers who smoked one to 50 (mean 17) cigarettes per day. The 154 participants who ceased smoking but formerly smoked at least one cigarette daily were classified as former smokers. The 101 nonsmokers had never smoked before. As exemplified before, all individuals were metabolically profiled by ESI-MS/MS to measure the fasting serum concentrations of up to 363 endogenous metabolites. Finally, 198 metabolites were used for subsequent analyses in the smoking study.

Multivariate analysis of the generated metabolic profiles allowed clear differentiation of the smoker group within the population sample from former smokers and nonsmokers, where significant changes were primarily detected for clusters of lipid metabolites. In addition, 23 defined lipid metabolites could be identified as nicotine-dependent biomarkers exhibiting higher levels in smokers compared with former and nonsmokers, with the notable exception of three acyl-alkyl-phosphatidylcholines (e.g. plasmalogens). Consistently, the ratios of plasmalogens to diacyl-phosphatidylcholines that are regulated by the enzyme alkylglycerone phosphate synthase (alkyl-DHAP) in both ether lipid and glycerophospholipid pathways were significantly reduced in smokers. These results were consistent with the smoker-specific strong downregulation of *AGPS* encoding alkyl-DHAP that has been identified formerly in a gene expression analysis of human lung tissues. The data suggested association of smoking with plasmalogen deficiency disorders, caused by decreased or complete loss of peroxisomal alkyl-DHAP activity.

Wang-Sattler *et al.* (2008) emphasized the physiological importance of lipids, which is reflected by a variety of diseases that are related to lipid abnormalities, like atherosclerosis, diabetes, obesity, and Alzheimer's disease. Lipids, representing major structural components of biological membranes, ensure cellular integrity and allow for subcellular compartmentalization in organelles. Accordingly, the findings that the most significant metabolite changes in this study were detected for lipid metabolite clusters and that the 23 identified biomarkers represented lipids could indicate that membrane damage is caused by components of the tobacco smoke. Such dysregulations were predicted to subsequently modify the concentrations of related metabolites like the identified biomarkers.

GWAS of serum metabolites: emphasizing the value of the GDM concept in larger studies

Early in 2010, the first high-ranking publication in the field of combined genomics–metabolomics approaches was published in *Nature Genetics*. Again, the group from the Helmholtz Zentrum Munich including Christian Gieger, Rui Wang-Sattler, Karsten Suhre, Hans-Erich Wiechmann, Jerzy Adamski, and Thomas Illig, also co-authors of the aforementioned two publications, was in charge of this study by Illig *et al.* (2010).

The authors presented a GWAS of metabolites measured in serum that was similar to that described in the work of Gieger *et al.* (2008); however, this time the discovery cohort comprised 1809 participants of the KORA study instead of only 284. Furthermore, the most significant association findings were subsequently replicated in an independent population consisting of 422 participants of the TwinsUK cohort. The KORA and the TwinsUK samples were individually genome-wide genotyped using the Affymetrix 6.0 GeneChip array and the Illumina Hap317 K chip respectively. The fasting serum concentrations of 163 metabolites that covered a defined panel of amino acids, sugars, acylcarnitines, and phospholipids were measured by ESI-MS/MS with the Biocrates AbsoluteIDQ targeted metabolomics technology. Based on their previous finding that use of metabolite concentration ratios as proxies for enzymatic reaction rates reduces the variance and yields robust statistical associations, the authors additionally tested all possible metabolite concentration ratios ($163 \times 162 = 26\,406$ traits) with a linear additive model for association with all SNPs passing their selection criteria. The corresponding estimated genome-wide significance level after correction for multiple testing was $P < 3{\cdot}64 \times 10^{-12}$. The advantage of this approach was the focus on pairs of metabolites that are more likely to be coupled either biochemically or physiologically.

A two-step discovery design in the KORA population was chosen: first, an initial discovery sample of 1029 male and female individuals from the KORA cohort was used, and all loci with P values of association $< 10^{-7}$ for metabolite concentrations and $P < 10^{-9}$ for concentration ratios in a GWAS were selected. Altogether, 32 loci fulfilled these criteria. Secondly, one SNP for each locus was tested in a second step in an independent sample of 780 participants selected from the remaining KORA population. Using data from all 1809 individuals, joint P values of association were then calculated, which resulted in the identification of 15 loci for which the strength of association increased when the additional data were added. These loci, exhibiting genome-wide significant P values of association smaller than $3{\cdot}64 \times 10^{-12}$, were used for replication in the TwinsUK cohort. Of the 15 loci tested, nine were clearly replicated ($P < 0{\cdot}05$) after Bonferroni correction for 15 tests. For eight of these, the lead SNPs were located within or close to genes (*FADS1*, *ELOVL2*, *ACADS*, *ACADM*, *ACADL*, *SPTLC3*, *ETFDH*, and *SLC16A9*) encoding enzymes or solute carriers with functions that matched the associated metabolic traits. The identified loci explained 5·6–36·3% of the observed total variance in metabolite concentrations.

For several of the associated loci detected by Illig *et al.* (2010) relations to clinically relevant phenotypes had already been reported: for instance, *FADS1* encoding fatty acid desaturase 1 has been described to be associated with attention-deficit/hyperactivity disorder (Brookes *et al.* 2006) as well as with cholesterol and triglyceride levels (Kathiresan *et al.* 2009), and *ACADS* encoding C-2 to C-3 short-chain acyl-CoA dehydrogenase represents a susceptibility locus for ethylmalonic aciduria (Tanaka *et al.* 2009). All in all, this study once more impressively demonstrated that serum metabolite concentrations provide a direct readout of biological processes and highlighted the importance of the GDMs for the development of diseases. Related GWAS studies, however, dealing with clearly more focused phenotypes, namely plasma levels of PUFAs (Tanaka *et al.* 2009) and of circulating sphingomyelin concentrations (Hicks *et al.* 2009), had already been published in 2009.

Characterization of the diabetes plasma metabolome and extension of the analytical tools

At the end of 2010, the Helmholtz team again demonstrated the strong biomedical impact of metabolomics even in the absence of complementary genomic data by publishing a study in which a metabolic signature of diabetes in plasma samples was characterized. In this work of Suhre *et al.* (2010), the concentrations of more than 420 metabolites were measured in overnight fasting blood of 40 individuals with self-reported diabetes and 60 control individuals without diabetes randomly selected from the participants of the KORA study, all male and over 54 years.

