Chapter 8: Genetics of the human metabolome, what is next?

Harish Dharuri

Ayşe Demirkan

Jan Bert van Klinken

Dennis Owen Mook-Kanamori

Cornelia M. Van Duijn

Peter A.C. ’t Hoen

Ko Willems van Dijk

Biochim Biophys Acta. 2014; 1842(10):1923-1931

Abstract

Increases in throughput and decreases in costs have facilitated large scale metabolomics studies, the simultaneous measurement of large numbers of biochemical components in biological samples. Initial large scale studies focused on biomarker discovery for disease or disease progression and helped to understand biochemical pathways underlying disease. The first population-based studies that combined metabolomics and genome wide association studies (mGWAS) have increased our understanding of the (genetic) regulation of biochemical conversions. Measurements of metabolites as intermediate phenotypes are a potentially very powerful approach to uncover how genetic variation affects disease susceptibility and progression. However, we still face many hurdles in the interpretation of mGWAS data. Due to the composite nature of many metabolites, single enzymes may affect the levels of multiple metabolites and, conversely, levels of single metabolites may be affected by multiple enzymes. Here, we will provide a global review of the current status of mGWAS. We will specifically discuss the application of prior biological knowledge present in databases to the interpretation of mGWAS results and discuss the potential of mathematical models. As the technology continuously improves to detect metabolites and to measure genetic variation, it is clear that comprehensive systems biology based approaches are required to further our insight in the association between genes, metabolites and disease.

Introduction

The “inborn errors of metabolism” as defined by Garrod at the beginning of the twentieth century depict the first clearly recognized examples of specific genetic defects leading to the accumulation of metabolites in body fluids [1]. For example, in alkaptonuria, a genetic defect in the enzyme homogentisate 1,2-dioxygenase leads to the accumulation of homogentisic acid and its oxide alkapton in plasma and urine. Detection of alkapton in urine is relatively simple in that exposure of urine from affected patients to air results in black discoloration that is readily detected by eye. Alkaptonuria is transmitted as a recessive Mendelian trait with near complete penetrance and is an example of a rare metabolic disease caused by rare genetic variants [2].

Changes in plasma metabolites are also pathogenic hallmarks of common metabolic diseases such as type-2 diabetes. The defining metabolic marker for type 2 diabetes is glucose, but hyperglycemia co-occurs with changes in a variety of additional metabolites including amino acids, lipids

and lipoproteins. The high heritability of type 2 diabetes is not explained by rare genetic variants segregating in families, but is thought to be caused by a variety of, and presumably combination of common genetic variants. This paradigm is referred to as “common disease-common variant” hypothesis and is pursued in so-called genome wide association studies (GWAS). In GWAS, genome wide genotyping platforms measure genotypes for hundred thousand to millions of single nucleotide polymorphisms (SNPs) with minor allele frequencies (MAF) generally larger than 0.05 and test each of those SNPs for association with a specific trait [3]. A large number of GWAS have been performed with a variety of both binary traits (e.g. type 2 diabetes) and quantitative traits (e.g. fasting glucose levels). These studies have successfully uncovered genetic variants that contribute to disease risk and also to the variation in quantitative phenotypes [4]. For example, for type-2 diabetes, thus far, more than 60 risk loci have been identified, giving novel insights into the complex pathophysiology of the disease. However, the risk attributed to individual SNPs in the vicinity of even the strongest candidate gene, transcription factor 7-like 2 (*TCF7L2*), are relatively modest (odds ratios of 1.5-1.7) [5]. Moreover, the combined genetic loci discovered to date explain only a small proportion (less than 5%) of the observed heritability of type 2 diabetes. Thus, a significant proportion of the observed heritability remains to be uncovered [6].

Since a large proportion of the SNPs discovered through GWAS are intergenic or lie within the intronic regions of genes, rather than in the protein coding sequences, the genetic basis for the association is often not obvious. It is possible that the SNPs discovered through GWAS are in linkage disequilibrium (LD) with the real causal variant that is not captured by the platform. This hypothesis to uncover “missing heritability” is currently being tested by many labs using next generation deep sequencing approaches to screen the whole genome or whole exome to locate the functional variants. Unfortunately, thus far, these approaches have met with relatively limited success. This lack of success may be associated with our inability to recognize the causative variants among the many detected variants. Alternatively, GWAS hits may constitute expression quantitative trait loci (eQTLs) influencing the expression level of one or more genes nearby (*cis*-eQTLs), or at a distant physical location (*trans*-eQTLs) [7, 8]. Recently, a combination of RNA and genome sequencing has provided in-depth insight into the relation between genetic variation and transcriptome variation and their association with functional variation [9].

Whereas it is often difficult to determine the effect of GWAS-discovered SNPs on nearby or distant genes, it is clear that many different genes and loci are involved in the pathogenesis of complex diseases such as type 2 diabetes. In addition, it is also clear that environmental factors including lifestyle (i.e. diet and physical activity) affect the development of diabetes. Therefore, it may be more appropriate to consider common metabolic disorders such as diabetes as the outcome of a variety and often combination of mild "inborn errors of metabolism" in conjunction with the environment. These mild "inborn errors of metabolism" would be reflected by differences in the concentrations of metabolites in cells and/or body fluids and could provide insight into the "missing heritability". The terms "genetically determined metabotype" (GDM) [10] and "genetically influenced metabotype" (GIM) have been coined for this [11]. GIM has been defined as relatively prevalent genetic variants that lead to substantial modification in the efficiency of metabolic conversions [12]. The combination of GIMs in any given individual determines his metabolic individuality and thus, in combination with environment and lifestyle, the risk for metabolic disorders such as type 2 diabetes.

Metabolomics measurements

The detection of GIMs has been facilitated by technological developments in the field of metabolomics, where it is now possible to simultaneously measure hundreds of metabolites in large sets of biological samples using automated procedures, and at relatively low cost (10s of euros per sample). A variety of metabolomics platforms are available, all having their own characteristics. Generally speaking, the metabolomics techniques can be divided in two types of platforms and two types of approaches. Metabolomics platforms based on mass spectrometry (MS) in general require extensive sample preparation and are used in-line with gas or liquid chromatography (GC-MS and LC-MS). In contrast, nuclear magnetic resonance (NMR) based platforms require relatively limited sample preparation and the samples can be analyzed without prior separation procedures. MS and NMR based platforms can be employed for targeted and/or non-targeted approaches. In a targeted approach, the platform is optimized for detection of a set of predefined metabolites and absolute or relative concentrations are determined using internal standards. In contrast, in a non-targeted approach, the platform is optimized to capture global snapshots of the test and reference samples and reports the differences. To subsequently identify the metabolites underlying the differential signal in the

untargeted approach, additional analyses are required that are frequently challenging. Therefore, metabolomics datasets from a non-targeted approach often contain a large number of 'unknown' compounds. The main characteristic of all metabolomics platforms is that a subset of compounds can be detected based on common chemical properties of these compounds rather than their biological relatedness. No single analytical technique exists that is suitable for the identification and quantification of all endogenous metabolites in a sample.

Excellent reviews on the possibilities and challenges of the different metabolomics platforms and approaches are available [13-15]. In general, NMR spectroscopy is highly reproducible and quantitative. However, NMR spectroscopy is relatively insensitive and metabolite identification relies on specialized and mostly proprietary spectral deconvolution algorithms. These algorithms may not always identify the same metabolites and may not always base the identification of a specific metabolite on the same spectral signal. In contrast, MS based platforms provide highly precise information on metabolite mass from which identity can often be inferred. However, metabolite quantification requires spiked internal standards. Thus, a common challenge in metabolomics on any platform is the reproducibility of reported metabolite levels across different laboratories. In addition to these platform-specific challenges, additional variability may be caused by differences in instrumentation and experimental setup conditions such as sample preparation and extraction method, collection protocols, source material (plasma, serum, urine, etc), but also sample storage conditions and batch effects. These aspects all require careful consideration when replicating observations and pooling metabolomics data for meta-analyses.

