Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Biotechnology

Computational approaches for systems metabolomics

Jan Krumsiek[1,2,4], Jörg Bartel[1,2,4] and Fabian J Theis[1,3]

Systems genetics is defined as the simultaneous assessment and analysis of multi-omics datasets. In the past few years, metabolomics has been established as a robust tool describing an important functional layer in this approach. The metabolome of a biological system represents an integrated state of genetic and environmental factors and has been referred to as a 'link between genotype and phenotype'. In this review, we summarize recent progresses in statistical analysis methods for metabolomics data in combination with other omics layers. We put a special focus on complex, multivariate statistical approaches as well as pathway-based and network-based analysis methods. Moreover, we outline current challenges and pitfalls of metabolomics-focused multi-omics analyses and discuss future steps for the field.

Addresses
[1] Institute of Computational Biology, Helmholtz Zentrum München, Neuherberg, Germany
[2] German Center for Diabetes Research (DZD e.V.), Germany
[3] Department of Mathematics, Technische Universität München, Garching, Germany

Corresponding author: Theis, Fabian J (fabian.theis@helmholtz-muenchen.de)
[4] These authors contributed equally to this work.

Current Opinion in Biotechnology 2016, **39**:198–206

This review comes from a themed issue on **Systems biology**

Edited by **Mark P Styczynski** and **Fabian J Theis**

For a complete overview see the Issue and the Editorial

Available online 30th April 2016

http://dx.doi.org/10.1016/j.copbio.2016.04.009

Introduction

In recent years, the biomedical research field has experienced tremendous advancements in high-throughput measurement technologies. Various layers of the central molecular dogma are now well covered by so-called 'omics' data, including the assessment of DNA variation, DNA modifications, transcript expression, protein abundances and modifications, as well as metabolite profiles. In human population studies, nowadays millions of molecular markers are screened across many omics levels in thousands of samples. The promise of such multi-omics datasets is to provide a holistic picture of the biological system in health and disease, giving rise to an exciting new branch of systems biology called 'systems genetics' [1••]. The central idea is that only by simultaneously assessing as many layers of the biological system as possible and, importantly, the complex interactions between them, we can develop a fundamental understanding of the underlying mechanisms between genotype and (patho)phenotype. The computational challenge is to develop statistical approaches that identify the additional knowledge expected to be buried in multi-omics datasets [2•].

Among the omics technologies, metabolomics plays a special role. The metabolome is the set of all small molecules, such as amino acids, sugars and lipids, in a biological system. It is considered to be an endpoint of biological processes and carries an imprint of all genetic, epigenetic and environmental factors [3]. It has therefore also been referred to as the 'link between genotype and phenotype' [4] (Figure 1a). As a consequence, the majority of biological and medical perturbations can be expected to be visible in the metabolome, making metabolites ideal biomarkers. Metabolomics has been particularly successful in the field of human epidemiology, with studies ranging from neurological disorders over type 2 diabetes to cardiovascular disease [5].