Of particular importance, this time a combination of three complementary analytical platforms covering nuclear magnetic resonance (NMR) and MS/MS was applied. Three different metabolomics providers performed metabolite detection and quantification: Biocrates Life Sciences AG, Chenomx Incorporated, and Metabolon Incorporated located in Austria (Innsbruck), Canada (Edmonton), and USA (Durham) respectively. In the case of the Biocrates platform, a targeted profiling scheme based on ESI-MS/MS using multiple reaction monitoring (MRM) as well as neutral loss and precursor ion scans was used for a quantitative screen for already known small molecule metabolites. Selective detection and quantification of 201 metabolites belonging to the compound classes amino acids, biogenic amines and polyamines, reducing mono- and oligosaccharides, glycerophospho- and sphingolipids, eicosanoids, and other oxidized PUFAs were carried out. Appropriate internal standards structurally identical but containing stable isotopes like deuterium, ^{13}C, or ^{15}N allowed the absolute quantitation of metabolites. The data set represented a subset of that used in the smoking study by Wang-Sattler *et al.* (2008) described

earlier. Altogether, 24 metabolites were identified in the EDTA plasma samples in the NMR analysis performed by the Chenomx platform. These included alcohols, amides, amines, amino acid derivatives, amino acids, aromatic compounds, fatty acids, food/drug components, organic acids, and sugars. Two separate ultra-high-performance liquid chromatography/MS/MS (UHPLC–MS/MS) injections and one gas chromatography (GC)–MS injection per sample were combined by the Metabolon platform, where the UHPLC injections were optimized for basic and acidic species. Subsequently, searches of the generated MS/MS data against an in-house generated authentic standard library were accomplished. The library encompassed retention time, molecular weight, preferred adducts, and in-source fragments for all molecules, including their associated MS/MS spectra. Using this platform, 257 small molecules could be detected. The final data set encompassed 482 distinct values of absolute (Biocrates and Chenomx) or relative (Metabolon) metabolite concentrations that were available for analysis. Remarkably, among these, only 50 metabolites were quantified on more than one platform.

The multiplatform approach allowed the identification of several novel deregulated metabolites that associate with diabetes. For instance, whereas several glycerophospholipids exhibited pronounced negative associations, corresponding phosphatidylethanolamines carrying similar lipid side chains were positively associated. Medium-chain length fatty acids as well as arachidonate were negatively associated, while several long-chain fatty acids including PUFAs like linolate and linolenate exhibited positive associations. Positive associations with diabetes were also found for the branched chain amino acids leucine, isoleucine, and valine; their gamma-glutamyl derivates; and the ketone body 3-hydroxybutyrate (BHBA). By contrast, acetate concentrations were lower, as indicated by the ratio between acetate and BHBA. Furthermore, perturbations of metabolic pathways related to kidney dysfunction (3-indoxyl sulfate) and the interaction with the gut microflora (bile acids) were detected. In addition, as could be predicted, up to 90% increased concentrations of glucose, mannose, deoxyhexose (primarily deoxyglucose), uronic acid (primarily glucuronic acid), dihexose (primarily maltose), and several products from the biosynthesis or the degradation of glycosylated proteins or glycolipids were detected in the diabetes group, demonstrating the strongest positive associations for sugar metabolites.

In sum, the study of Suhre *et al*. demonstrated a plasma metabolome signature of diabetes that included dysregulated carbohydrate metabolism, perturbated lipid metabolism, mild signals of ketosis, and early signals of impaired renal function. Besides the major finding that metabolic markers might be useful in the early detection of diabetes-related complications already under subclinical conditions, the results furthermore clearly demonstrated that combining different techniques allows a pronounced extension of the analytical potential of metabolomics approaches.

GWAS of urine metabolites

During 2011, a real boost in higher ranking studies on metabolomics/genomics-related topics could be observed. In June of that year, the Helmholtz team, together with colleagues from the Greifswald University in Northern Germany, again successfully published a study in *Nature Genetics*. This work of Suhre *et al*. (2011*a*,*b*) represented the first GWAS on metabolic traits measured in human urine instead of serum or plasma. Urine was chosen because it represents the main vehicle for metabolite and protein excretion and it provides access to a part of the body with very different metabolic capacities compared with serum or plasma. Furthermore, urine is the biomaterial primarily used in diagnosis of kidney diseases. Therefore, the study of Suhre *et al*. was designed to identify GDMs detectable by analyzing human urine to investigate the detoxification capacity of the human body and to identify genetic loci predisposing for chronic kidney diseases (CKDs).

Urine samples from male participants of the population-based Study of Health in Pomerania (SHIP) were collected and 400 MHz ^{1}H-NMR spectra were obtained from these specimens. Metabolite concentrations were annotated and quantified manually in these NMR spectra. For the SHIP cohort, genome-wide individual genotyping data generated using the Affymetrix Human SNP Array 6.0 were available. First, altogether 59 metabolite concentrations and 1661 ratios between metabolite concentrations were tested for associations with 645 249 autosomal SNPs in 862 male SHIP participants serving as the discovery cohort. After correction for multiple testing, five loci exhibited significant associations at a genome-wide significance level of $P<4{\cdot}51\times10^{-11}$. Because the strength of association for the next-best locus was three orders of magnitude lower, the study was limited to these five hits. Next, the positive association findings were replicated using an additional 870 female SHIP participants as well as 992 KORA samples of both genders representing an independent cohort. All five loci were replicated in KORA and showed comparable effect sizes in the SHIP females group, resulting in joint P values of association ranging between $3{\cdot}2\times10^{-19}$ and $2{\cdot}1\times10^{-182}$. These results explained metabolite variances between 3·2 and 33·4%. Finally, for verification of the association results over time, samples from 170 male participants of the discovery cohort recruited in a 5-year follow-up study were analyzed. Of the five loci tested, four were verified in this less powered verification study, whereas the fifth locus only nearly missed the significance level. Three loci were already described to be of biomedical relevance: the genes *SLC7A9*, *NAT2*, and *SLC6A20* represented known risk loci for CKD, coronary artery disease (CAD) as well as genotype-dependent response to drug toxicity, and iminoglycinuria respectively. In addition, one SNP in *AGXT2* encoding alanine–glyoxylate aminotransferase 2 was identified as the genetic basis of hyper-β-aminoisobutyric aciduria.

As one example, the lead SNP of the *SLC7A9* locus is located upstream of the gene that encodes member 9 of the solute carrier family 7. This protein represents a light subunit of a high-affinity, sodium-independent transporter of cystine and neutral and dibasic amino acids. Mutations in *SLC7A9* cause non-type I cystinuria, a disease leading to cystine stones in the urinary system and might therefore be associated with CKD (Evan *et al.* 2006, Mattoo & Goldfarb 2006). Two recent GWAS identified an association of the *SLC7A9* locus with the estimated glomerular filtration rate (eGFRCrea), with serum creatinine, and with CKD (Chambers *et al.* 2010, Köttgen *et al.* 2010). In the study of Suhre *et al.*, a test of the ratio between lysine and valine produced a quite strong association with the locus lead SNP, followed in strength by associations with the lysine/pyroglutamate and lysine/asparagine ratios. These results convincingly suggested lysine as a substrate of the transporter encoded by *SLC7A9*, consistent with the hypothesis that an SNP in this gene confers different reabsorption efficiency for lysine, leading to inverse effects on the metabolism of valine, pyroglutamate, and asparagines that are obviously linked to the development of CKD.