Genome wide association studies of metabolomics data

Since metabolomics data are (semi)quantitative, they are suited for metabolomics GWAS (mGWAS), uncovering genetic variants that affect metabolite levels. One of the first studies employed an MS-based platform that could identify and quantify up to 363 metabolites in 284 individuals [10]. The study reported that common SNPs explained up to 12% of the observed variance in metabolite levels. Moreover, the study determined that the explained variance could be dramatically increased by considering ratios of metabolites. This is because analyzing ratios of metabolite concentrations potentially reduces the variation in the dataset when the pair of metabolites is related to the substrate and product of a given enzymatic reaction. Furthermore, where a SNP impacts such a metabolic reaction, consideration

of ratios leads to a dramatic reduction in p-value of association. For example, rs174548, a SNP in an intron of the fatty acid delta-5 desaturase 1 (*FADS1*) gene is associated with a phosphatidylcholine moiety, PC C36:4 (36 denotes the number of carbons in the side chains and 4 denotes the number of double bonds) levels with a p-value of 4.52×10^{-8}, slightly above the genome-wide threshold. However, association of the same SNP with the ratio of PC C36:4 / PC C36:3 has a p-value of 2.4×10^{-22}, a reduction by 14 orders of magnitude. The FADS1 enzyme introduces a double bond in long chain polyunsaturated fatty acids and the moities PC C36:3 and PC C36:4 are related to the substrate and product of this enzymatic reaction.

A consistent theme that has emerged from mGWAS is that significant SNP-metabolite associations point to the underlying biological mechanism. This is in contrast to GWAS of clinical endpoints where unravelling the underlying mechanism is often much more challenging. In addition to FADS1, several other associations have shown that the functional nature of the gene matches with the biochemical characteristics of the associated metabolite. For example, SNPs in the gene *GLS2* (glutamine synthase 2) have been found associated with glutamine [16, 17].This is a biologically plausible association because the enzyme GLS2 catalyses the hydrolysis of glutamine. Furthermore, genome-wide hits with unknown gene function offer an opportunity to infer novel biological mechanism underlying the SNP-metabolite association. For example, as a proof of principle, Suhre et al experimentally investigated the association of the SNP rs7094971 in the solute carrier family 16, member 9 (*SLC16A9*) with carnitine. The study validated that the hitherto uncharacterized protein was indeed a carnitine transporter in *Xenopus* oocytes [17]. This result underscores the utility of mGWAS in uncovering novel functions and identifying candidate genes for further study.

Table 1 provides an overview of published mGWAS, their characteristics and main findings. It is obvious that the number of highly significant associations is overwhelming and that many of these associations have yet to be interpreted in their proper pathophysiological context. The heritability of small metabolites and amino acids has been reported to vary between 23% and 55%. The heritability of lipids and lipoproteins is somewhat higher ranging, respectively, from 48% to 62% and 50% to 76% [16]. A recent report from a community based cohort indicates that for the majority of metabolites, heritability explains > 20 % of inter-individual variation and that variation attributable to heritable factors is greater than that attributable to clinical factors [18]. The non-heritable proportion of the variation in

Table 1. Overview of published mGWAS datasets. Characteristic features of the study are shown in the table. PC: phosphatidylcholine, SM: sphingomyelin, PE: phosphatidylethanolamine, HDL: high-density lipoprotein, LDL: low-density-lipoprotein, VLDL: very-low-density-lipoprotein, TG: triglycerides, PL: phospholipid, LPC: lysophosphatidylcholine, TC: total cholesterol, PUFA: polyunsaturated fatty acid, Fischer's ratio: (valine + leucine + isoleucine)/(phenylalanine + tyrosine), uk: unknown (not-assigned) peak . *not defined, **reported by two different study groups at the same time.

Study	Platform	Metabolites	MAF	Genotypes	Sample	Discovery	Replication	Novel hits	Known hits	Correction	mGWAS P-value
[10] Gieger, 2008	ESI-MS/MS	metabolites and ratios (n=363)	>0.05	Affy GeneChip Human Mapping 500K	serum	KORAF3 (n=284)	none	reported suggestively significant: PC–*FADS1*, SM–*FADS1*, PE–*LIPC*, SM–*LIPC*, SM–*PLEK*, lysine–*PARK2*, SM–*ANKRD30A*, propionylcarnitine/butyrylcarnitine–*SCAD*, lauroylcarnitine/octanoylcarnitine–*MCAD*	none	Bonferroni	1.3×10^{-9}
[69] Tanaka, 2008	GC	n-3 and n-6 fatty acids (n=6)	*	Illumina Infinium HumanHap550	plasma	InCHIANTI (n=1075)	GOLDN (n=1076)	none	*FADS1* and *ELOVL2*	none	1.0×10^{-7}
[70] Hicks, 2009	ESI-MS/MS	sphingolipids and proportions (n=76)	*	Illumina Infinium HumanHap300	plasma, serum	EUROSPAN (n=4110)	none	ceramide–*ATP10D*, SM–*SGPP1*, SM–*LASS4*, ceramide–*LASS4*, ceramide–*SPTLC3*	SM–*FADS1*	Bonferroni (n=76)	1.0×10^{-10}
[71] Chasman, 2009	1H-NMR (LipoProtein-I and II assays) and direct assay	Lipoproteins (n=22)	>0.01	HumanHap300 Duo chips	plasma	WGHS (n=17296)	PROCARDIS (n=200), FHS (n=2700)	small HDL–*PCCB*, TG–*BTNL2*, medium VLDL–*PPP1R3B*, small LDL–*KLF14*, total HDL–*PAH*, large HDL–*DNAH10*, medium HDL–*WIPI1*	apoB–*PCSK9*, VLDL–*ANGPTL3*, apoB–*CELSR2/PSRC1/SPRT1*, medium HDL–*APOA2*, small VLDL–*APOB*, TG–*GCKR*, LDL cholesterol–*ABCG5*, HDL cholesterol–*COBLL1/GRB14*, LDL cholesterol–*HMGCR*, TG–*BTNL2*, TG–*MLXIPL*, medium VLDL–*PPP1R3B*, medium VLDL–*LPL*, small LDL–*TRIB1*, apoA1–*ABCA1*, small VLDL–*ABO*, large HDL–*FADS1-3*, medium VLDL–*APOA1-5*, LDL cholesterol–*HNF1A*, large HDL–*LIPC*, large HDL–*CETP*, apoA1–*LIPG*, LDL cholesterol–*LDLR*, total LDL–*APOC1-APOE*, apoA1–*HNF4A*, small HDL–*PLTP*	none	5.0×10^{-8}
[19] Illig, 2010	Biocrates	metabolites (n=163) and ratios	>0.1	Affymetrix 6.0 GeneChip and Illumina Hap317K	serum	KORAF4 (n=1029)	KORA F4 (n=780),female TWINSUK (422)-3 steps design	reported suggestive loci : glycine/PC–*CPS1*, valine/Isovalerylcarnitine–*SLC22A4*, PC and PC ratios–*SYNE2(SGPP1)* and significant loci : PC and PC ratios–*FADS1*, PC and PC ratios–*ELOVL2*, carnitine ratios–*SCAD(ACADS)*, carnitine	none	Bonferroni*	3.6×10^{-12}

