Genome-wide association studies with metabolomics

Jerzy Adamski

Abstract

Genome-wide association studies (GWAS) analyze the genetic component of a phenotype or the etiology of a disease. Despite the success of many GWAS, little progress has been made in uncovering the underlying mechanisms for many diseases. The use of metabolomics as a readout of molecular phenotypes has enabled the discovery of previously undetected associations between diseases and signaling and metabolic pathways. In addition, combining GWAS and metabolomic information allows the simultaneous analysis of the genetic and environmental impacts on homeostasis. Most success has been seen in metabolic diseases such as diabetes, obesity and dyslipidemia. Recently, associations between loci such as FADS1, ELOVL2 or SLC16A9 and lipid concentrations have been explained by GWAS with metabolomics. Combining GWAS with metabolomics (mGWAS) provides the robust and quantitative information required for the development of specific diagnostics and targeted drugs. This review discusses the limitations of GWAS and presents examples of how metabolomics can overcome these limitations with the focus on metabolic diseases.

Keywords complex disease, genetics, metabolomics, metabolic traits, pathway inference

Complex diseases: omics and genome-wide association studies

Common, severe human diseases such as cancer, diabetes, asthma, or mental and cardiovascular disorders have complex etiologies and complex mechanisms. To uncover the causal events leading to these diseases, information on the factors that challenge human health and the immediate responses to these challenges is needed. Yet, unfortunately, the dataset is never complete. In most cases, studies of humans are restricted to observations after a disease has occurred, except in clinical cases when individuals with particular diseases are treated or take part in randomized controlled intervention trials. Outside clinical trials, longitudinal studies (observational studies tracking the same individual) that analyze phenotypes can also be undertaken. Both of these types of studies are hampered by unknown and uncontrolled exposure to the environment (such as differences in nutrition, medication, environmental endocrine disruptors and lifestyle) even in well-phenotyped cohorts (where weight, height and health status, for example, are known).

Cohorts can be analyzed for specific features such as genomic variance (variants in the DNA sequence) or metric parameters (concentrations or comparative levels) of RNA, proteins or metabolites. If the features analyzed and disease phenotypes coincide (and the frequency of coincidence is biostatistically valid), then it would be possible to identify the pathways involved. Therefore, a current approach to unveiling the etiology and mechanism of complex diseases is to employ sophisticated analysis methodologies (omics) that allow for the integration of multiple layers of molecular and organismal data. Data acquired with omics have already contributed considerably to the understanding of homeostasis in health and disease. Genome-wide association studies (GWAS), in particular, have contributed substantially to the field in the past 6 years [1]. This approach has identified numerous genetic loci that are associated with complex diseases. However, the number of genetic mechanisms that have been identified to explain complex diseases has not increased significantly [2].

In this review, I will highlight the current limitations of GWAS and how issues such as the large sample size required can be overcome by adding metabolomics information to these studies. I will explain the principles behind the combination of metabolomics and GWAS (mGWAS) and how together they can provide a more
powerful analysis. I conclude by exploring how mGWAS has been used to identify the metabolic pathways involved in metabolic diseases.

Aims and limitations of GWAS

GWAS analyze the association between common genetic variants and specific traits (phenotypes). The phenotypes originally included weight (or body mass index), height, blood pressure or frequency of a disease. More recently, specific traits in the transcriptome, proteome or metabolome have been included, and these are usually quantitative (for example, concentration). GWAS can also be used to explore whether common DNA variants are associated with complex diseases (for example, cancer or type 2 diabetes mellitus). The common variants might be single nucleotide polymorphisms (SNPs), copy number polymorphisms (CNPs), insertions/deletions (indels) or copy number variations (CNVs), but most GWAS employ SNPs [3]. At present, SNPs are used most frequently because of coverage of a large fraction of genome, throughput of assay, quality assurance and cost-effectiveness. Because the concept of GWAS is hypothesis-free, the analyses of GWAS are generally genetically unbiased, but they assume a genetic cause that might not be the most significant contributor.