Importantly, the work of Suhre *et al.* for the first time described a focus shift of the combined metabolomics/genomics approaches away from metabolic homeostasis, as can be observed in human serum or plasma, toward the renal detoxification capacities, as evidenced by excreted metabolites in human urine. Furthermore, that study represented the first GWAS on metabolic traits that were determined by NMR spectroscopy, a profiling technology complementary to MS that provides a comprehensive coverage of the human metabolome, given that the compounds to be analyzed are present in higher concentrations (micromolar range).

Comparison of serum and plasma metabolomes

A study published in July 2011 that was led by the Helmholtz team described the comparison of the two main specimens used in metabolome analyses up to then, serum and plasma. In this study of Yu *et al.* (2011) genomic data were not considered because its aim was to characterize the influence of different collecting procedures as well as of the coagulation cascade on measured protein and metabolite concentrations: serum, on the one hand, is obtained from blood that has coagulated. By centrifugation, fibrin clots that formed during coagulation, along with blood cells and related coagulation factors, are separated from serum, where this process is parallelized by the release of proteins like cytokines from platelets into the latter. On the other hand, plasma is obtained by adding anticoagulants like EDTA or heparin before removing the blood cells.

The concentrations of 163 metabolites were analyzed using a commercially available metabolomics kit (AbsoluteIDQ kit p150, Biocrates Life Sciences AG, Innsbruck, Austria) based on Flow Injection Analysis (FIA)-MS in EDTA plasma and serum samples that were collected simultaneously from 377 fasting individuals of the KORA study. Samples were measured separately in ten plates. After excluding 41 metabolites due to low measurement stability in order to ensure the quality of the data, plasma and corresponding serum samples from 83 randomly chosen individuals were additionally remeasured in the same plates and mean correlation coefficients (r) of all metabolites between the duplicates were determined. These amounted to 0·83 in plasma and 0·80 in serum, which indicated significantly better stability of plasma compared with serum, although the absolute differences in r were small. The metabolite profiles from plasma and serum were demonstrated to be clearly distinct, where 104 metabolites exhibited significantly higher concentrations in serum. In particular, nine metabolites showed relative concentration differences larger than 20%. Altogether, although there were measurable differences in absolute concentrations between the two sample sources, overall correlation was high for most metabolites (mean $r=0{\cdot}8160$), thereby reflecting a proportional change in concentration. Of note, the comparison of two groups of individuals with different phenotypes using both serum and plasma data revealed that more metabolites exhibited significantly different concentrations in serum. For example, when 51 type 2 diabetes (T2D) patients were compared with 326 non-T2D individuals, 15 more significantly different metabolites were identified in serum in addition to 25 detected for both specimens. Comparison of male and female individuals as well as of smokers and nonsmokers produced similar results: in each case, serum samples contained larger numbers of significantly different metabolites.

Importantly, the study of Yu *et al.* (2011) demonstrated generally good reproducibility in both plasma and serum and even better reproducibility in plasma. It was shown that similar results should be obtained whether the same source of metabolites is used in metabolome studies; however, the higher concentrations in serum might allow for more sensitive biomarker detection.

Identification of gender-specific metabolomic differences

The study from Mittelstrass *et al.* (2011) that was published in August 2011 for the first time described the identification of sexual dimorphisms in metabolic and genetic biomarkers. Until then, most studies that associated metabolite concentrations and ratios of measured metabolites with defined phenotypes had not considered sexual dimorphism and did not perform data stratification by gender. However, there are well-known correlations between gender and incidence, prevalence, age of onset, symptoms, and severity of diseases, as well as reaction to drugs. Nevertheless, a survey of studies published in 2004 of nine different medical journals found that only 37% of the participants were women – even only 24% when restricted to drug trials – and only 13% of studies analyzed the generated data by sex (Kim *et al.* 2010). Therefore, in the study of Mittelstrass *et al.* (2011) that was once more led by the Helmholtz team,

gender-specific differences of serum metabolite concentrations as well as the underlying genetic background, particularly genotypic metabotype differences between males and females, were analyzed.

For the discovery and replication analyses, the authors used more than 3300 independent individuals from the KORA study with available measurements of 131 biologically relevant metabolites that covered a biologically relevant panel, divided into five subgroups involving amino acids, phosphatidylcholines, sphingomyelins, acylcarnitines, and C6-sugars. Out of these, significant concentration differences between males and females were identified for 102 metabolites by a linear regression approach (P values $<3{\cdot}8\times10^{-4}$). At least one metabolite of each mentioned subgroup exhibited significant sex-specific differences in metabolite concentrations. The results suggested that gender-specific concentration differences are not randomly spread over the different metabolites but affect defined metabolic pathways. Of particular interest was the finding that most sphingomyelins were significantly lower in men compared with women, because there is direct experimental evidence for a role of sphingolipids (sphingomyelins and ceramides) in the development of several common complex chronic diseases including atherosclerotic plaque formation, myocardial infarction, cardiomyopathy, pancreatic β-cell failure, insulin resistance, coronary heart disease, and T2D (Holland & Summers 2008, Yeboah *et al.* 2010). Because other lipid-derived molecules like bile acids were already demonstrated not to be sex specific (Rodrigues *et al.* 1996), it might be concluded that especially sphingomyelins represent important intermediate phenotypes involved in the modulation of gender-specific disease susceptibility.

The complementary GWAS on gender-specific genetic variations in human metabolism was performed using data from 2000 KORA participants. Here, sex-stratified GWAS adjusted for age and body mass index (BMI) were performed for logarithmic concentrations of all metabolites. Gender differences were identified by testing the estimated SNP effects for heterogeneity between men and women. This GWAS indeed demonstrated gender-specific differences in the effects of genetic variations on metabolites in men and women: altogether, eight SNPs on chromosome 2 showed genome-wide significant differences in SNP effects (β-estimates) between men and women for association with glycine. The absolute estimates of all these SNPs were higher in women ($\beta=20{\cdot}2$) compared with men ($\beta=-0{\cdot}067$), where SNP rs715 showed the strongest gender difference with a P value of $3{\cdot}65\times10^{-10}$ for the test of β-estimate differences. SNP rs715 is located within the 5′-UTR of *CPS1* encoding the mitochondrial enzyme carbamoyl-phosphate synthase 1, which catalyzes the first committed step of the hepatic urea cycle. Besides others, the *CPS1* locus had already been identified in the aforementioned study of Illig *et al.* (2010) to be genome-wide significantly associated with glycine concentrations. However, these results were not gender stratified and consequently could not demonstrate the sex-specific differences, again demonstrating the necessity of such stratification.