								ratios–*MCAD(ACADM)*, carnitine ratios–*ACADL(LCAD)*, PC and PC ratios *PLEKHH1*, SM and SM ratios–*SPTLC3*, carnitine ratios–*ETFHD*, carnitine–*SLC16A9*			
[17] Suhre, 2011	Metabolon (UHPLC/MS/MS 2 and GC-MS)	>250 metabolites and ratios yielding (n=37,000)	>0.01	Affymetrix 6.0 GeneChip (discovery), HumanHap300, HumanHap610Q, 1M-Duo and 1.2MDuo 1M(replication) followed by HapMap2 imputations	serum	KORAF4 (n=1768)	TWINSUK (N=1052)	5-oxoproline–*OPLAH*, myristate/myritsoleate–*SCD*, androsterone sulphate–*CYP3A4*, 10-nonadeceonate/ 10-undeceonate–*CYP4A*, glycine–*CPS1*, succinylcarnitine–*LACTB*, isobutyrlcarnitine–*SLC22A1*, serine–*PHGDH*, fibrinogen cleavage peptide ratio–*ENPEP*, andosterone sulphate/epiandrosterone sulphate–*AKR1C*, inosine–*NT5E*, proline–*PRODH*, alpha-hydroxyisovalerate–*HPS5*, fibrinogen cleavage peptine ratio–*ALPL*, bradykinin–*KLKB1*, glutamine–*GLS2*, cafeine/quinate–*AHR*, lactate/isovelarylcarnitine–*IVD*, decanoylcarnitine–*ETFDH*, glutaorylcarnitine/lysine–*SLC7A9*, docosahexaenoic acid/eicosapentaenoic acid–*ELOVL2*, carnitine–*SLC16A9*, isoleucine/tyrosine–*SLC16A10*	butyrylcarnitine/propinylcarnitine–*ACADS*, N-acetylornithine–*NAT8*, PC and PC ratios–*FADS1*, bilirubin/oleolylcarnitine–*UGT1A*, hexaonlycarnitine/oleate–*ACADM*, glucose/mannose–*GCKR*, methylxanthine/4-acetamidobutanoate–*NAT2*, fibrinogen cleavage peptide ratio–*ABO*, urate–*SLC2A9*, eicosenoate/tetradecanedionate–*SLC01B1*, fibrinogen cleavage peptide ratio–*FUT2*, aspartyl-phenyalanine–*ACE*, isovalerylcarnitine–*SLC22A4*, LPC and LPC ratios–*PDXDC1*	Bonferroni (n=37,000)	2.0×10^{-12}
[72]Nicholson, 2011	1H-NMR and Biocrates (FIA-MS)	526 redundant NMR peaks and Biocrates metabolites (n=163)	>0.01	Hapmap 2	Plasma, urine	MolTWIN (n=142) females, longitudinal	MolOBB (n=69)	trimethylamine–*PYROXD2*, dimethylamine–*PYROXD2*, N-acetylated compound–*NAT8*	3-aminoisobutyrate–*AGXT2***		
[73] Suhre, 2011	1H-NMR followed by CHENOMX NMR suit 6.1	56 metabolites and 1661 ratios	*	Affymetrix Human SNP array 6.0	urine	SHIP (males, n=862)	SHIP(n=1032), Longitudinal sample(n=170), KORAF4 (n=992)	formate/succinate–*NAT2*, lysine/valine–*SLC7A9*	3-aminoisobutyrate–*AGXT2***, 2-hydroxyisobutyrate–*WDR66*, alanine/dimethylglycine–*SLC6A20*	Bonferroni (n=1720)	4.5×10^{-11}
[74] Lemaitre 2011	GC	n-3 fatty acids (n=4)	>0.01	HapMap 2	plasma	CHARGE (n=8866)	none	docosapentanoic acid–*GCKR*	alpha-linolenic acid–*FADS1-2*, eicosapentaenoic acid–*FADS1-2*, docosapentanoic acid–*FADS1-2*, eicosapentaenoic acid, docosapentanoic acid, docosahexaenoic acid–*ELOVL2*	none	5.0×10^{-8}

[20] Demirkan, 2012	ESI-MS/MS	Sphingolipids, Phospholipids, (n=115) and their proportions	none	Hapmap 2	plasma, serum	EUROSPAN (n=4034)	none	PE–*PAQR9*, PC–*AGPAT1*, LPC–*PKD2L1*, SM–*PLD2*, SM–*APOE*	PC–*GCKR*, PC–*ELOVL2*, PC–*FADS1-2*, PC–*APOA5*, PC–*PLEKHH1*, PE–*LIPC*, ceramide–*ATP10D*, SM–*SGPP1*, SM–*LASS4*, SM–*SPTLC3*, LPC–*PDXDC1*	Bonferroni (n=23 independent vectors)	2.2×10^{-9}
[16] Kettunen, 2012	1H-NMR, LIPO and LMWM extraction windows	Lipoproteins, lipids, small metabolites (n=117) and 99 selected ratios.	*	7.7 M imputed SNPs from a custom made reference set of HapMap 3 and 1000G	serum	NFBC1966, YF, HBCS, GenMets, DILGOM, Twins (n=8330)	none	alanine/valine–*SLC1A4*, Fischer's ratio–*PPM1K*, phenylalanine–*F12*, glutamine/histidine–*DHDPSL*, phenylalanine/tyrosine–*TAT*, Fischer's ratio–*SLC2A4*, citrate–*SLC25A1*, x-large HDL–*FCGR2B*, albumin–*ALB*, xx-large VLDL–*PPP1R11*, linoleic acid/PUFA–*CPT1A*.	alanine/glutamine–*GCKR*, glucose–*G6PC2*, histidine/valine–*KLKB1*, alanine/tyrosine–*SLC16A10*, glucose–*MTNR1B*, glutamine/glucose–*GLS2*, large LDL free cholesterol–*PCSK9*, mobile lipids–*ANGPTL3*, VLDL diameter–*MLXIPL*, medium VLDL PL–*LPL*, TC/esterified cholesterol–*ABCA1*, linoleic acid/PUFA–*FADS1*, valine/TG–*APOA1*, x-large HDL TG–*LPL*, linoleic acid/PUFA–*PDXDC1*, HDL cholesterol–*CETP*, x-large HDL TG–*LIPG*, medium LDL cholesterol/medium LDL PL–*LDLR*, large LDL free cholesterol–*APOE*, large HDL lipid content/medium HDL lipid content–*PLTP*	Bonferroni (n=216)	2.3×10^{-10}
[75] Wu, 2013	GC	fatty acids (n=4)	>0.01	HapMap 2	plasma	CHARGE Consortium (n=8961)	none	palmitic acid, stearic acid–*ALG14*, palmitoleic acid , oleic acid and stearic acid–*FADS1*, stearic acid–*LPAGT1*, palmitoleic acid–*GCKR*, palmitoleic acid –*PKD2L1* and a locus on chromosome 2.	none	none	5.0×10^{-8}
[76] Rueedi, 2014	1H-NMR	redundant NMR features(n=1276)	*	HapMap 2 (discovery) Illumina OmniQuad (replication)	urine	Colaus (n=835)	TasteSensomics (n=601)	uk–*PYROXD2*, fucose–*FUT*	N-acetylated compounds–*ALMS1(NAT8)*, uk–*ACADL*, 3-aminoisobutyrate–*AGXT2*, uk–*NAT2*, uk–*ABO*, trimethylamine–*PYROXD2*, uk–*PYROXD2*, uk–*ACADS*, 2-hydroxyisobutyrate–*PSDM9*, lysine–*SLC7A9*	Bonferroni (n=125)	5.7×10^{-10}

metabolite levels is likely due to factors such as age, gender, menopause, medication, smoking, nutrition and underlying diseases. The relative contribution and interplay of each of these factors requires larger mGWAS and modeling of gene x environment interactions.