In the past, candidate gene and pedigree analyses were very successful in the study of diseases of monogenetic origin: heritable dysregulation of certain metabolomic traits (inborn errors of metabolism) were among the first to be associated with specific genes [4]. However, these approaches are not useful in complex diseases because candidate regions contain too many genes or there are no groups of related individuals with a clear inheritance pattern of the disease phenotype. Inspired by the success of the Mendelian inheritance (genetic characteristics passed from the parent organism to offspring) approach, a great effort was undertaken to generate a human reference database of common genetic variant patterns based on a haplotype survey - the haplotype map (HapMap) [5]. This resource indeed improved, through linkage disequilibrium (LD) analyses, both the quality and the speed of GWAS, but it has not solved the major issue of study outcome. The common limitation of GWAS is that they do not provide mechanisms for disease; in other words, GWAS are unable to detect causal variants. Specifically, a GWAS provides information about an association between a variant (for example, SNP) and a disease, but the connection between a SNP and a gene is sometimes unclear. This is because annotated genes in the vicinity of a SNP are used in an attempt to explain the association functionally. However, proximity to a gene (without any functional analyses) should not be taken as the only sign that the identified gene contributes to a disease.

It should be further noted that the current analysis tools for SNPs do not include all possible variants, but rather only common ones with a major allele frequency greater than 0.01. SNPs with frequencies of less than 1% are not visible (or hardly discernible) in GWAS at present [3], and therefore some genetic contributions might remain undiscovered. So far, associations discovered by GWAS have had almost no relevance to clinical prognosis or treatment [6], although they might have contributed to risk stratification in the human population. However, common risk factors fail to explain the heritability of human disease [7]. For example, a heritability of 40% had been estimated for type 2 diabetes mellitus [8,9], but only 5 to 10% of the type 2 diabetes mellitus heritability can be explained by the more than 40 confirmed diabetes loci identified by GWAS [9,10].

Overcoming the limitations

There are several ways to improve GWAS performance. Instead of searching for a single locus, multiple independent DNA variants are being selected to identify those responsible for the occurrence of a disease [2]. Odds ratios could be more useful than P-values for the associations [6] in the interpretation of mechanisms and the design of replication or functional studies. This is especially true if highly significant (but spurious) associations are observed in a small number of samples, which might originate from a stratified population. The design of GWAS is also moving from tagging a single gene as a cause of disease to illuminating the pathway involved. This pathway might then be considered as a therapeutic target. In this way, GWAS comes back to its roots. The term ‘post-GWAS’ is used to describe GWAS-inspired experiments designed to study disease mechanisms. This usually involves exploration of expression levels of genes close to the associated variants, or knockout experiments in cells or animals [11]. In other words, post-GWAS analyses bring functional validation to associations [12].

Although omics approaches are powerful, they do not provide a complete dataset. Each omic technology provides a number of specific features (for example, transcript level fold change, protein identity or metabolite concentration, concentration ratios). At present, experimental datasets consisting of thousands of features unfortunately do not encompass all the features present in vivo. With incomplete data, only imperfect conclusions can be expected. However, the coverage of different omics features is expanding rapidly to overcome both genetic and phenotypic limitations of GWAS. As for the genetic aspects, progress in whole genome sequencing (for example, the 1000 Genomes Project [13,14]) is beginning to provide more in-depth analyses for less frequent (but still significant), and multiple, co-existing
Metabolomic approaches for GWAS

Metabolomics is a very useful tool that complements classical GWAS for several reasons. These include quantification of metabolites, unequivocal identification of metabolites, provision of longitudinal (time-resolved) dynamic datasets, high throughput (for example, 1,000 samples per week on a single NMR spectrometer, gas chromatography-mass spectrometer (GC-MS) or liquid chromatography-tandem mass spectrometer (LC-MS/MS)) [20,25]. The number of metabolites identified varies depending on the tissue and is usually between 300 (blood plasma) and 1,200 (urine) [26]. The major advantage of non-targeted metabolomics is its unbiased approach to the metabolome. The quantification is a limiting issue in non-targeted metabolomics as it provides the differences in the abundance of metabolites rather than absolute concentrations. In silico analyses (requiring access to public [27-30] or proprietary [31,32] reference databanks) are required to annotate the NMR peaks, LC peaks or ion traces to specific metabolites. Therefore, if a metabolite mass spectrum is not available in the databases, the annotation is not automatic but requires further steps. These may include analyses under different LC conditions, additional mass fragmentation or high-resolution (but slow) NMR experiments.

Targeted metabolomics work with a defined set of metabolites and can reach a very high throughput (for example, 1,000 samples per week on a single LC-MS/MS). The set might range from 10 to 200 metabolites in a specific (for example, only for lipids, prostaglandins, steroids or nucleotides) GC-MS or LC-MS/MS assay [33-37]. To cover more metabolites, samples are divided into aliquots and parallel assays are run under different conditions for GC- or LC-MS/MS. In each of the assays the analyzing apparatus is tuned for one or more specific chemical classes and stable isotope labeled standards are used to facilitate concentration determination. The major advantages of targeted metabolomics are the throughput and absolute quantification of metabolites.