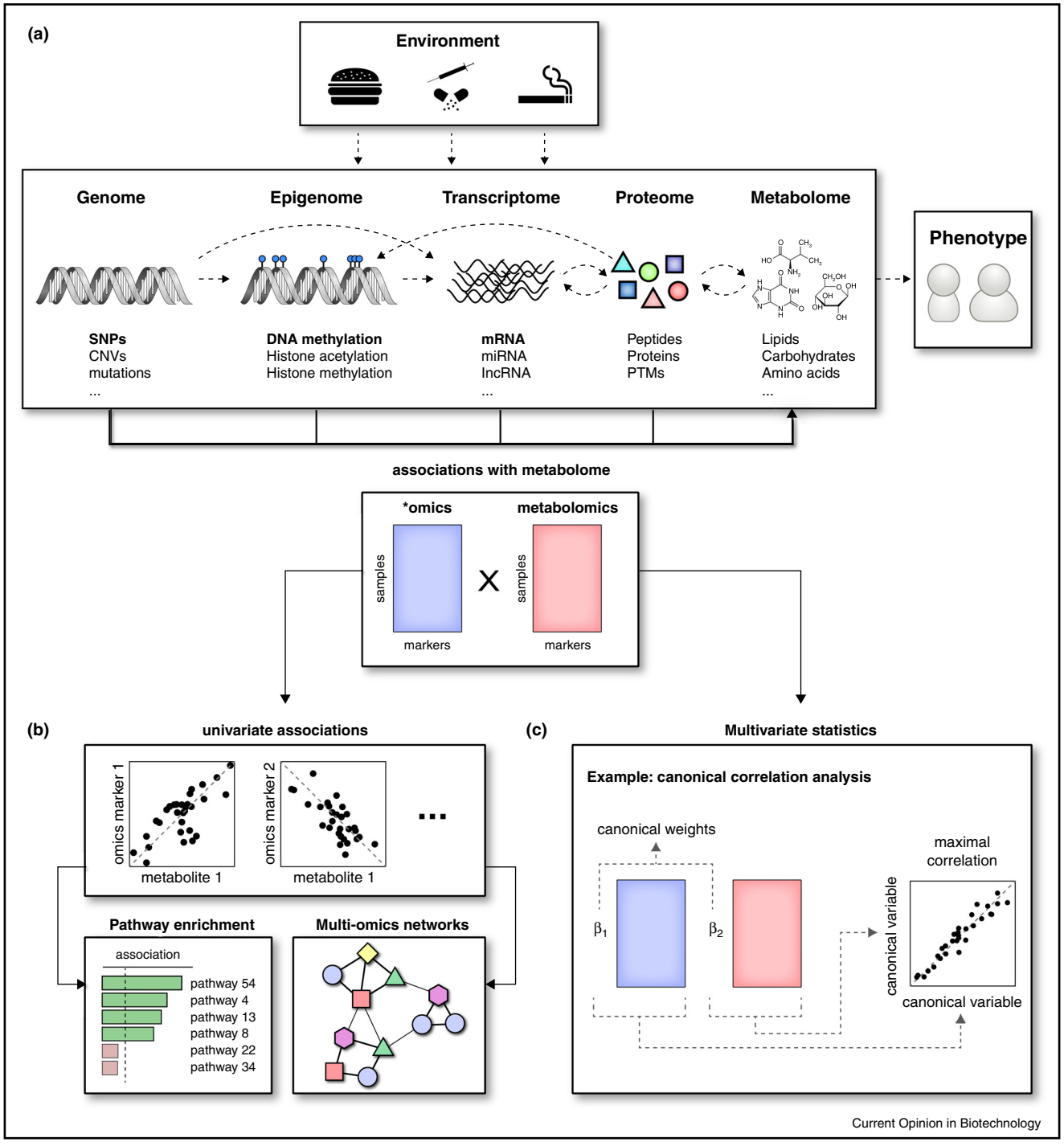
In this review, we summarize recent papers and developments in the analysis of metabolomics data with other molecular omics layers, with a special focus on studies in the human system. We particularly discuss statistical and computational methods, including pathway analysis, networks and multivariate integration methods. We deliberately omit the discussion of public resources, such as metabolic pathway and protein–protein interaction databases, since this would be beyond the scope of this review. For this topic, we refer the interested reader to Ng *et al.* [6•].

Metabolomics and DNA variation

High-throughput genotyping methods gave rise to large-scale genome-wide association studies (GWAS) over a decade ago [7], with the promise to elucidate the genetic basis of complex diseases. Many traits have since been correlated with single nucleotide polymorphisms (SNPs), including metabolomics measurements from human cohorts. Compared to other traits, a substantial amount of metabolites in human blood have been reported with remarkable heritability, showing exceptionally high fractions of variance explained by common genetic variants [8].

The first association study between genetic variation in the general population and metabolic traits in blood measured by mass-spectroscopy was performed by Gieger *et al.* [9], based on 363 metabolites measured in 284 male

Figure 1



Systems metabolomics. **(a)** Complex interplay between molecular omics, environment and the phenotype. The different layers of the biological system are nowadays well covered by various omics technologies. The metabolome is of special interest, since it integrates all molecular and environmental effects. It is to be noted that this chart represents a simplified view of information flow which is still subject of active debate. **(b)** Univariate associations between single metabolites and other omics markers are usually computed as a first line of analysis. These associations can then either be visualized and further analyzed as multi-omics networks, or grouped into overrepresented pathways using pathway enrichment analysis. **(c)** Multivariate methods exploit the usually high covariation of measurements within and between omics layers. Canonical correlation analysis is shown as an example. It seeks to find canonical weight vectors, such that the resulting canonical variables $\beta_1 X$ and $\beta_2 Y$ (X and Y representing the two data matrices), are maximally correlated. It then searches for the next pair of variables orthogonal to the first ones, and so on.