Challenges associated with mGWAS

Metabolomics platforms generally yield information on the levels of one to several hundreds of metabolites. Consideration of all metabolites results in a severe multiple testing burden. This precludes genuine SNP-metabolite pairs from being considered when they fail to reach the stringent statistical threshold for significance. This problem is further exacerbated when considering metabolite ratios. The p-value threshold for a single outcome GWAS is determined by the number of independent genomic loci. Due to the intricate LD structure of the human genome, this p-value is typically set at $p < 5 \times 10\text{-}8$. Similar to SNPs in LD, a significant proportion of the metabolites are highly correlated to other often similar metabolites and cannot be considered as independent. To account for multiple test correction, some groups have computed the Bonferroni correction by counting all the metabolites [10, 17, 19], while a few other groups have adopted a less stringent strategy by taking into account the number of independent metabolites as determined by a principle component analysis [20]. A standardized approach to deal with the multiple testing issue in mGWAS remains to be formulated. Another issue relates to the reporting of novel hits. In conventional GWAS, a hit for a specific phenotype is novel if it is independent from previously reported SNPs that are associated with the phenotype. In mGWAS, some of the hits associate with closely related yet non-identical metabolites/phenotypes. In these cases, the association but not the SNP is novel.The mGWAS that have been reported so far followed the classical GWAS approach to uncover genetic variants affecting metabolites and metabolite ratios. The selection of metabolite ratios for GWAS has been done based on selected prior knowledge or simply by analyzing all possible combinations. For example, Illig et al. analyzed the whole ratio matrix of 163 metabolites quantified by a commercial targeted array designed to capture a selection of sugars, amino-acids, acyl-carnitines and phospholipids. Despite the burden of multiple testing inherent to this approach, they still were able to capture associations below the significance threshold, particularly for the *FADS* locus [19]. This is likely due to the fact that both the substrate and the product of the FADS enzymes were present in the platform, which may not be always the case for other metabolites and enzymes. Our group followed a

similar approach and performed an mGWAS for phospholipids and sphingolipids. To decrease the burden of multiple testing we used the proportion of each metabolite within its own class, in addition to its absolute concentration and reported additional 6 new loci for these molecules [20]. However, these unbiased but naive approaches seem insufficient to fully exploit the data generated by the metabolomics platforms. Although increasing the sample size will reveal novel genes affecting metabolite levels, additional novel approaches that utilize knowledge of biological relatedness between the molecules are required to take mGWAS one step further.

Various genes that have been identified thus far to affect metabolite levels have also been identified in GWAS of conventional metabolic traits, such as glucose and total plasma lipids. For example, variation in the FADS gene cluster is associated with the fatty acid composition of phospholipids, but also fasting glucose levels, triglycerides and total cholesterol (table 1 and [21]). In addition, the FADS gene cluster has also been associated with the intermediary outcome intima media thickness [20]. These data are in agreement with the notion that phospholipids are somehow causally involved in one of the first steps leading to disturbances in glucose and/or lipid metabolism and subsequent cardio-metabolic disease. However, numerous hypotheses can be formulated to link phospholipids with cardio-metabolic disease. These hypotheses include changes in cellular membrane properties and thus receptor function, but also changes in lipoprotein surface properties and function. Whether any or all of these potential mechanisms play a role in the link between the FADS gene cluster and disease remains to be determined and experimentally validated. However, detailed insight into the specific pathways that are affected by variation in phospholipids is a required first step to select the most likely hypotheses.

Pathway analysis of mGWAS data

Pathway analysis is exquisitely suited to increase the statistical power to identify biologically plausible loci and simultaneously improve our understanding of the underlying biological mechanisms. Pathway-based approaches examine test statistics for a group of genes in contrast to single-marker analysis. The ‘group of genes’ is an expert defined set that is functionally related to the phenotype. The utility of this technique to identify novel and biologically meaningful loci has already been shown in GWAS with clinical endpoints [22-25]. Furthermore, pathway based approaches are uniquely suited to mGWAS owing to the abundance of knowledge on

proteins involved in metabolite conversion and secretion, as captured in various databases of metabolic pathways and reactions.

The term 'pathway' in a pathway analysis is usually referring to a set of functionally related genes participating in a common biological process. The resources of prior knowledge that are commonly used in pathway analysis include controlled vocabularies like Gene Ontology [26], manually curated gene sets from MSigDB [27] and the pathway databases like KEGG[28], BioCyc[29] and REACTOME [30]. Metabolic pathways offer the ideal knowledge resource for pathway analysis in mGWAS due to the direct relationship between entities represented in these databases and compounds measured on metabolomics platforms.

Metabolic pathways consist of three tiers of information: 1) metabolites at the lowest level; 2) reactions built from metabolites and the enzymes that drive these reactions; and 3) pathways built upon reactions [31]. Pathway databases like KEGG, BioCyc and Reactome have extended our knowledge of human metabolism. However, no single database captures all relevant biochemical knowledge and conceptual differences between the databases pose a serious challenge to knowledge integration efforts [31, 32]. For example, a study [33] published in 2011 found that the consensus among five major pathway databases at the level of the genes is 13%, at the level of enzyme commission (EC) numbers is 18%, at the level of metabolites is 9% and at the level of the reactions is merely 3%. The lack of consensus in metabolite specific databases extends to resources like HMDB [34] and ChEBI[35] due to differing representation of common metabolites and reactions. Three recent efforts namely BKM-react [36], MetRxn [37], and MNXref [32] attempt to automate the reconciliation of metabolite and reaction information.

Pathway analysis entails selecting a pre-defined set of genes or pathways to test for enrichment. This selection is generally based on the relevance of the test set to the phenotype being assessed by the GWAS. The generation of gene sets relevant to metabolites requires a systematic interrogation of metabolite databases and depends heavily on the accessibility and download formats made available by the database. Furthermore, it is important that the software developed to generate such gene sets is easy to use. To address these issues, we have developed tools to systematically interrogate on-line databases using Taverna [38], a workflow-based management system. Taverna allows users access to remote data resources like KEGG, BioCyc, Ensembl [39] and NCBI [http://www.ncbi.nlm.nih.gov/] and data

management systems like Biomart [40] through implementation of web services. To generate a gene set for each of the metabolites measured on a metabolomics platform, we designed workflows to interrogate pathway databases and retrieve genes from pathways and reactions relevant to the metabolite [41]. A corresponding SNP set (SNPs present in ±25 kb flanking region of the genes) was generated for each of the metabolites. As a proof of principle, we investigated the utility of the reduced and biologically relevant SNP set to identify known and novel association from a published GWA dataset by Illig et al [19]. The smaller SNP set reduced the multiple-testing threshold by around two orders of magnitude. This reduction helped us discover novel SNP-metabolite associations in the Illig et al GWAS datasets [41]. For example, a SNP in the gene *ALDH1L1* (aldehyde dehydrogenase 1 L1) was found associated with the ratio of serine/glycine. The original study missed this association because the p-value cut-off in the discovery stage of the study precluded this association from being considered in the replication stage. ALDH1L1 is an important component of the one-carbon pool pathway and acts upstream of SHMT (serine hydroxy methyl transferase) enzyme that mediates the bulk of glycine to serine conversion in the cell. This reaffirms the notion that a method that relies on background knowledge present in pathway databases has the ability to reduce the multiple test burden and thereby facilitate the discovery of true positives in GWAS results. It should be noted that assignment of SNPs to genes represents a challenge in itself. It is common to include only SNPs in the coding region of the gene or within a certain, more or less arbitrary, distance threshold. However, Hong et al [42] note that the reliable conversion of SNPs to representative genes is not trivial and that positional gene clustering if not corrected for can lead to spurious results in a pathway analysis. Properly accounting for LD structure and knowledge on eQTLs will help to link SNPs to the right genes.

Our pathway analysis approach to alleviate the multiple-testing burden through selective testing of SNPs can be seen as complementary to conventional GWAS analysis. However, pathway analysis can also be used in a post-GWAS setting to identify enriched pathways within the identified significantly associated SNPs. We have reported [20] a pathway analysis designed to identify enriched pathways using web accessible software made available by ConsensusPathDB [43]. The latter is a database that integrates pathways and interaction resources made available by databases like KEGG, BioCyc and Reactome. The study reported the enrichment of the following pathways for phospholipid traits: glycerolipid metabolism, chylomicron-mediated lipid transport, triglyceride biosynthesis and metabolism of lipids

and lipoproteins. The list of enriched pathways functionally matches the traits, thus reinforcing the importance of pathway analysis in such studies.