Convincingly, the study of Mittelstrass *et al.* (2011) demonstrated that there are significant differences between the metabolite profiles of males and females. In addition, sexual dimorphism for specific genetic variants in metabolism-related genes was described. This work laid basis for monitoring in personalized medicine differentiating, e.g. amino acid or lipid metabolism in woman and man. The findings were of special interest because they underscored the imperative necessity of considering gender-specific effects as well in the design as in interpretation of such studies.

Introducing ^{1}H-NMR spectroscopy in metabolome analysis

Finally, two studies of special interest in the field of combined genomics/metabolomics were published in September 2011. Nicholson *et al.* presented a GWAS of the ^{1}H-nuclear magnetic resonance spectroscopy (^{1}H-NMR) metabolome in human urine and plasma samples collected from two cohorts of individuals of European descent, where one cohort consisted of female twins donating samples longitudinally. These authors newly introduced the term ‘metabolite quantitative trait locus (mQTL) analysis’ in this field to characterize studies that allow the identification of genetic polymorphisms with sufficiently strong effects on metabolism by performing GWAS of metabolite profiles. Accordingly, mQTL analyses will allow for the identification of GDMs.

^{1}H-NMR represents an untargeted, discovery-driven approach covering many important metabolites. Using this technique, Nicholson *et al.* addressed the question about the presence of ^{1}H-NMR-detectable metabolites in urine or plasma strongly influenced by genetic polymorphisms. To this end, metabolite concentrations were quantified by ^{1}H-NMR and subsequently tested for genome-wide association with SNPs. Plasma and urine samples were collected from participants across the MolTWIN and the MolOBB cohorts. The MolTWIN cohort comprised 142 postmenopausal female twins of Northern European descent (51 monozygotic and 19 dizygotic pairs) and two unrelated individuals between 45 and 76 years of age who donated samples longitudinally. The MolOBB cohort comprised 69 participants selected from the Oxford Biobank (OBB) on the basis of case/control status for metabolic syndrome. The final set of subjects encompassed 42 controls (17 females and 25 males) and 27 cases (12 females and 15 males). For all participants, ^{1}H-NMR spectra on plasma and urine samples were generated using a 600 MHz spectrometer. Genome-wide individual genotyping data were measured using the Illumina 317 K BeadChip SNP array where from the MolTWIN cohort only one monozygotic twin from each pair was genotyped, while both members of each dizygotic twin pair were genotyped. After completed quality control and imputation, data on 2 541 644 autosomal SNPs were available for association analysis on a total of 155 individuals: 68 from the MolOBB cohort and 87 from MolTWIN.

The concentrations of four metabolites, namely trimethylamine, 3-aminoisobutyrate, an *N*-acetylated compound, and dimethylamine, exhibited significant and replicable association with SNPs ($8\times10^{-11}<P<2{\cdot}8\times10^{-23}$). Whereas the first three of these were measured in urine, dimethylamine was determined in plasma. Trimethylamine and dimethylamine mapped to a single genetic region; therefore, finally three distinct genetic loci were detected to be associated. Interestingly, two of these three loci were located within haplotype blocks at *2p13.1* and *10q24.2* carrying the genetic signature of strong recent positive selection in European populations, indirectly pointing toward biomedical relevance. The genes *NAT8* encoding a putative *N*-acetyltransferase and *PYROXD2* encoding a protein with a pyridine nucleotide-disulphide oxidoreductase domain represent putative candidates causatively mediating the corresponding mQTL associations, although the precise functions of these proteins is not clear. In the case of *PYROXD2*, the haplotype relatively advantageous in European populations turned out to be associated with a decreased expression of the gene and an increased concentration of trimethylamine in urine and dimethylamine in plasma. In the case of *NAT8*, consistent with the observed different concentrations in *N*-acetylated compounds in urine, the encoded cysteinyl-conjugate *N*-acetyltransferase CCNAT catalyzes *N*-acetylation, and the *NAT8* locus has already been associated with renal function (Juhanson *et al.* 2008, Chambers *et al.* 2010, Köttgen *et al.* 2010).

A detailed variance components analysis of the sources of population variation in the metabolite levels was possible because of the longitudinal twin design of the study. According to the results of this analysis, the mQTLs explained between 40 and 64% of biological population variation in the concentrations of the corresponding metabolites. These effect sizes were stronger than those reported in the earlier study of metabolites by Illig *et al.* (2010) described earlier that used the targeted metabolomics Biocrates platform for serum analysis.

A further aim of the study of Nicholson *et al.* was to provide further support for the findings described by Illig *et al.* (2010). Therefore, they performed a replication analysis of these results using the Biocrates platform to assay their own set of plasma samples. The plasma sample metabolite concentrations for both cohorts were determined by MS where MRM detection combined with the use of stable isotope labeled and other internal standards enabled data quantification. All metabolite concentrations were initially calculated in mM. Altogether, 12 of the 15 mQTLs described by Illig *et al.* (2010) were replicated, with four additional mQTLs beyond the nine replicated by Illig *et al.*(2010) themselves, which means that a total of 13 of the 15 mQTLs identified by Illig *et al.* (2010) were successfully replicated by Nicholson *et al.*

Convincingly, the study of Nicholson *et al.* demonstrated the additional benefit of ^{1}H-NMR spectroscopy in metabolome analyses: the Biocrates platform represents a targeted approach focusing on a specifically selected set of lipids and amino acids. By contrast, the use of the untargeted ^{1}H-NMR spectroscopy approach allows for quantification of the most abundant 50–100 metabolites in a biofluid that typically exhibit concentrations higher than 10 μM. Only five plasma metabolites, namely the amino acids glutamine, glycine, leucine, tyrosine, and valine, were targeted by the Biocrates platform and also quantified by ^{1}H-NMR spectroscopy. This demonstrated that the two approaches are complementary, detecting widely nonoverlapping metabolite sets.