individuals of the German KORA cohort. The authors discovered that by using ratios of metabolites as metabolic traits, the associations with SNPs increase dramatically. Later studies steadily extended the number of associating loci by increasing statistical power with thousands of samples [10], larger panels of measured metabolites [8,11••], urine as an alternative biomarker fluid [12,13], NMR-based metabolite measurements [14] and non-Caucasian cohorts [15]. Metabolomics GWAS studies have identified hundreds of mQTLs, metabolomics quantitative trait loci, involving a variety of metabolic pathways. Notably, many genes occurring as top hits in these studies have previously been described as disease-related genes and drug targets [11••]. A recent, comprehensive review of the metabolomics GWAS field can be found in Kastenmüller *et al.* [16•].

Extending the common one SNP/one metabolite analyses, several authors suggested to associate correlated modules of metabolites with genetic variation. Such an approach increases statistical power (due to a reduced number of tests) while simultaneously removing unwanted variation in the data. For example, Inouye *et al.* [17] used canonical correlation analysis (Figure 1c), a multivariate extension of regular pairwise correlation, to associate correlated metabolite modules with SNPs in a GWAS. They identified several novel genetic loci that had not been linked with metabolic traits before. In a similar fashion, Ried *et al.* [18] defined metabolite modules from partial correlation networks and associated them with genetic variation using a permutation-based approach called 'phenotype set enrichment analysis'.

Several studies incorporated prior knowledge into the analysis to functionally characterize the substantial number of hits found in a metabolomics GWAS. The straight-forward and commonly used approach for this is gene set enrichment analysis (see e.g. Subramanian *et al.* [19], Figure 2a). It assesses whether SNPs identified in metabolomics GWAS tend to statistically accumulate in genes of predefined pathways. Two recent studies followed this approach, identifying systematic associations of lipid metabolites with the glycerolipid pathway [20] and the acetyl-CoA dehydrogenase pathway [21•].

Finally, extending the concept of set-based analysis to a more fine-grained level, our group used correlation analysis and Gaussian graphical models to establish a large-scale map (an 'atlas') of statistical associations between SNPs and metabolic pathways. This provided a detailed *in vivo* picture of the influence of genetic variation on metabolism in humans [11••,22•] (Figure 2b). Future studies could query this atlas, for example, to identify candidate genes for metabolites that have been associated with a disease phenotype.