Pathway analysis approaches for GWAS can be categorized based on the type of input data and the specific null hypothesis that is being tested [44]. With regard to input data, there are two types of approaches; one approach uses SNP p-values and the other approach uses the effect sizes derived from SNP phenotype data (beta's) to calculate pathway-level statistics. With regard to the null hypothesis being tested, two approaches are available: competitive tests and self-contained tests. A competitive test compares the test statistic of a gene set to a standard defined by its complement. In contrast, a self-contained test compares the test statistic of the gene set to a fixed standard and does not take into account genes in other gene sets. The issues and solutions to SNP-to-gene mapping and tests for gene set enrichment are common to all GWAS and we would like to direct the readers to other excellent reviews [44-46].

Gaussian Graphical Modelling

Gaussian Graphical Modelling (GGM) is an unbiased and database independent approach to reconstruct metabolic networks from large-scale metabolomics data sets [47]. GGMs are undirected probabilistic graphical models, in which pairwise correlations between metabolites are conditioned against the correlations with all other metabolites in the dataset. Krumsiek et al. [47] demonstrated that the high partial correlations represent direct interactions and that groups of metabolites that score highly in the correlation matrix can be attributed to reaction steps in known pathways. As indicated earlier, non-targeted metabolomics platforms also quantify many 'unknown' metabolites. This issue was addressed in a recent work by Krumsiek et al [48] who demonstrated that unknown metabolites can be identified by integrating GGMs with mGWAS results. Their method exploits partial correlations between known and unknown metabolites in addition to their association to specific loci in order to generate a hypothesis regarding the identity of the unknown metabolites. Through experimental validation the study provided genetic and biochemical evidence for classification of several unknown metabolites. These studies demonstrate that GGMs in combination with mGWAS could potentially facilitate metabolite classification and also provide a more comprehensive elucidation of enzyme-metabolite relationship and metabolic pathways.

Pleiotropy in mGWAS

Most metabolomics platforms measure numerous metabolites that are highly related and correlated to each other. For example, the Biocrates Absolute IDQ© p150 mass-spectrometry based platform measures up to 163 metabolites belonging to the classes of amino acids, carnitines, and phospholipids. Of the phospholipids, 90 different PCs are measured that only differ based on alkyl/acyl bonds, number of single/double bonds and length of the side chains. Genes that affect the levels or degree of saturation of fatty acids also influence the phospholipid pool. Hence, several loci that participate in fatty acid metabolism associate with multiple phosphatidylcholines [10, 19, 20].

Pleiotropy, the association of a genotype with multiple phenotypes, represents an opportunity to increase the power to identify novel loci and gain insight in metabolic pathways. However, GWAS based on univariate statistical analysis does not take pleiotropy into account. A few groups have developed algorithms and software to exploit the potential of increased statistical power using multivariate statistical analysis [49-54]. Ried et al. [49] developed a method called “Phenotype Set Enrichment Analysis” (PSEA) for the analysis of gene effects on iron-related and blood count traits. The aim of PSEA is to test if a predefined set of phenotypes is associated with a gene. The advantage of such a joint analysis is two-fold: first, the combined analysis of multiple phenotypes can provide insight into the underlying genetic basis and second, it leads to improved statistical power in comparison to association analysis of single phenotypes. PSEA is based on the idea of gene set enrichment analysis for the investigation of phenotype sets. The analysis consists of four steps: i) generate a gene-wise test statistic per phenotype; ii) determine an enrichment score for each combination of phenotype set and gene; iii) a permutation test to determine the enrichment of a phenotype set; and iv) determine the statistical significance of the phenotype set and account for multiple test correction. In another study, Stephens et al. [50] report a unified framework that extensively relies on Bayesian statistics for association analysis of multiple related phenotypes. The utility of the method is illustrated with an application to a genome-wide association study of blood lipid traits from the Global Lipids consortium. To identify novel associations the study applied a two-stage process where in the first stage promising SNPs were identified by applying univariate and multivariate tests to every SNP and in the second stage a Bayesian analysis was performed on the set of promising SNPs. The method could identify 18 potentially novel genetic

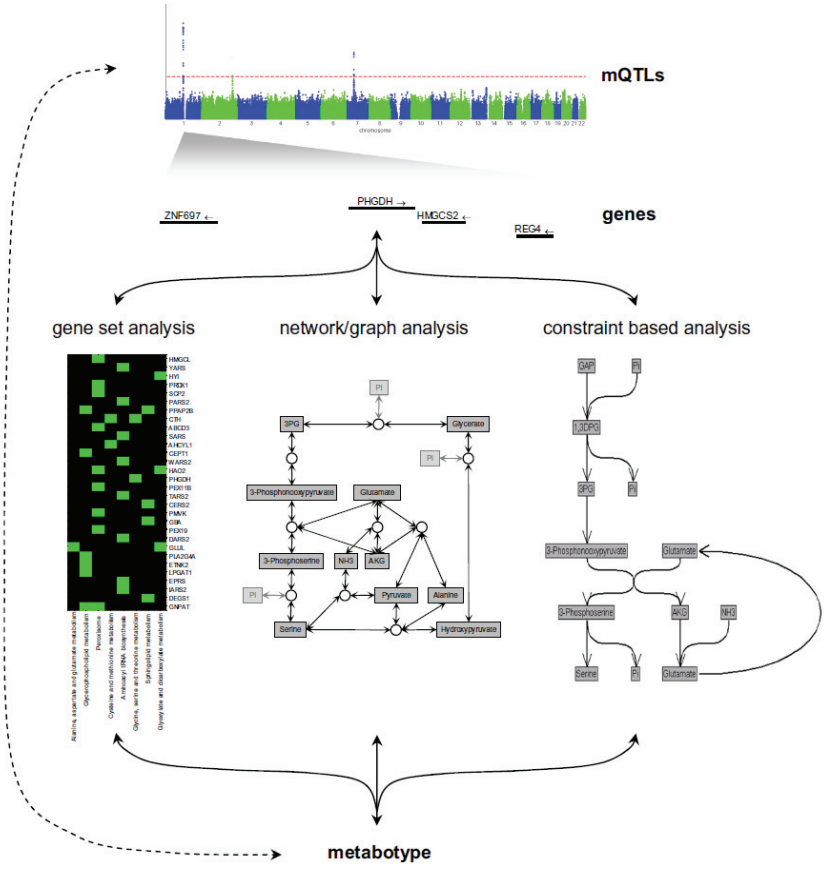
associations that were not identified by the traditional univariate analysis. In general, however, a limitation of multivariate algorithms is that they operate only for a modest number of phenotypes. Inouye et al [54] report a multivariate analysis that utilizes the correlation structure of the 130 metabolites measured on their NMR platform. An unsupervised algorithm is used to identify metabolic networks and in the next step a multivariate test of association for each of the networks with the SNP panel is performed. The authors report 7 new loci using this method. These results indicate that mGWAS analyses profit from a shift from the univariate analysis paradigm to joint modeling of phenotypes to improve the power in identification of novel loci as well as to improve our understanding of the biological function for known loci.

Towards mechanistic models

To completely understand the relations between different metabolites in the various tissues and cell types, it is essential to have a full description of all relevant metabolic reactions and the involved enzymes and transporter proteins. This knowledge can be utilized to develop mathematical models that describe the fluxes through the metabolic system. Furthermore, these models can then be used to predict how fluxes and metabolite levels change as a consequence of genetic variation. This modelling approach is generally referred to as 'bottom-up' systems biology [55].