Even more large-scale GWAS of serum metabolites

The last study to be discussed in this overview that primarily deals with publications of outstanding interest in the metabolomics/genomics context is that of Suhre *et al.*, which was published in September 2011 in *Nature*. Based on the finding that susceptibility loci increasing the risk for the development of chronic complex diseases that are identified in the hundreds by GWAS typically exert only small effects and that information underlying the physiological mechanisms is frequently lacking, the authors concluded that associations with metabolic traits as functional intermediates – by definition the aforementioned mQTLs – can overcome these problems, potentially informing individualized therapy. Accordingly, their study described a comprehensive GWAS of genotype-dependent metabolic phenotypes using nontargeted metabolomics to analyze a panel of small molecules encompassing 250 metabolites from 60 biochemical pathways in serum samples from 2820 individuals from two large population-based European cohorts, the German KORA study ($n=1768$) and the British TwinsUK study ($n=1052$). Consequently, this represents the largest GWAS on metabolite traits performed until today, exceeding that of Illig *et al.* (2010), which included 2231 individuals and 163 measured metabolites.

Genotyping of the KORA participants was performed using the Affymetrix GeneChip array 6.0, whereas the probands of the TwinsUK study were genome-wide individually genotyped with a combination of Illumina arrays (HumanHap300, HumanHap610Q, 1M-Duo and 1.2MDuo 1M). The metabolome analysis was performed using UHPLC and GC separation, coupled with MS/MS. With 24 min per sample and a low median process variability of $<12\%$, the achieved profiling was very efficient. As discussed earlier, metabolite concentration ratios can strengthen association signals and may provide valuable information concerning possible involved metabolic pathways. Therefore, all ratios were calculated and included in the statistical analysis. This finally resulted in more than 37 000 metabolic traits as defined by concentrations and concentration ratios that had to be associated with around 600 000 SNPs. In order to reduce the required computational as well as data storage burden, the analysis started with an initial screening stage for selecting promising signals, where the association was calculated by fitting linear models separately in both cohorts to log-transformed metabolic traits. Adjustment was carried out for

age, gender, and family structure. Subsequently, all signals that exhibited suggestive evidence of association with a metabolic trait in both cohorts were selected, which meant that the *P* values had to be $<10^{-6}$ in both cohorts or $<10^{-3}$ in one and $<10^{-6}$ in the other. Reassessment of the association signals for these loci was performed through fixed-effects inverse variance meta-analysis of both cohorts for all 37 000 metabolic traits using imputed SNPs relative to HapMap2. For each locus, those combinations of SNPs and metabolic traits that produced the smallest *P* values in this meta-analysis were finally chosen. The calculated threshold *P* value for genome-wide significance that resulted after application of a conservative Bonferroni correction amounted to $<2\times10^{-12}$.

This analysis resulted in the identification of 37 genetic loci significantly associated with blood metabolite concentrations at a stringent genome-wide threshold. Among these, 23 represented newly identified GDMs, whereas 14 replicated already known associations. At 30 loci, the lead SNP mapped to a gene encoding a protein that is obviously biochemically linked to the associating metabolites, for instance is involved in their synthesis, degradation, or metabolism. Altogether, 25 loci exhibited effect sizes unusually high for GWAS that accounted for 10–60% of differences in metabolite levels per allele copy.

Subsequently, Suhre *et al.* performed an extensive literature and database search to identify those among these 37 loci that were previously reported as being associated with a clinical endpoint, a medically relevant intermediate phenotype, or a pharmacogenetic effect. Based on the respective lead SNP or a proxy in pronounced linkage disequilibrium with published disease-associated SNPs, such relationships were identified in 15 cases including CVD, kidney disease, Crohn's disease, gout, cancer, adverse reactions to drug therapy, and predisposing risk factors for T2D and CVD. With the exception of three loci, all SNPs represented common SNPs with an MAF $>10\%$.

As one example, the *N*-acetylation of compounds represents an important known mechanism to detoxify nephrotoxic medications as well as environmental toxins. A reduced detoxification capacity for such substances could result in impaired kidney function. A key GDM in this context is the *N*-acetyltransferase 8 encoding *NAT8* locus previously reported to associate with kidney function, as already mentioned in the discussion of the work of Nicholson *et al.* (Juhanson *et al.* 2008, Chambers *et al.* 2010, Köttgen *et al.* 2010, Nicholson *et al.* 2011). A highly significant association of variation at the *NAT8* locus with *N*-acetylornithine was detected by Suhre *et al.*, who subsequently investigated whether *N*-acetylornithine concentrations were associated with kidney function. In both cohorts, a clear association with estimated glomerular filtration rate (eGFR) was found where higher levels of *N*-acetylornithine were correlated with lower eGFR ($p_{KORA}=7{\cdot}6\times10^{-4}$, $p_{TwinsUK}=3{\cdot}6\times10^{-8}$). The identified risk allele was associated with higher *N*-acetylornithine concentrations. As enhanced *N*-acetylation of ornithine mediated by a variant of the *NAT8* gene product is obviously involved in the causative etiology of CKD, this finding could serve as a starting point for future follow-up analyses.

Similarly, six loci identified by Suhre *et al.* as mQTLs were previously described to affect the risk of CAD (Schunkert *et al.* 2011). These loci, namely *ABO*, *NAT2*, *CPS1*, *NAT8*, *ALPL*, and *KLKB1*, exhibited in part only weak evidence for association with CAD ($P<0{\cdot}01$). However, a putative role of the identified metabolic traits in the context of heart disease could be plausibly derived from their biochemical properties. For example, *NAT8*, which encodes the aforementioned *N*-acetyltransferase 8, might link CAD to CKD via ornithine acetylation. Furthermore, Kallikrein B plasma (Fletcher factor) 1 encoded by *KLKB1* is known to be involved in the bradykinin-mediated blood pressure regulation. Suhre *et al.* were now able to demonstrate an association between bradykinin concentrations and a genetic variant in *KLKB1*. In addition, they confirmed the predicted association between bradykinin and hypertension in both cohorts ($p_{KORA}=1{\cdot}7\times10^{-9}$, $p_{TwinsUK}=0{\cdot}0495$). In the cases of *ABO* encoding the blood group-specific glycosyltransferases and *ALPL* encoding alkaline phosphatase, both loci were associated with fibrinogen A-α phosphorylation. This prompted the authors to speculate that genetically determined differences in fibrinogen A-α phosphorylation and resulting blood coagulation properties may represent the basis of these associations with CAD. Similar results were obtained for diabetes-related traits, venous thromboembolism risk loci, and associations related to the context of pharmacogenomics.