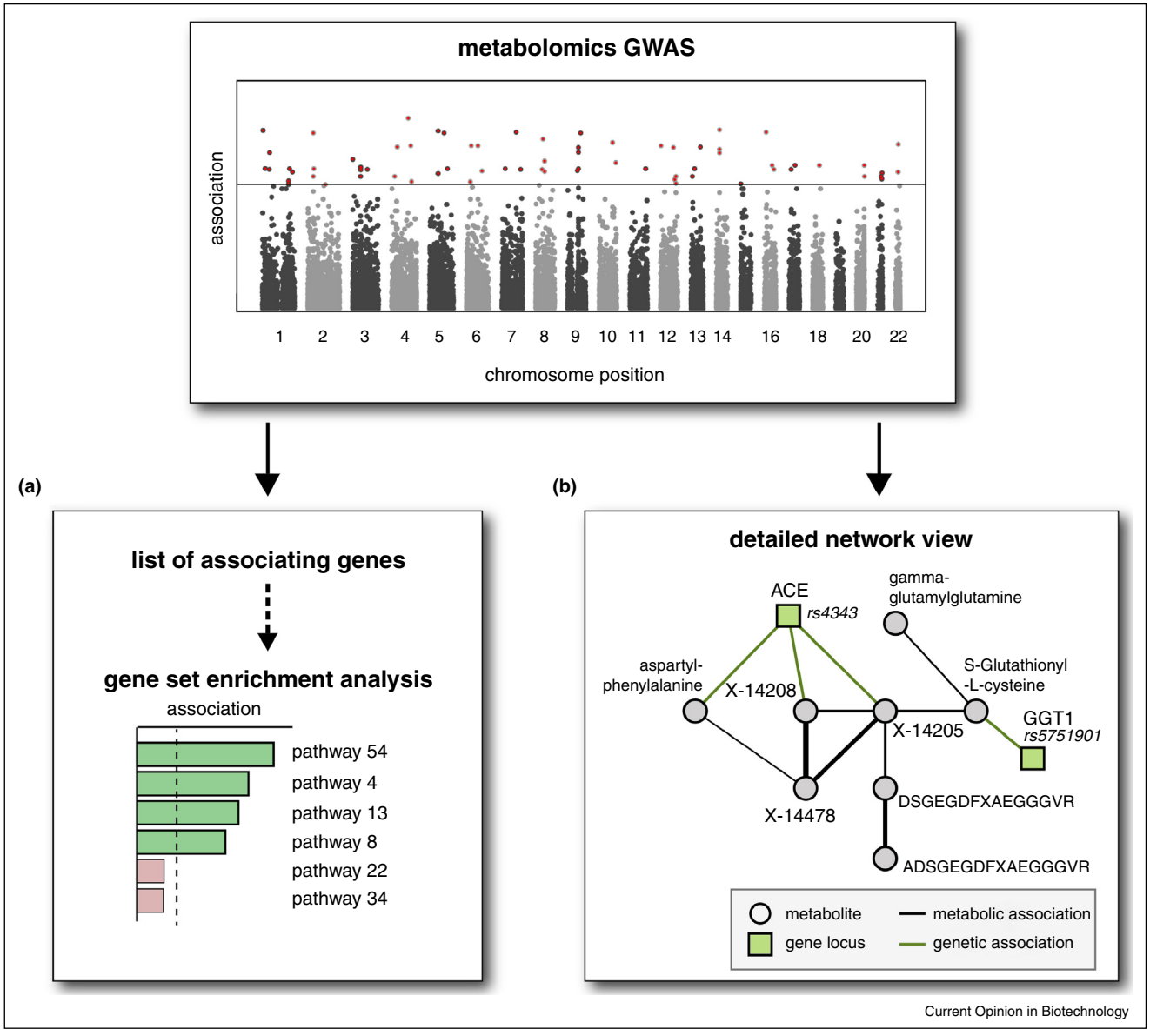
Metabolomics and transcriptomics

Another molecular layer frequently correlated with metabolomics data are gene expression profiles. Transcriptomics is a well-established measurement technology enabling the genome-wide assessment of gene expression. Coding transcripts are frequently used as proxies for protein levels, the direct physical interaction partners of metabolites. Moreover, in contrast to genetic data, transcript profiles also reflect environmental and lifestyle influences. Pioneering efforts for integrated metabolomics/transcriptomics analyses have been made in plant research, with papers that are now highly acknowledged in the field [23,24]. A recent review summarizes different strategies and methods for the integration of metabolomics and transcriptomics data [25•], including challenges with respect to experimental design and general multi-omics analysis strategies.

Multivariate statistical methods, which model the shared covariance between datasets, have become particularly important to associate high-dimensional, quantitative omics measurements. Simple univariate analyses, that is, comparing one metabolite and one transcript at a time, do not account for correlation structures within each omics level and are thus not able to capture the complex interplay between them. The general concept of multivariate methods is to find highly correlating directions (components, clusters) of markers within each dataset, which can then be associated between datasets (cf. Figure 1c). These methods include partial least squares (PLS) regression [26], O2-PLS [27] and canonical correlation analysis [28]. Griffin *et al.* used PLS regression to analyze time-resolved metabolomics and transcriptomics data of rat liver tissue to study fatty liver disease [26]. The authors identified systematic changes in lipid metabolism affected at both the transcriptional and metabolic level. Jozefczuk *et al.* used canonical correlation analysis to analyze the time-dependent metabolic and transcriptomic stress response of *E. coli* [28]. They were able to monitor metabolic adaptations to environmental conditions over time, identifying numerous condition-dependent associations between metabolites and transcripts. In a related approach, Wahl *et al.* used 'weighted correlation network analysis' (WGCNA) to separately define clusters of metabolites and transcripts, which were then correlated and later associated with weight gain in humans [29].

Extending the above-mentioned, purely statistical approaches, several authors suggested the use of prior knowledge in pathway-based methods for the joint analysis of metabolomics and transcriptomics. Su *et al.* mapped correlations computed in the NCI60 cancer cell line dataset onto a metabolic network model to identify correlating metabolite/transcript pairs involved in the same metabolic reactions or the same metabolic pathways [30]. Similarly, Çakir *et al.* mapped metabolites and transcripts to a genome-scale metabolic model in yeast

Figure 2



Integration of metabolomics data and SNPs. GWAS with metabolic traits generate significantly correlating SNP-metabolite pairs. **(a)** To determine which pathways accumulate a significant amount of metabolite-associating SNPs, gene set enrichment analysis can be performed (example in [21•]). **(b)** Integrating metabolite and SNP associations into a network view provides a detailed picture of *in vivo* pathway associations (panel B adapted from [22•]).

in order to identify different modes of metabolic regulation [31]. The need for pathway-based analysis methods also gave rise to publically available toolboxes. One of the first projects was the 'Integrated Molecular Pathway-Level Analysis' (IMPaLA) web tool, which provides pathway enrichment tools for joint metabolomics/transcriptomics datasets [32].