Recently, in a global research effort several genome-scale metabolic models (GSMMs) have been merged into a consensus model for human metabolism [56]. The key difference between this model and pathway databases is that GSMMs are represented mathematically and have typically undergone additional curation steps that enable mathematical analysis of these models. Most importantly, curation consists of 1) ensuring that all reactions are mass balanced, 2) filling the gaps in the model such that the network is fully connected, and 3) checking that the model is functionally valid, i.e. it has to faithfully predict which metabolic conversions an organism (or tissue) is capable of. A detailed protocol for the reconstruction and curation of organism and tissue-specific GSMMs has recently been described by Thiele and Palsson [57].

Given a fully functional GSMM, its behavior can be analyzed in terms of the space of feasible steady state fluxes through the network, using a set of techniques commonly referred to as constraint-based analysis (CBA) [58-61]. An application of CBA that is of particular interest to metabolic research in



Box 1.

Pathway analysis of mGWAS results. The first step of pathway analysis consists of mapping the locus associated with the metabolite level to a set of candidate genes. Candidate gene selection may be based on LD, vicinity to the associated locus or the fact that the locus affects the expression of a gene at some distance (eQTL). An alternative approach for selecting can- didate genes from mGWAS results is to aggregate the *p*-values of all SNPs that lie close to a gene into a gene-wise *p*-value and subsequently consider significant genes. The next step in- volves integrating the selected genes with knowledge from pathway databases and/or metabolic models. Three separate approaches can be distinguished at this point: (1) gene set analysis, (2) network or graph analysis and (3) constraint based analysis. (1) Gene set analysis employs expert derived gene sets representing biological pathways and processes to de- termine whether certain sets are statistically enriched for the selected genes. (2) Network analysis uses the topology of a biological network to identify enriched submodules. The most commonly used biological networks are Protein–Protein Interaction (PPI) networks and metabolic networks that consist of graphs with edges between metabolites and enzymatically catalyzed reactions (represented by squares and circles, respectively, in the diagram). (3) Constraint based analysis (CBA) provides a set of mathematical techniques that characterize the functional capacity of a metabolic network in terms of the feasible fluxes through the network. In contrast to traditional network analysis, CBA takes into account the steady state and thermodynamic constraints that are imposed by the set of reactions. That is, internal metabolites may not be net produced or consumed and the flux through irreversible reactions must be non-negative [57–61]. CBA requires well curated genome scale models such as developed by Thiele et al. [56]. In the diagram the Manhattan plot of a GWAS on serine levels is shown, focusing on the locus inside the PHGDH gene that was first discovered by Sühre et al. [17]. The enzyme encoded by PHGDH catalyzes the conversion of 3-phospho-D-glycerate (3PG) to 3-phosphonooxypyruvate. Mapping this gene to the pathway gene sets defined in KEGG shows that it occurs in the “glycine, serine and threonine metabolism” pathway (rn00260), which provides a direct link to serine. Using network analysis, several possible paths are found between the reaction catalyzed by PHGDH and serine that could explain the association between gene and metabolite. Finally, CBA gives a more specific result and shows that PHGDH plays a role in serine biosynthesis

humans is to simulate changes in the flux distribution in response to perturbations that reflect pathological or drug treated states. Shlomi et al [62] and Thiele et al [56] have used this method to predict metabolite biomarkers for inborn errors of metabolism. Their approach consisted of predicting the variation in metabolite concentrations and comparing this variation between the healthy case, in which fluxes could pass through the reaction associated with the gene of interest, and the disease case, for which this reaction was blocked. Applying this method on the consensus model of human metabolism, Thiele et al [56] were able to predict directional changes in metabolite biomarkers with an accuracy of 77%. See Box 1 for a comprehensive overview of the different approaches to perform pathway analyses of mGWAS results.

Recently, the use of human GSMMs as a scaffold for the integration and interpretation of omics data has been pioneered by Lewis et al [63], Jerby and Ruppin [64], and Mardinoglu et al [65]. However, GSMMs have not yet been used in the analysis and interpretation mGWAS results. The main advantage of CBA is that it goes beyond traditional methods of pathway analysis where pathways are either represented as pre-defined gene sets or as reaction chains that follow from graph-based searches. Therefore, its application to the mGWAS setting has great potential for providing true mechanistic insight into the links between genetic loci and metabolic phenotypes and constitutes a promising direction for future research. Ultimately integration of GSMMs with genetic data and expression and clinical phenotypes will help unravel disease patho-physiology and identify optimal individualized treatment strategies [66, 67].

Conclusions

The first waves of metabolomics and genetic analyses by mGWAS have provided a wealth of insight into the genetic basis of metabolic individuality and risk factors for common metabolic disorders, even with modest sample sizes and conventional and conservative statistical approaches. However, true understanding of the interrelation between common metabolic disorders, metabolites and genetic variation requires in depth insight into the associated pathways and their regulation. One approach to gain this insight is to mine available pathway databases using statistical tools and this approach has already proven its value in mGWAS. The next step in pathway analyses is to include stoichiometric and kinetic parameters and complement the statistical analyses with a more comprehensive systems biology bases approach using mathematical modelling. GSMMs are a first step towards that

direction, but thus far lack quantitative information. Inclusion of quantitative data on the regulation of enzyme activity and reaction kinetics will be vital for developing more accurate predictive models [68]. The combined efforts of numerous research groups around the world to address these issues will pave the way for the application of comprehensive systems biology based approaches to gain insight into the genetics of the human metabolome and especially its relation to disease.

Acknowledgements

The authors are receiving funds from the European Community's Seventh Framework Programme (FP7/2007–2013) 202272 ENGAGE (201413), the Centre for Medical Systems Biology (CMSB) and the Netherlands Consortium for Systems Biology (NCSB), both within the framework of the Netherlands Genomics Initiative (NGI)/Netherlands Organisation for Scientific Research (NWO) and the European Commission Seventh Framework Programme Wf4Ever (Digital Libraries and Digital Preservation area ICT-2009.4.1 project reference 270192). DOMK is supported by a personal VENI-grant fromNWO (016.146.023).

References

1. Garrod SAE: **Inborn errors of metabolism**; 1909.
2. Scriver CR: **Garrod's Croonian Lectures (1908) and the charter 'Inborn Errors of Metabolism': albinism, alkaptonuria, cystinuria, and pentosuria at age 100 in 2008**. *J Inherit Metab Dis* 2008, **31**(5):580-598.
3. Siva N: **1000 Genomes project**. *Nature biotechnology* 2008, **26**(3):256.
4. Stranger BE, Stahl EA, Raj T: **Progress and promise of genome-wide association studies for human complex trait genetics**. *Genetics* 2011, **187**(2):367-383.
5. Groves CJ, Zeggini E, Minton J, Frayling TM, Weedon MN, Rayner NW, Hitman GA, Walker M, Wiltshire S, Hattersley AT *et al*: **Association analysis of 6,736 U.K. subjects provides replication and confirms TCF7L2 as a type 2 diabetes susceptibility gene with a substantial effect on individual risk**. *Diabetes* 2006, **55**(9):2640-2644.

6. Hara K, Fujita H, Johnson TA, Yamauchi T, Yasuda K, Horikoshi M, Peng C, Hu C, Ma RC, Imamura M *et al*: **Genome-wide association study identifies three novel loci for type 2 diabetes**. *Hum Mol Genet* 2013.

7. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ: **Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS**. *PLoS genetics* 2010, **6**(4):e1000888.

8. Zhernakova DV, de Klerk E, Westra HJ, Mastrokolias A, Amini S, Ariyurek Y, Jansen R, Penninx BW, Hottenga JJ, Willemsen G *et al*: **DeepSAGE reveals genetic variants associated with alternative polyadenylation and expression of coding and non-coding transcripts**. *PLoS genetics* 2013, **9**(6):e1003594.

9. Lappalainen T, Sammeth M, Friedlander MR, t Hoen PA, Monlong J, Rivas MA, Gonzalez-Porta M, Kurbatova N, Griebel T, Ferreira PG *et al*: **Transcriptome and genome sequencing uncovers functional variation in humans**. *Nature* 2013, **501**(7468):506-511.