Both approaches (that is, targeted and non-targeted) reveal a large degree of common metabolite coverage [38] or allow for quantitative comparisons of the same metabolites [21,39]. Metabolomics generates large-scale datasets, in the order of thousands of metabolites, which are easily included in bioinformatics processing [40,41].
GWAS with metabolomics traits

The outcome of GWAS depends very much on the sample size and the power of the study, which increases with the sample size. Some criticisms of GWAS have addressed this issue by questioning whether GWAS are theoretically big enough to overcome the threshold of P-values and associated odds ratios. Initial GWAS for a single metabolic trait (that is, plasma high-density lipoprotein (HDL) concentration [42]) were unable to detect the genetic component even with 100,000 samples. This indicates low genetic penetrance for this trait and suggests that another approach should be used to delineate the underlying mechanism. More recently, metabolomics was found to reveal valuable information when combined with GWAS. Studies with a much smaller sample size (284 individuals) but with a larger metabolic set (364 featured concentrations) demonstrated the advantage of GWAS combined with targeted metabolomics [34]. In this study the genetic variants were able to explain up to 28% of the metabolic ratio variance (that is, the presence or absence of a genetic variant coincided with up to 28% of changes in concentration ratios of metabolites from the same pathway). Moreover, the SNPs in metabolic genes were indeed functionally linked to specific metabolites converted by the enzymes, which are gene products of the associated genes.

In another study on the impact of genetics in human metabolism [35], involving 1,809 individuals but only 163 metabolic traits, followed by targeted metabolomics (LC-MS/MS), it was shown that in loci with previously known clinical relevance in dyslipidemia, obesity or diabetes (FADS1, ELOVL2, ACADS, ACADM, ACADL, SPTLC3, ETFDH and SLC16A9) the genetic variant is located in or near genes encoding enzymes or solute carriers whose functions match the associating metabolic traits. For example, variants in the promoter of FADS1, a gene that encodes a fatty acid desaturase, coincided with changes in the conversion rate of arachidonic acid. In this study, the metabolite concentration ratios were used as proxies for enzymatic reaction rates, and this yielded very robust statistical associations, with a very small P-value of $6.5 \times 10^{-19}$ for FADS1. The loci explained up to 36% of the observed variance in metabolite concentrations [35].

In a recent fascinating study on the genetic impact on the human metabolome and its pharmaceutical implications with GWAS and non-targeted metabolomics (GC or LC-MS/MS), 25 genetic loci showed unusually high penetrance in a population of 1,768 individuals (repli­cated in another cohort of 1,052 individuals) and accounted for up to 60% of the difference in metabolite levels per allele copy. The study generated many new hypotheses for biomedical and pharmaceutical research [21] for indications such as cardiovascular and kidney disorders, type 2 diabetes, cancer, gout, venous thrombosis and Crohn’s disease.

A specific subset of the metabolome dealing with lipids termed lipidomics has provided important insights into how genetics contributes to modulated lipid levels. This area is of particular interest for cardiovascular disease research, as about 100 genetic loci (without causal explanation as yet) are associated with serum lipid concentrations [42]. Lipidomics increases the resolution of mGWAS over that with complex endpoints such as total serum lipids (for example, HDL only). For example, a NMR study showed that eight loci (LIPC, CETP, PLTP, FADS1, -2, and -3, SORT1, GCKR, APOB, APOA1) were associated with specific lipid subfractions (for example, chylomicrons, low-density lipoprotein (LDL), HDL), whereas only four loci (CETP, SORT1, GCKR, APOA1) were associated with serum total lipids [43]. GWAS have already enabled tracing of the impact of human ancestry on n-3 polyunsaturated fatty acid (PUFA) levels. These fatty acids are an important topic in nutritional science in trying to explain the impact of PUFA levels on immunological responses, cholesterol biosynthesis and cardiovascular disease [44–47]. It has been shown that the common variation in n-3 metabolic pathway genes and in the GCKR locus, which encodes the glucose kinase regulator protein, influences the levels of plasma phospholipid of n-3 PUFAs in populations of European ancestry, whereas in other ancestries (for example, African or Chinese) there is an impact on the influences in the FADS1 locus [48]. This explains the mechanisms of different responses to diet in these populations. GWAS with NMR-based metabolomics can also be applied to large cohorts. An example is the analysis of 8,330 individuals in whom significant associations ($P < 2.31 \times 10^{-10}$) were identified at 31 loci, including 11 new loci for cardiometabolic disorders (among these most were allocated to the following genes: SLC1A4, PPM1K, F12, DHDSL, TAT, SLC2A4, SLC25A1, FCGR2B, FCGR2A) [49]. A comparison of 95 known loci with 216 metabolite concentrations uncovered 30 new genetic or metabolic associations ($P < 5 \times 10^{-8}$) and provides insights into the underlying processes involved in the modulation of lipid levels [50]. mGWAS can also be used in the assignment of new functions to genes. In metabolite quantitative trait locus (mQTL) analyses with non-targeted NMR-based metabolomics, a previously uncharacterized familial component of variation in metabolite levels, in addition to the heritability contribution from the corresponding mQTL effects, was discovered [38]. This study demonstrated that the so-far functionally unannotated genes NAT8 and PYROID2 are new candidates for the mediation of changes in the metabolite levels of triethylamine and dimethylamine.