Only rather recently, several groups started to integratively analyze metabolomics and transcriptomics data from well-powered, large-scale human population cohorts. For instance, Inouye *et al.* were the first to build a large-scale correlation network from whole blood gene expression data of more than 500 individuals [33••]. The authors investigated the association between a leukocyte-specific gene expression module and circulating metabolites, mostly lipid parameters, in blood.

In a similar study, our group systematically analyzed the interplay between serum metabolites and whole blood

transcriptomics data from more than 700 individuals on a substantially wider metabolomics panel [34••]. We constructed a correlation network, the human 'Blood Metabolome-Transcriptome Interface', containing 114 metabolites and 522 transcripts. This network was then subjected to a series of bioinformatics analyses for further characterization. In brief, we showed that blood correlations between metabolites and transcripts represent parts of true metabolic pathways, especially lipid transport and immune system processes. We then identified signatures of coregulation by performing transcription-factor binding site enrichment analysis on all transcripts associated with a given metabolite or metabolic pathway. Finally, mapping statistical associations with HDL-C, LDL-C and TG, known risk factors for common diseases, revealed various pathways affected pathways in the network.

Metabolomics and proteomics

In contrast to the combination with transcriptomics data, integrative analysis studies of metabolomics and proteomics are still rare. While microarray-based and modern sequencing-based methods assess the transcriptome putatively in a genome-wide fashion, the coverage of mass-spectrometry-based peptide profiling is still limited [35]. Nevertheless, even at this early stage, integrating metabolomics and proteomics data provides valuable insights by investigating physically interacting molecules (metabolites and proteins) which are tightly connected to the phenotype. Importantly, merely investigating transcripts as proxies for proteins is not always sufficient due to the well-described discrepancy between protein concentrations and their respective coding mRNAs levels [36].

First studies on joint metabolomics/proteomics measurements also originated in the field of plant biology, for example to assign new enzymatic functions to proteins [37] and to gain mechanistic insights into plant-specific biochemical processes [38]. Wienkoop *et al.* used independent component analysis (ICA) combined with correlation networks to study the stress-induced associations between metabolites and proteins in *Arabidopsis thaliana* [39]. In the human system, Oberbach *et al.* used ICA to analyze time-resolved serum metabolomics and proteomics data from lean and obese healthy participants of a challenging study, revealing molecules and pathways associated with obesity [40•] (Figure 3a).

Similar to other omics combinations, specific methods have been developed to incorporate prior knowledge into the analysis process. For instance, integrative omics-metabolic analysis (IOMA) quantitatively integrates metabolomics and proteomics data into genome-wide metabolic models to predict kinetic rates under given experimental conditions [41]. To this end, it formulates a quadratic programming optimization problem to compute a set of metabolic fluxes which are compatible with rates inferred from the metabolomics and proteomics measurements. As another example, our group developed a method to identify regulated regions in metabolic pathways which show differential concentrations of both metabolites and proteins under different experimental conditions [42•]. The method is based on a random walk algorithm and attempts to combine 'seed' nodes with strong differential changes in the network into modules (Figure 3b). We applied the method to identify regulated pathways in Jurkat T cells stimulated with an environmental pollutant.