Exemplarily, the protein member 9 of the solute carrier family 16, also known as MCT9, which is encoded by *SLC16A9*, was experimentally verified as a carnitine efflux transporter as one proof-of-principle validation of the new discoveries. The association of SNP rs7094971 in *SLC16A9* with carnitine indicated that this metabolite represents the not yet identified substrate of this predicted monocarboxylic acid transporter. Therefore, the authors tested [^{3}H]-carnitine uptake by *SLC16A9*-expressing *Xenopus* oocytes. It could be clearly demonstrated that *SLC16A9* encodes a pH-independent carnitine efflux transporter, which is possibly responsible for carnitine efflux from absorptive epithelia into the blood. To facilitate future functional studies as well as the clinical interpretation of GWAS results, the authors furthermore established a knowledge-base resource, which is publically available via a web server.

In sum, among all other studies discussed so far, this work most impressively demonstrated how the combination of GWAS and metabolomics to identify GDMs is able to advance the knowledge about the genetic basis of human metabolic individuality and the origins of common diseases, allowing the generation of many new hypotheses for biomedical and pharmaceutical research. Suhre *et al.* noted that by discussing only associations supported by two independent studies at genome-wide significance, they chose a very conservative approach because, based on QQ-plots and coarse assumptions, they estimated that more

than 500 (!) loci with association signals below that conservative threshold may be confirmed as GDMs in more highly powered studies. The authors finally prognosticated that future GWAS, which will combine multiple 'omics' technologies including transcriptomics, proteomics, metabolomics, and even epigenomics in a single study, will likely be the next big step toward a full understanding of the interaction between genetic predispositions and environmental factors in the development of complex chronic diseases, their diagnosis, prevention, and safe and efficient therapy.

Toward a combination of transcriptomics and metabolomics: first attempts and perspectives

Whereas classical GWAS analyze the association of genetic variants – mostly SNPs – with specific phenotypes in terms of measurable continuous variables or in case–control experimental designs, the situation is different and quite more difficult when considering whether such phenotypes are related to gene expression profiles in larger epidemiological cohorts. Here, the essentially linear model of causality that explains the relationship between genetic variation and phenotypic expression only holds true with restrictions: nongenetic sources of variation, like physical activity, diet, medication, or smoking status, might also influence the individual measured transcriptome, not or only partially depending on underlying genetic factors. This is of particular relevance whether the transcriptomes of large numbers of individuals are measured using RNA prepared from an easily accessible tissue as whole blood, as it is frequently done in larger cohort studies. Especially, blood cells are exposed to many nongenetic variables that can be predicted to modify their gene expression patterns; for example, in addition to the aforementioned factors, the individual immune status will strongly influence the measured whole-blood cell transcriptome. Furthermore, it has to be kept in mind that the whole-blood transcriptome represents only an incomplete substitute for corresponding gene expression profiles of other body tissues or organs; for instance, not all genes expressed in liver or brain will also necessarily be active in the different blood cell types and vice versa. Notwithstanding, the whole-blood transcriptome definitively represents a valuable source of information as a first approximation of an individual's specific 'global' gene expression pattern, reflecting not only the impact of genetic polymorphisms but also that of nongenetic, in the broadest sense 'lifestyle-dependent' factors.

Still, to the best of our knowledge, there are no large, well-powered studies in the field of genetic epidemiology analyzing associations between metabolic traits and genome-wide gene expression patterns. For instance, the GWAS of metabolic traits described earlier that related concentrations of plasma, serum, or urine metabolites or ratios between metabolite concentrations to genetic polymorphisms could, in theory, be complemented by corresponding transcriptome analyses using RNA prepared from whole blood, associating metabolome and transcriptome data. On the other hand, studies associating phenotype data, genome-wide genotyping data, and genome-wide transcriptome data were already published (Dixon *et al.* 2007, Göring *et al.* 2007, Stranger *et al.* 2007, Emilsson *et al.* 2008, Schadt *et al.* 2008, Idaghdour *et al.* 2010, Zeller *et al.* 2010), although still in limited number when compared with classical GWAS studies.

As one example, the study of Zeller *et al.* that was published in 2010 analyzed the transcriptomes of circulating monocytes as key cells involved in immunity-related diseases and atherosclerosis in 1490 unrelated individuals and investigated associations with 675 000 SNPs and ten common cardiovascular risk factors. As one further reason for using monocytes instead of whole blood, these authors stated that the use of a single cell type should reduce the complexity of transcriptome data and may avoid possible biases resulting from the heterogeneous cell type distribution in different samples as in the case when using whole-blood RNA. To minimize the danger of introducing artifacts, fresh samples were collected and processed in a short period of time. Monocytes were obtained from 730 women and 760 men, aged 35–74 years, recruited in the German epidemiological Gutenberg Heart Study (GHS), a community-based project. Individual transcriptomes were measured using the Illumina Human HT-12 expression BeadChips, whereas individual genome-wide genotyping was performed using Affymetrix 6.0 arrays, resulting in 675 350 SNPs that were available for association analyses after completed data quality control.

At a study-wise threshold of significance correcting for the number of SNPs and expressions ($P < 5{\cdot}78 \times 10^{-12}$), 37 403 associations, involving 29 912 SNPs and 2745 expression traits, were identified. Out of 12 808 genes exhibiting significant expression in monocytes, the authors detected 2745 loci where genetic variation influenced gene-specific mRNA amounts, representing so-called expression quantitative trait loci (eQTLs). The majority of these eQTLs (90%) exerted their effects in *cis*, which means that the polymorphisms were associated with different mRNA amounts of directly neighboring genes, located within 1 Mb of either their 5′- or 3′-ends.

Of note, extensive analyses demonstrated that associations identified by former GWAS of lipids, BMI, or blood pressure were rarely compatible with the idea of causative modulation of the respective phenotype by the monocyte gene expression level at the locus. In the case of CAD, the strongest association identified by GWAS involves SNPs in the *9p21* region (Schunkert *et al.* 2008). It has been reported that in mice, deletion of the region that is orthologous to the human *9p21* CAD locus affects the expression of the neighboring genes *CDNK2A* and *CDNK2B*, both encoding cyclin-dependent kinase inhibitor proteins, as well as the proliferation properties of vascular cells (Visel *et al.* 2010). Because Zeller *et al.* did not detect *CDKN2A* expression in monocytes, they focused their interest on *CDKN2B* and tested all SNPs available in the GHS cohort and encompassing the CAD locus for association

with *CDKN2B* expression. It turned out that *CDKN2B* expression was indeed strongly associated with several SNPs located in a region upstream of the gene; however, these SNPs were not associated with CAD, whereas proxies of the CAD-associated SNPs were not associated with *CDKN2B* expression. Remarkably, the SNPs associated with *CDKN2B* expression are located within the sequence of the noncoding alternatively spliced gene *ANRIL* (also named *CDKN2BAS*) whose implication in the association with CAD has been hypothesized (Jarinova *et al*. 2009).