Serum-based GWAS with LC/MS targeted metabolomics
has also contributed to field of function annotation: SLC16A9, PLEKHH1 and SYNE2 have been assigned to transport of acylcarnitine C5 and metabolism of phosphatidylcholine PCae36:5 and PCaa28:1, respectively [34,35].

mGWAS has recently contributed to knowledge on how to implement personalized medicine by analysis of the background of sexual dimorphism [51]. In 3,300 independent individuals 131 metabolite traits were quantified, and this revealed profound sex-specific associations in lipid and amino acid metabolism - for example, in the CPS1 locus (carbamoyl-phosphate synthase 1; \( P = 3.8 \times 10^{-10} \)) for glycine. This study has important implications for strategies concerning the development of drugs for the treatment of dyslipidemia and their monitoring; an example would be statins, for which different predispositions should now be taken into account for women and men.

GWAS and metabolic pathway identification

By integrating genomics, metabolomics and complex disease data, we may be able to gain important information about the pathways that are involved in the development of complex diseases. These data are combined in systems biology [52] and systems epidemiology evaluations [53,54]. For example, SNP rs1260326 in GCKR lowers fasting glucose and triglyceride levels and reduces the risk of type 2 diabetes [55]. In a recent mGWAS [35], this locus was found to be associated with different ratios between phosphatidylcholines, thus providing new insights into the functional background of the original association. The polymorphism rs10830963 in the melatonin-receptor gene MTNR1B has been found to be associated with fasting glucose [56], and the same SNP associates with tryptophan:phenylalanine ratios in mGWAS [35]: this is noteworthy because phenylalanine is a precursor of melatonin. This may indicate a functional relationship between the phenylalanine-melatonin pathway and the regulation of glucose homeostasis. The third example is SNP rs964184 in the apolipoprotein cluster APOA1-APOC3-APOA4-APOA5, which associates strongly with blood triglyceride levels [57]. The same SNP associates with ratios between different phosphatidylcholines in mGWAS [35]: these are biochemically connected to triglycerides by only a few enzymatic reaction steps.

Conclusions

By combining metabolomics as a phenotyping tool with GWAS, the studies gain more precision, standardization, robustness and sensitivity. Published records worldwide illustrate the power of mGWAS. They provide new insights into the genetic mechanisms of diseases that is required for personalized medicine.