Future perspectives and challenges

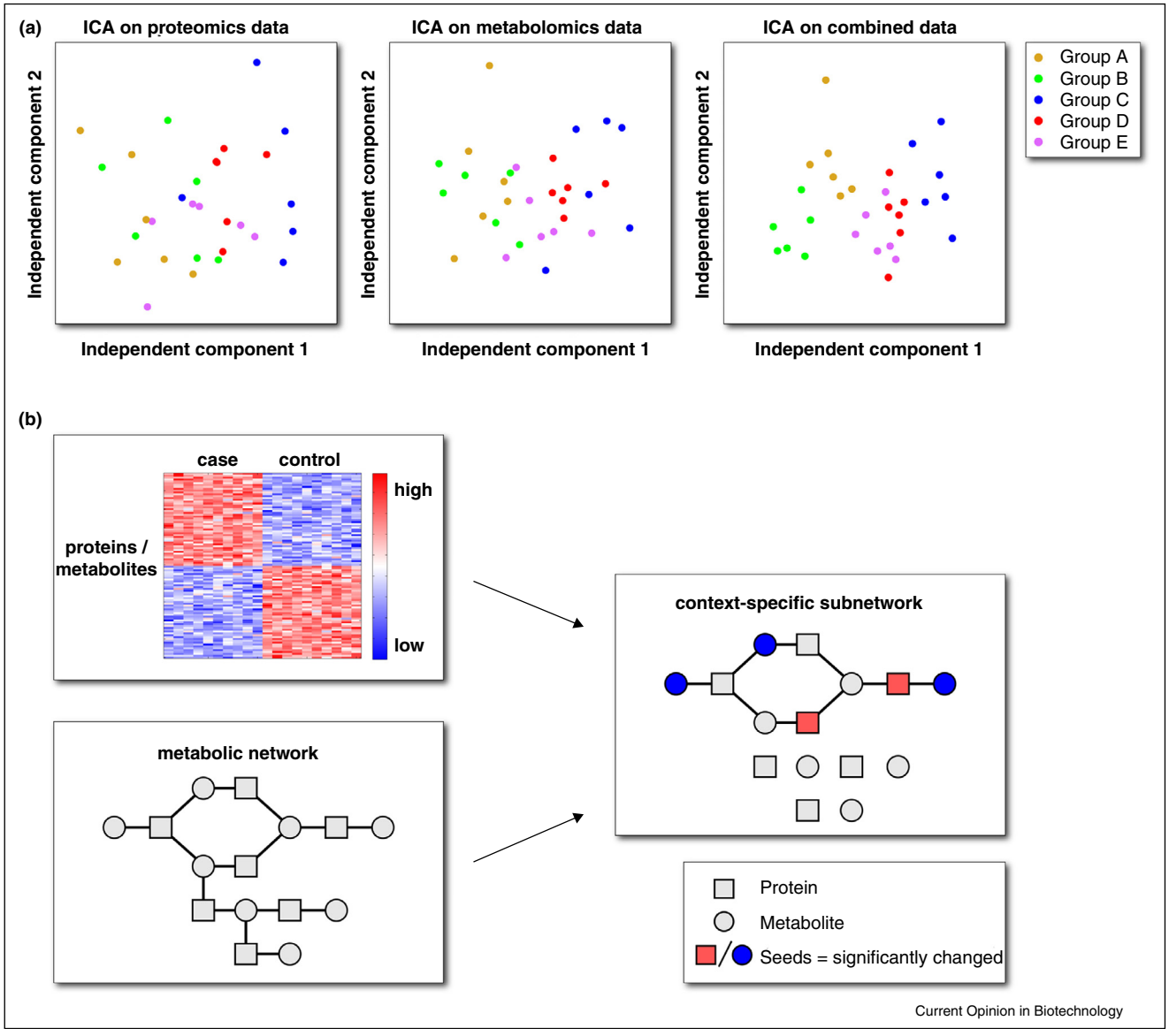
The field of metabolomics is now firmly established in systems genetics, and substantial improvements in multi-omics analysis methods have already been achieved in the past years. In the following, we will outline future directions and current challenges.

Commonly in biomedical research, major advances in the field can be expected from more efficient and precise measurement technology. For example, there are major efforts to further develop mass-spectrometry-based techniques [43•], in order to provide true 'metabolomes' covering the entirety of molecules in the biological system. In addition, more fine-grained omics technologies are currently being established, for instance covering post-translational modifications, such as glycoproteomics [44] and quantitative phosphoproteomics [45]. Integrating these novel technologies with metabolomics data will create interesting new computational challenges.

A crucial prerequisite to further advance the multi-omics research field will be the establishment of data processing standards and data depositing infrastructure. For example, in contrast to microarray-based and sequencing-based transcriptomics, there are no standardized pipelines for the normalization and quality control of mass-spectrometry-based metabolomics data. Different groups follow different approaches, and standardization efforts would facilitate reproducibility of results and increase overall research quality. On a related note, for gene expression data there are central data repositories, such as the Gene Expression Omnibus, GEO [46], which gives scientists open access to the raw data of published studies. For metabolomics, valuable first steps in this direction have been made with the MetaboLights project [47•], but more comprehensive efforts are required to achieve a higher coverage of publically available data.