Zeller *et al*. furthermore demonstrated significant association of altogether 1662 expression traits (13·0%) with at least one of the risk factors age, gender, BMI, HDL- and LDL-cholesterol, triglycerides, systolic and diastolic blood pressure, smoking, and plasma C-reactive protein (CRP) at a study-wide significance level ($P<3\times10^{-7}$). Gender and age were the two major factors influencing expression levels, but BMI, smoking, and CRP levels were also correlated with numerous expression traits, whereas only few associations with blood pressure and lipids were observed. Furthermore, genome-wide interaction analyses suggested that genetic variability and risk factors mostly acted additively on gene expression, where the structure of correlation among expression traits suggested that the variability of risk factors could be characterized by a limited set of independent gene expressions, putatively of biological and clinical relevance. Expression traits associated with cigarette smoking were, for instance, more strongly associated with carotid atherosclerosis than smoking itself.

The study of Zeller *et al*. exemplarily demonstrates the interplay of genetic and nongenetic factors in shaping individual transcriptomes, in this case shown on the example of the monocyte fraction prepared from whole blood. It can be predicted that the number of similar studies will rise continuously as more and more epidemiological cohort studies start to incorporate expression profiling approaches routinely in their standard data collection programs. Parallelizing the development in the GWAS field, this represents a direct consequence of the sinking costs that have to be raised for the underlying array-based expression profiling technology. Therefore, it is expected that in the near future, the connection of GWAS, genome-wide expression profiling, and genome-wide metabolome profiling will be accomplished.

Declaration of interest

The authors declare that there is no conflict of interest that could be perceived as prejudicing the impartiality of the research reported.

Funding

This research did not receive any specific grant from any funding agency in the public, commercial or not-for-profit sector.

Acknowledgements

The authors are grateful to Michael Lalk for critical reading of the manuscript and his helpful comments.

References

Adamski J 2012 Genome-wide association studies with metabolomics. *Genome Medicine* **4** 34. (doi:10.1186/gm333)

Brookes KJ, Chen W, Xu X, Taylor E & Asherson P 2006 Association of fatty acid desaturase genes with attention-deficit/hyperactivity disorder. *Biological Psychiatry* **60** 1053–1061. (doi:10.1016/j.biopsych.2006.04.025)

Chambers JC, Zhang W, Lord GM, van der Harst P, Lawlor DA, Sehmi JS, Gale DP, Wass MN, Ahmadi KR, Bakker SJ *et al*. 2010 Genetic loci influencing kidney function and chronic kidney disease. *Nature Genetics* **42** 373–375. (doi:10.1038/ng.566)

Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, Wong KC, Taylor J, Burnett E, Gut I, Farrall M *et al*. 2007 A genome-wide association study of global gene expression. *Nature Genetics* **39** 1202–1207. (doi:10.1038/ng2109)

Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, Zhu J, Carlson S, Helgason A, Walters GB, Gunnarsdottir S *et al*. 2008 Genetics of gene expression and its effect on disease. *Nature* **452** 423–428. (doi:10.1038/nature06758)

Evan AP, Coe FL, Lingeman JE, Shao Y, Matlaga BR, Kim SC, Bledsoe SB, Sommer AJ, Grynpas M, Phillips CL *et al*. 2006 Renal crystal deposits and histopathology in patients with cystine stones. *Kidney International* **69** 2227–2235. (doi:10.1038/sj.ki.5000268)

Gieger C, Geistlinger L, Altmaier E, Hrabé de Angelis M, Kronenberg F, Meitinger T, Mewes HW, Wichmann HE, Weinberger KM, Adamski J *et al*. 2008 Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. *PLoS Genetics* **4** e1000282. (doi:10.1371/journal.pgen.1000282)

Göring HH, Curran JE, Johnson MP, Dyer TD, Charlesworth J, Cole SA, Jowett JB, Abraham LJ, Rainwater DL, Comuzzie AG *et al*. 2007 Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nature Genetics* **39** 1208–1216. (doi:10.1038/ng2119)

Hicks AA, Pramstaller PP, Johansson A, Vitart V, Rudan I, Ugocsai P, Aulchenko Y, Franklin CS, Liebisch G, Erdmann J *et al*. 2009 Genetic determinants of circulating sphingolipid concentrations in European populations. *PLoS Genetics* **5** e1000672. (doi:10.1371/journal.pgen.1000672)

Holland WL & Summers SA 2008 Sphingolipids, insulin resistance, and metabolic disease: new insights from *in vivo* manipulation of sphingolipid metabolism. *Endocrine Reviews* **29** 381–402. (doi:10.1210/er.2007-0025)

Idaghdour Y, Czika W, Shianna KV, Lee SH, Visscher PM, Martin HC, Miclaus K, Jadallah SJ, Goldstein DB, Wolfinger RD *et al*. 2010 Geographical genomics of human leukocyte gene expression variation in Southern Morocco. *Nature Genetics* **42** 62–67. (doi:10.1038/ng.495)

Illig T, Gieger C, Zhai G, Römisch-Margl W, Wang-Sattler R, Prehn C, Altmaier E, Kastenmüller G, Kato BS, Mewes HW *et al*. 2010 A genome-wide perspective of genetic variation in human metabolism. *Nature Genetics* **42** 137–141. (doi:10.1038/ng.507)

Jarinova O, Stewart AF, Roberts R, Wells G, Lau P, Naing T, Buerki C, McLean BW, Cook RC, Parker JS *et al*. 2009 Functional analysis of the chromosome *9p21.3* coronary artery disease risk locus. *Arteriosclerosis, Thrombosis, and Vascular Biology* **29** 1671–1677. (doi:10.1161/ATVBAHA.109.189522)

Juhanson P, Kepp K, Org E, Veldre G, Kelgo P, Rosenberg M, Viigimaa M & Laan M 2008 *N*-acetyltransferase 8, a positional candidate for blood pressure and renal regulation: resequencing, association and *in silico* study. *BMC Medical Genetics* **9** 25. (doi:10.1186/1471-2350-9-25)

Kathiresan S, Melander O, Guiducci C, Surti A, Burtt NP, Rieder MJ, Cooper GM, Roos C, Voight BF, Havulinna AS *et al*. 2008 Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nature Genetics* **40** 189–197. (doi:10.1038/ng.75)

Kathiresan S, Willer CJ, Peloso GM, Demissie S, Musunuru K, Schadt EE, Kaplan L, Bennett D, Li Y, Tanaka T *et al.* 2009 Common variants at 30 loci contribute to polygenic dyslipidemia. *Nature Genetics* **41** 56–65. (doi:10.1038/ng.291)

Keinan A & Clark AG 2012 Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* **336** 740–743. (doi:10.1126/science.1217283)