10. Gieger C, Geistlinger L, Altmaier E, Hrabe de Angelis M, Kronenberg F, Meitinger T, Mewes HW, Wichmann HE, Weinberger KM, Adamski J *et al*: **Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum**. *PLoS genetics* 2008, **4**(11):e1000282.

11. Suhre K, Gieger C: **Genetic variation in metabolic phenotypes: study designs and applications**. *Nature reviews Genetics* 2012, **13**(11):759-769.

12. Adamski J, Suhre K: **Metabolomics platforms for genome wide association studies--linking the genome to the metabolome**. *Curr Opin Biotechnol* 2013, **24**(1):39-47.

13. Serkova NJ, Standiford TJ, Stringer KA: **The emerging field of quantitative blood metabolomics for biomarker discovery in critical illnesses**. *Am J Respir Crit Care Med* 2011, **184**(6):647-655.

14. Kell DB: **Systems biology, metabolic modelling and metabolomics in drug discovery and development**. *Drug Discov Today* 2006, **11**(23-24):1085-1092.

15. Griffiths WJ, Koal T, Wang Y, Kohl M, Enot DP, Deigner HP: **Targeted metabolomics for biomarker discovery**. *Angew Chem Int Ed Engl* 2010, **49**(32):5426-5445.

16. Kettunen J, Tukiainen T, Sarin AP, Ortega-Alonso A, Tikkanen E, Lyytikainen LP, Kangas AJ, Soininen P, Wurtz P, Silander K *et al*: **Genome-wide association study identifies multiple loci influencing human serum metabolite levels**. *Nature genetics* 2012, **44**(3):269-276.

17. Suhre K, Shin SY, Petersen AK, Mohney RP, Meredith D, Wagele B, Altmaier E, CardioGram, Deloukas P, Erdmann J *et al*: **Human metabolic individuality in biomedical and pharmaceutical research**. *Nature* 2011, **477**(7362):54-60.

18. Rhee EP, Ho JE, Chen MH, Shen D, Cheng S, Larson MG, Ghorbani A, Shi X, Helenius IT, O'Donnell CJ *et al*: **A genome-wide association study of the human metabolome in a community-based cohort**. *Cell Metab* 2013, **18**(1):130-143.

19. Illig T, Gieger C, Zhai G, Romisch-Margl W, Wang-Sattler R, Prehn C, Altmaier E, Kastenmuller G, Kato BS, Mewes HW *et al*: **A genome-wide perspective of genetic variation in human metabolism**. *Nature genetics* 2010, **42**(2):137-141.

20. Demirkan A, van Duijn CM, Ugocsai P, Isaacs A, Pramstaller PP, Liebisch G, Wilson JF, Johansson A, Rudan I, Aulchenko YS *et al*: **Genome-wide association study identifies novel loci associated with circulating phospho- and sphingolipid concentrations**. *PLoS genetics* 2012, **8**(2):e1002490.

21. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA: **Potential etiologic and functional implications of genome-wide association loci for human diseases and traits**. *Proceedings of the National Academy of Sciences of the United States of America* 2009, **106**(23):9362-9367.

22. Wang K, Zhang H, Kugathasan S, Annese V, Bradfield JP, Russell RK, Sleiman PM, Imielinski M, Glessner J, Hou C *et al*: **Diverse genome-wide association studies associate the IL12/IL23 pathway with Crohn Disease**. *American journal of human genetics* 2009, **84**(3):399-405.

23. Zhong H, Yang X, Kaplan LM, Molony C, Schadt EE: **Integrating pathway analysis and genetics of gene expression for genome-wide association studies**. *American journal of human genetics* 2010, **86**(4):581-591.

24. Torkamani A, Topol EJ, Schork NJ: **Pathway analysis of seven common diseases assessed by genome-wide association**. *Genomics* 2008, **92**(5):265-272.

25. Elbers CC, van Eijk KR, Franke L, Mulder F, van der Schouw YT, Wijmenga C, Onland-Moret NC: **Using genome-wide pathway analysis to unravel the etiology of complex diseases**. *Genetic epidemiology* 2009, **33**(5):419-431.

26. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nature genetics* 2000, **25**(1):25-29.

27. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP: **Molecular signatures database (MSigDB) 3.0**. *Bioinformatics* 2011, **27**(12):1739-1740.

28. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M: **KEGG for integration and interpretation of large-scale molecular data sets**. *Nucleic acids research* 2012, **40**(Database issue):D109-114.

29. Caspi R, Altman T, Dreher K, Fulcher CA, Subhraveti P, Keseler IM, Kothari A, Krummenacker M, Latendresse M, Mueller LA *et al*: **The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases**. *Nucleic acids research* 2012, **40**(Database issue):D742-753.

30. D'Eustachio P: **Pathway databases: making chemical and biological sense of the genomic data flood**. *Chemistry & biology* 2013, **20**(5):629-635.

31. Altman T, Travers M, Kothari A, Caspi R, Karp PD: **A systematic comparison of the MetaCyc and KEGG pathway databases**. *BMC bioinformatics* 2013, **14**:112.

32. Bernard T, Bridge A, Morgat A, Moretti S, Xenarios I, Pagni M: **Reconciliation of metabolites and biochemical reactions for metabolic networks**. *Briefings in bioinformatics* 2012.

33. Stobbe MD, Houten SM, Jansen GA, van Kampen AH, Moerland PD: **Critical assessment of human metabolic pathway databases: a stepping stone for future integration**. *BMC systems biology* 2011, **5**:165.

34. Wishart DS, Knox C, Guo AC, Eisner R, Young N, Gautam B, Hau DD, Psychogios N, Dong E, Bouatra S *et al*: **HMDB: a knowledgebase for the human metabolome**. *Nucleic acids research* 2009, **37**(Database issue):D603-610.

35. Hastings J, de Matos P, Dekker A, Ennis M, Harsha B, Kale N, Muthukrishnan V, Owen G, Turner S, Williams M *et al*: **The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013**. *Nucleic acids research* 2013, **41**(Database issue):D456-463.

36. Lang M, Stelzer M, Schomburg D: **BKM-react, an integrated biochemical reaction database**. *BMC biochemistry* 2011, **12**:42.

37. Kumar A, Suthers PF, Maranas CD: **MetRxn: a knowledgebase of metabolites and reactions spanning metabolic models and databases**. *BMC bioinformatics* 2012, **13**:6.

38. Wolstencroft K, Haines R, Fellows D, Williams A, Withers D, Owen S, Soiland-Reyes S, Dunlop I, Nenadic A, Fisher P *et al*: **The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud**. *Nucleic acids research* 2013, **41**(Web Server issue):W557-561.

39. Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S *et al*: **Ensembl 2013**. *Nucleic acids research* 2013, **41**(Database issue):D48-55.

40. Kasprzyk A: **BioMart: driving a paradigm change in biological data management**. *Database : the journal of biological databases and curation* 2011, **2011**:bar049.

41. Dharuri H, Henneman P, Demirkan A, van Klinken JB, Mook-Kanamori DO, Wang-Sattler R, Gieger C, Adamski J, Hettne K, Roos M *et al*: **Automated workflow-based exploitation of pathway databases provides new insights into genetic associations of metabolite profiles**. *BMC genomics* 2013, **14**:865.

42. Hong MG, Pawitan Y, Magnusson PK, Prince JA: **Strategies and issues in the detection of pathway enrichment in genome-wide association studies**. *Human genetics* 2009, **126**(2):289-301.

43. Kamburov A, Stelzl U, Lehrach H, Herwig R: **The ConsensusPathDB interaction database: 2013 update**. *Nucleic acids research* 2013, **41**(Database issue):D793-800.

44. Wang K, Li M, Hakonarson H: **Analysing biological pathways in genome-wide association studies**. *Nature reviews Genetics* 2010, **11**(12):843-854.