References

1. Ku CS, Loy EE, Pawan Y, Chia KS. The pursuit of genome-wide association studies: where are we now? J Hum Genet. 2010, 55:195-206.
2. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. Am J Hum Genet. 2012, 90:7-24.
3. Altmhuler O, Daly MJ, Lander ES. Genetic mapping in human disease. Science 2008, 322:881-888.
4. Palossari PM, Kilpoom JM, Hiltunen JK. Peroxidial diseases. Ann Med 1992, 24:163-166.
5. de Bakker PI, Yelensky R, Peter I, Gabriel SB, Daly MJ, Altshuler D. Efficiency and power in genetic association studies. Nat Genet. 2005, 37:1217-1223.
6. McClellan J, King MC. Genetic heterogeneity in human disease. Cell 2010, 141:210-217.
7. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM. Finding the missing heritability of complex diseases. Nature 2009, 461:747-753.
8. So HC, Yip BH, Sham PC. Estimating the total number of susceptibility variants underlying complex diseases from genome-wide association studies. PLoS One 2010, 5:e13989.
9. Park KS. The search for genetic risk factors of type 2 diabetes mellitus. Diabetes Metabol J 2011, 35:12-22.
10. McCarthy MI. Genomics, type 2 diabetes, and obesity. N Engl J Med 2010, 363:2339-2350.
11. Freedman ML, Monteiro AN, Gayther SA, Coetzee GA, Risch A, Plais C, Casey G, De Basi M, Carlson C, Duggan D, James M, Liu P, Tichelaar JW, Wiks HG, You M, Mills IG. Principles for the post-GWAS functional characterization of cancer risk loci. Nat Genet 2011, 43:513-518.
12. Doucett M. Select: GWAS gets functional. Cell 2010, 143:177-178.
13. Via M, Gingeron C, Burchard EG. The 1000 Genomes Project: new opportunities for research and social challenges. Genome Med 2010, 2:3.
14. Huang J, Ellinghaus D, Franke A, Howie B, Li Y. 1000 Genomes-based imputation identifies novel and refined associations for the Wellcome Trust Case Control Consortium phase 1 Data. Eur J Hum Genet. 2012, doi: 10.1038/ejhg.2012.3.
15. Rankin VC, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. Nat Rev Genet 2011, 12:529-541.
16. Bataille V, Lens M, Spector TD. The use of the twin model to investigate the genetics and epigenetics of skin diseases with genomic, transcriptomic and methylation data. J Eur Acad Dermatol Venereol 2012, doi: 10.1111/j.1468-3083.2011.04444.x.
17. Krueger F, Krecek B, Franke A, Andrews SR. DNA methylome analysis using short bisulfite sequencing data. Nat Methods 2012, 9:145-151.
Polonikov A, Peters A, Theis FJ, Kronenberg F, Weidinger S, Wichmann HE, Suhe K, Wang-Sattler R, Adamski J, Illig T: Discovery of sexual dimorphisms in metabolic and genetic biomarkers. PLoS Genet 2011, 7:e1002215.

52. Hood L, Heath JR, Phelps ME, Lin B: Systems biology and new technologies enable predictive and preventative medicine. Science 2004, 306:640-643.

53. Adourian A, Jennings E, Balasubramanian R, Hines WM, Damian D, Plasterer TN, Clish CB, Stroobant P, McBurney R, Verheij ER, Bobeldijk I, van der Greef J, Lindberg J, Kenne K, Andersson U, Hellmold H, Nilsson K, Salter H, Schuppe-Koistinen I: Correlation network analysis for data integration and biomarker selection. Mol Biosyst 2008, 4:249-259.

54. Haring R, Wallaschofski H: Diving through the “-omics”: the case for deep phenotyping and systems epidemiology. OMICS 2012. doi: 10.1089/omi.2011.0108

55. Vaxillaire M, Cavalcanti-Proenca C, Dechaume A, Tichet J, Marre M, Balkau B, Froguel P: The common P446L polymorphism in GCKR inversely modulates fasting glucose and triglyceride levels and reduces type 2 diabetes risk in the DESIR prospective general French population. Diabetes 2008, 57:2253-2257.

56. Prokopenko I, Langenberg C, Florez JC, Saxena R, Soranzo N, Thorleifsson G, Loos RJ, Manning AK, Jackson AU, Aulchenko Y, Potter SC, Erdos MR, Sanna S, Hottenga JJ, Wheeler E, Kaakinen M, Lyssenko V, Chen WM, Ahmadi K, Beckmann JS, Bergman RN, Bochud M, Bonnycastle LL, Buchanan TA, Cao A, Cervino A, Con L, Collins FS, Crisponi L, de Geus EJ, et al: Variants in MTNR1B influence fasting glucose levels. Nat Genet 2009, 41:77-81.

57. Kathiresan S, Melander O, Guiducci C, Surti A, Burtt NP, Rieder MJ, Cooper GM, Roos C, Voight BF, Havulinna AS, Wahlstrand B, Hedner T, Corella D, Tai ES, Ordovas JM, Berglund G, Vartiainen E, Jou slahti P, Hedblad B, Taskinen MR, Newton-Cheh C, Salomaa V, Pettonen L, Groop L, Altshuler DM, Orho-Melander M: Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. Nat Genet 2008, 40:189-197.

doi:10.1186/gm333
Cite this article as: Adamski J. Genome-wide association studies with metabolomics. Genome Medicine 2012, 4:34.