On the analysis side, the metabolomics field should further adopt and extend methods from other 'omics disciplines. For example, substantial efforts have been made to annotate and prioritize GWAS hits using functional annotations from ENCODE and related projects [48•]. Similar approaches could be used to gain further insights into the nature and functional relevance of mQTLs. Moreover, extended GWAS methods are currently being developed,

Figure 3



Integration of metabolomics and proteomics data. **(a)** Independent component analysis on proteomics data, metabolomics data, and the combined data set. While there is no clear separation of experimental groups with any of the two omics alone, the combination of both provides a substantially stronger signal. Inspired by results from Oberbach *et al.* [40•]. **(b)** Module identification by a random walk algorithm. The algorithm identifies pathway modules with a high abundance of significantly changed proteins and metabolites, here called 'seed' nodes. Algorithm was used in Baumann *et al.* [42•].

such as prior knowledge-based ranking methods to identify pathways shared across different diseases [49]. Another branch of methods that should be further extended for multi-omics studies are machine-learning approaches, such as support vector machines, Bayesian networks and neural networks. For instance, Zhu *et al.* integrated metabolomics and transcriptomics data using Bayesian networks to construct causal regulatory networks in yeast [50].

One layer of molecular biology that we have not systematically addressed in this review are epigenetic modifications. Systematic high-throughput assessment of DNA methylation is possible today by microarray-based and sequencing technologies [51]. To the best of our knowledge, only a single epigenome-wide association study (EWAS) with systematic metabolomics measurements has been published to date [52•]. The authors reported 15 hits between metabolome and epigenome in humans and, importantly, point out various challenges in the assessment and interpretation of such associations. For example, correlations between CpG sites and metabolites are particularly confounded by underlying genetic variation and

environmental factors, thereby complicating direct functional interpretations. Since the metabolomics/epigenetics field is still young, pathway-based and network-based methods will have to be developed and applied after solving the fundamental issue of result interpretation.

Another upcoming omics layer not addressed here is the gut microbiome. Results from recent studies integrating metabolomics with metagenomics data from the human intestinal surface [53] or fecal samples [54] already revealed strong interactions between metabolism and gut microbiota. However, the field of metagenomics is still in its infancy, and further experimental and statistical methods need to be developed to infer true microbial compositions from DNA sequencing data.

Following up on result interpretation, associations between omics layers often raise the question of effect direction and causality. For genetic data, effect direction is a trivial issue due to the obvious immutability of the germline DNA. For any other omics level, assessment of causality poses a highly complex problem. If the effect directionality is not given by the study design, for example, in longitudinal studies, the possibilities to assess directionality become quite limited. Previous studies have attempted to model causality between two omics layers using SNPs as instrumental variables in structural equation modeling (SEM) [55,56] or Mendelian randomization [34••,57]. These approaches exploit the natural randomization of genotypes in an attempt to mimic randomized controlled trials, the gold standard to assess causality. However, most studies still lack statistical power and make strong, possibly false assumptions on the absence of confounding factors. Thus, substantial developments are still necessary to truly infer causal links from data.

Leaving the field of purely observational analyses, mechanistic models that define precise kinetic relationships of enzymatic reactions and transport processes will be a major next step in the metabolomics/multi-omics integration field. For unicellular model organisms, such as *E. coli* and yeast, detailed insights into the regulation of cellular metabolism have been established in the past years. For example, Fendt *et al.* [58] provide a detailed analysis of the relationship between enzyme capacity and metabolic concentrations in *S. cerevisiae*. In the human system, first steps have been taken towards the development of systematic mechanistic models, such as a whole-cell kinetic model of the human erythrocyte [59]. Future studies will have to face the challenge of a complex, multi-compartmental system in higher organisms, especially in the light of multi-omics datasets.

The ultimate application of system genetics will be large-scale, possibly longitudinal studies of pathophenotypes with simultaneous measurements for all omics layers, including phenotypic information (the phenome) [60]. This represents the transition from 'systems genetics' to 'systems medicine', that is, the patient-centric view on multi-omics data [61•]. A recent, famous study in this direction was the 'integrative personal omics profile, iPOP' study [62], where a single individual was monitored over a 14-month period with time-resolved measurements of genomics, transcriptomics, proteomics, metabolomics and clinical data. As outlined in this review, integrating and analyzing two omics layers at a time has already proved complicated, and the actual benefit and novel insights of true multi-omics datasets still remains to be demonstrated. Novel computational methods must be developed to process and analyze the massive data sets that will be produced.