45. Sun YV: **Integration of biological networks and pathways with genetic association studies**. *Human genetics* 2012, **131**(10):1677-1686.

46. Khatri P, Sirota M, Butte AJ: **Ten years of pathway analysis: current approaches and outstanding challenges**. *PLoS computational biology* 2012, **8**(2):e1002375.

47. Krumsiek J, Suhre K, Illig T, Adamski J, Theis FJ: **Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data**. *BMC systems biology* 2011, **5**:21.

48. Krumsiek J, Suhre K, Evans AM, Mitchell MW, Mohney RP, Milburn MV, Wagele B, Romisch-Margl W, Illig T, Adamski J *et al*: **Mining the unknown: a systems approach to metabolite identification combining genetic and metabolic information**. *PLoS genetics* 2012, **8**(10):e1003005.

49. Ried JS, Doring A, Oexle K, Meisinger C, Winkelmann J, Klopp N, Meitinger T, Peters A, Suhre K, Wichmann HE *et al*: **PSEA: Phenotype Set Enrichment Analysis--a new method for analysis of multiple phenotypes**. *Genetic epidemiology* 2012, **36**(3):244-252.

50. Stephens M: **A unified framework for association analysis with multiple related phenotypes**. *PloS one* 2013, **8**(7):e65245.

51. van der Sluis S, Posthuma D, Dolan CV: **TATES: efficient multivariate genotype-phenotype analysis for genome-wide association studies**. *PLoS genetics* 2013, **9**(1):e1003235.

52. O'Reilly PF, Hoggart CJ, Pomyen Y, Calboli FC, Elliott P, Jarvelin MR, Coin LJ: **MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS**. *PloS one* 2012, **7**(5):e34861.

53. Park SH, Lee JY, Kim S: **A methodology for multivariate phenotype-based genome-wide association studies to mine pleiotropic genes**. *BMC systems biology* 2011, **5 Suppl 2**:S13.

54. Inouye M, Ripatti S, Kettunen J, Lyytikainen LP, Oksala N, Laurila PP, Kangas AJ, Soininen P, Savolainen MJ, Viikari J *et al*: **Novel Loci for metabolic networks and multi-tissue expression studies reveal genes for atherosclerosis**. *PLoS genetics* 2012, **8**(8):e1002907.

55. Bruggeman FJ, Westerhoff HV: **The nature of systems biology**. *Trends in microbiology* 2007, **15**(1):45-50.

56. Thiele I, Swainston N, Fleming RM, Hoppe A, Sahoo S, Aurich MK, Haraldsdottir H, Mo ML, Rolfsson O, Stobbe MD *et al*: **A community-**

driven global reconstruction of human metabolism. *Nature biotechnology* 2013, **31**(5):419-425.

57. Thiele I, Palsson BO: **A protocol for generating a high-quality genome-scale metabolic reconstruction**. *Nature protocols* 2010, **5**(1):93-121.

58. Varma A, Palsson BO: **Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110**. *Applied and environmental microbiology* 1994, **60**(10):3724-3731.

59. Edwards JS, Palsson BO: **Metabolic flux balance analysis and the in silico analysis of Escherichia coli K-12 gene deletions**. *BMC bioinformatics* 2000, **1**:1-10.

60. Schuster S, Fell DA, Dandekar T: **A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks**. *Nature biotechnology* 2000, **18**(3):326-332.

61. Schilling CH, Letscher D, Palsson BO: **Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective**. *Journal of theoretical biology* 2000, **203**(3):229-248.

62. Shlomi T, Cabili MN, Ruppin E: **Predicting metabolic biomarkers of human inborn errors of metabolism**. *Molecular systems biology* 2009, **5**:263.

63. Lewis NE, Schramm G, Bordbar A, Schellenberger J, Andersen MP, Cheng JK, Patel N, Yee A, Lewis RA, Eils R *et al*: **Large-scale in silico modeling of metabolic interactions between cell types in the human brain**. *Nature biotechnology* 2010, **28**(12):1279-1285.

64. Jerby L, Ruppin E: **Predicting drug targets and biomarkers of cancer via genome-scale metabolic modeling**. *Clinical cancer research : an official journal of the American Association for Cancer Research* 2012, **18**(20):5572-5584.

65. Mardinoglu A, Agren R, Kampf C, Asplund A, Nookaew I, Jacobson P, Walley AJ, Froguel P, Carlsson LM, Uhlen M *et al*: **Integration of clinical data with a genome-scale metabolic model of the human adipocyte**. *Molecular systems biology* 2013, **9**:649.

66. Mardinoglu A, Nielsen J: **Systems medicine and metabolic modelling**. *Journal of internal medicine* 2012, **271**(2):142-154.

67. Varemo L, Nookaew I, Nielsen J: **Novel insights into obesity and diabetes through genome-scale metabolic modeling**. *Frontiers in physiology* 2013, **4**:92.

68. Jamshidi N, Palsson BO: **Formulating genome-scale kinetic models in the post-genome era**. *Molecular systems biology* 2008, **4**:171.

69. Tanaka T, Shen J, Abecasis GR, Kisialiou A, Ordovas JM, Guralnik JM, Singleton A, Bandinelli S, Cherubini A, Arnett D *et al*: **Genome-wide association study of plasma polyunsaturated fatty acids in the InCHIANTI Study**. *PLoS genetics* 2009, **5**(1):e1000338.

70. Hicks AA, Pramstaller PP, Johansson A, Vitart V, Rudan I, Ugocsai P, Aulchenko Y, Franklin CS, Liebisch G, Erdmann J *et al*: **Genetic determinants of circulating sphingolipid concentrations in European populations**. *PLoS genetics* 2009, **5**(10):e1000672.

71. Chasman DI, Pare G, Mora S, Hopewell JC, Peloso G, Clarke R, Cupples LA, Hamsten A, Kathiresan S, Malarstig A *et al*: **Forty-three loci associated with plasma lipoprotein size, concentration, and cholesterol content in genome-wide analysis**. *PLoS genetics* 2009, **5**(11):e1000730.

72. Nicholson G, Rantalainen M, Li JV, Maher AD, Malmodin D, Ahmadi KR, Faber JH, Barrett A, Min JL, Rayner NW *et al*: **A genome-wide metabolic QTL analysis in Europeans implicates two loci shaped by recent positive selection**. *PLoS genetics* 2011, **7**(9):e1002270.

73. Suhre K, Wallaschofski H, Raffler J, Friedrich N, Haring R, Michael K, Wasner C, Krebs A, Kronenberg F, Chang D *et al*: **A genome-wide association study of metabolic traits in human urine**. *Nature genetics* 2011, **43**(6):565-569.

74. Lemaitre RN, Tanaka T, Tang W, Manichaikul A, Foy M, Kabagambe EK, Nettleton JA, King IB, Weng LC, Bhattacharya S *et al*: **Genetic loci associated with plasma phospholipid n-3 fatty acids: a meta-analysis of genome-wide association studies from the CHARGE Consortium**. *PLoS genetics* 2011, **7**(7):e1002193.

75. Wu JH, Lemaitre RN, Manichaikul A, Guan W, Tanaka T, Foy M, Kabagambe EK, Djousse L, Siscovick D, Fretts AM *et al*: **Genome-wide association study identifies novel loci associated with concentrations of four plasma phospholipid fatty acids in the de novo lipogenesis pathway: results from the Cohorts for Heart and Aging Research in**

Genomic Epidemiology (CHARGE) consortium. *Circ Cardiovasc Genet* 2013, **6**(2):171-183.

76. Rueedi R, Ledda M, Nicholls AW, Salek RM, Marques-Vidal P, Morya E, Sameshima K, Montoliu I, Da Silva L, Collino S *et al*: **Genome-wide association study of metabolic traits reveals novel gene-metabolite-disease links**. *PLoS genetics* 2014, **10**(2):e1004132.