Kim AM, Tingen CM & Woodruff TK 2010 Sex bias in trials and treatment must end. *Nature* **465** 688–689. (doi:10.1038/465688a)

Köttgen A, Pattaro C, Böger CA, Fuchsberger C, Olden M, Glazer NL, Parsa A, Gao X, Yang Q, Smith AV *et al.* 2010 New loci associated with kidney function and chronic kidney disease. *Nature Genetics* **42** 376–384. (doi:10.1038/ng.568)

Mattoo A & Goldfarb DS 2006 Cystinuria. *Seminars in Nephrology* **28** 181–191. (doi:10.1016/j.semnephrol.2008.01.011)

Mittelstrass K, Ried JS, Yu Z, Krumsiek J, Gieger C, Prehn C, Roemisch-Margl W, Polonikov A, Peters A, Theis FJ *et al.* 2011 Discovery of sexual dimorphisms in metabolic and genetic biomarkers. *PLoS Genetics* **7** e1002215. (doi:10.1371/journal.pgen.1002215)

Nicholson G, Rantalainen M, Li JV, Maher AD, Malmodin D, Ahmadi KR, Faber JH, Barrett A, Min JL, Rayner NW *et al.* 2011 A genome-wide metabolic QTL analysis in Europeans implicates two loci shaped by recent positive selection. *PLoS Genetics* **7** e1002270. (doi:10.1371/journal.pgen.1002270)

Pe'er I, Yelensky R, Altshuler D & Daly MJ 2008 Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genetic Epidemiology* **32** 381–385. (doi:10.1002/gepi.20303)

Rodrigues CM, Kren BT, Steer CJ & Setchell KD 1996 Formation of Δ22-bile acids in rats is not gender specific and occurs in the peroxisome. *Journal of Lipid Research* **37** 540–550.

Schadt EE, Molony C, Chudin E, Hao K, Yang X, Lum PY, Kasarskis A, Zhang B, Wang S, Suver C *et al.* 2008 Mapping the genetic architecture of gene expression in human liver. *PLoS Biology* **6** e107. (doi:10.1371/journal.pbio.0060107)

Schunkert H, Götz A, Braund P, McGinnis R, Tregouet DA, Mangino M, Linsel-Nitschke P, Cambien F, Hengstenberg C, Stark K *et al.* 2008 Repeated replication and a prospective meta-analysis of the association between chromosome *9p21.3* and coronary artery disease. *Circulation* **117** 1675–1684. (doi:10.1161/CIRCULATIONAHA.107.730614)

Schunkert H, König IR, Kathiresan S, Reilly MP, Assimes TL, Holm H, Preuss M, Stewart AF, Barbalic M, Gieger C *et al.* 2011 Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nature Genetics* **43** 333–338. (doi:10.1038/ng.784)

Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, Ingle CE, Dunning M, Flicek P, Koller D *et al.* 2007 Population genomics of human gene expression. *Nature Genetics* **39** 1217–1224. (doi:10.1038/ng2142)

Suhre K, Meisinger C, Döring A, Altmaier E, Belcredi P, Gieger C, Chang D, Milburn MV, Gall WE, Weinberger KM *et al.* 2010 Metabolic footprint of diabetes: a multiplatform metabolomics study in an epidemiological setting. *PLoS ONE* **5** e13953. (doi:10.1371/journal.pone.0013953)

Suhre K, Shin SY, Petersen AK, Mohney RP, Meredith D, Wägele B, Altmaier E, CARDIoGRAM , Deloukas P, Erdmann J *et al.* 2011*a* Human metabolic individuality in biomedical and pharmaceutical research. *Nature* **477** 54–60. (doi:10.1038/nature10354)

Suhre K, Wallaschofski H, Raffler J, Friedrich N, Haring R, Michael K, Wasner C, Krebs A, Kronenberg F, Chang D *et al.* 2011*b* A genome-wide association study of metabolic traits in human urine. *Nature Genetics* **43** 565–569. (doi:10.1038/ng.837)

Tanaka T, Shen J, Abecasis GR, Kisialiou A, Ordovas JM, Guralnik JM, Singleton A, Bandinelli S, Cherubini A, Arnett D *et al.* 2009 Genome-wide association study of plasma polyunsaturated fatty acids in the InCHIANTI Study. *PLoS Genetics* **5** e1000338. (doi:10.1371/journal.pgen.1000338)

Visel A, Zhu Y, May D, Afzal V, Gong E, Attanasio C, Blow MJ, Cohen JC, Rubin EM & Pennacchio LA 2010 Targeted deletion of the *9p21* non-coding coronary artery disease risk interval in mice. *Nature* **464** 409–412. (doi:10.1038/nature08801)

Wang-Sattler R, Yu Y, Mittelstrass K, Lattka E, Altmaier E, Gieger C, Ladwig KH, Dahmen N, Weinberger KM, Hao P *et al.* 2008 Metabolic profiling reveals distinct variations linked to nicotine consumption in humans – first results from the KORA study. *PLoS ONE* **3** e3863. (doi:10.1371/journal.pone.0003863)

Wellcome Trust Case Control Consortium, Craddock N, Hurles ME, Cardin N, Pearson RD, Plagnol V, Robson S, Vukcevic D, Barnes C, Conrad DF *et al.* 2010 Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* **464** 713–720. (doi:10.1038/nature08979)

Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, Clarke R, Heath SC, Timpson NJ, Najjar SS, Stringham HM *et al.* 2008 Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nature Genetics* **40** 161–169. (doi:10.1038/ng.76)

Yeboah J, McNamara C, Jiang XC, Tabas I, Herrington DM, Burke GL & Shea S 2010 Association of plasma sphingomyelin levels and incident coronary heart disease events in an adult population: Multi-Ethnic Study of Atherosclerosis. *Arteriosclerosis, Thrombosis, and Vascular Biology* **30** 628–633. (doi:10.1161/ATVBAHA.109.199281)

Yu Z, Kastenmüller G, He Y, Belcredi P, Möller G, Prehn C, Mendes J, Wahl S, Roemisch-Margl W, Ceglarek U *et al.* 2011 Differences between human plasma and serum metabolite profiles. *PLoS ONE* **6** e21230. (doi:10.1371/journal.pone.0021230)

Zeller T, Wild P, Szymczak S, Rotival M, Schillert A, Castagne R, Maouche S, Germain M, Lackner K, Rossmann H *et al.* 2010 Genetics and beyond – the transcriptome of human monocytes and disease susceptibility. *PLoS ONE* **5** e10693. (doi:10.1371/journal.pone.0010693)

Received in final form 6 June 2012
Accepted 10 July 2012
Made available online as an Accepted Preprint
10 July 2012