We believe that metabolomics, as an established, strong link between genotype and phenotype, will continue to play a major role in the systems genetics field. Advancements both on the measurement and especially on the analysis side will produce exciting novel insights in the years to come.

Acknowledgements

We thank Erik van den Akker for valuable comments on the manuscript.

This work was supported by funding from the European Union's Seventh Framework Programme [FP7-Health-F5-2012] under grant agreement n° 305280 (MIMOmics), and by a grant from the German Federal Ministry of Education and Research (BMBF), Grant No. 01ZX1313C (project e:Athero-MED).

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. •• Civelek M, Lusis AJ: **Systems genetics approaches to understand complex traits**. *Nat Rev Genet* 2014, **15**:34-48 http://dx.doi.org/10.1038/nrg3575.
Comprehensive review of multi-omics datasets and the systems genetics field.

2. • Berger B, Peng J, Singh M: **Computational solutions for omics data**. *Nat Rev Genet* 2013, **14**:333-346 http://dx.doi.org/10.1038/nrg3433.
Overview of computational and statistical approaches for the multi-omics field.

3. Griffin JL: **The Cinderella story of metabolic profiling: does metabolomics get to go to the functional genomics ball?** *Philos Trans R Soc B Biol Sci* 2006, **361**:147-161 http://dx.doi.org/10.1098/rstb.2005.1734.

4. Fiehn O: **Metabolomics – the link between genotypes and phenotypes**. *Plant Mol Biol* 2002, **48**:155-171.

5. Kaddurah-Daouk R, Kristal BS, Weinshilboum RM: **Metabolomics: a global biochemical approach to drug response and disease**. *Annu Rev Pharmacol Toxicol* 2008, **48**:653-683 http://dx.doi.org/10.1146/annurev.pharmtox.48.113006.094715.

6. • Ng A, Bursteinas B, Gao Q, Mollison E, Zvelebil M: **Resources for integrative systems biology: from data through databases to networks and dynamic system models**. *Brief Bioinform* 2006, **7**:318-330 http://dx.doi.org/10.1093/bib/bbl036.
Overview of prior knowledge databases, including protein–protein interaction networks, gene regulatory networks and metabolic pathways.

7. Hirschhorn JN, Daly MJ: **Genome-wide association studies for common diseases and complex traits**. *Nat Rev Genet* 2005, **6**:95-108 http://dx.doi.org/10.1038/nrg1521.

8. Suhre K, Shin S-Y, Petersen A-K, Mohney RP, Meredith D, Wägele B *et al.*: **Human metabolic individuality in biomedical and pharmaceutical research**. *Nature* 2011, **477**:54-60 http://dx.doi.org/10.1038/nature10354.

9. Gieger C, Geistlinger L, Altmaier E, Hrabé de Angelis M, Kronenberg F, Meitinger T *et al.*: **Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum**. *PLoS Genet* 2008, **4**:e1000282 http://dx.doi.org/10.1371/journal.pgen.1000282.

10. Illig T, Gieger C, Zhai G, Römisch-Margl W, Wang-Sattler R, Prehn C *et al.*: **A genome-wide perspective of genetic variation in human metabolism**. *Nat Genet* 2010, **42**:137-141 http://dx.doi.org/10.1038/ng.507.

11. •• Shin S-Y, Fauman EB, Petersen A-K, Krumsiek J, Santos R, Huang J *et al.*: **An atlas of genetic influences on human blood metabolites**. *Nat Genet* 2014 http://dx.doi.org/10.1038/ng.2982.
Most comprehensive mQTL study to date, associating genome-wide SNPs with ~500 blood metabolites in ~8000 European individuals. This paper established a large-scale network view, an 'atlas' of *in vivo* genome-metabolome associations.

12. Nicholson G, Rantalainen M, Li JV, Maher AD, Malmodin D, Ahmadi KR *et al.*: **A genome-wide metabolic QTL analysis in Europeans implicates two loci shaped by recent positive selection**. *PLoS Genet* 2011, **7**:e1002270 http://dx.doi.org/10.1371/journal.pgen.1002270.

13. Suhre K, Wallaschofski H, Raffler J, Friedrich N, Haring R, Michael K *et al.*: **A genome-wide association study of metabolic traits in human urine**. *Nat Genet* 2011, **43**:565-569 http://dx.doi.org/10.1038/ng.837.

14. Kettunen J, Tukiainen T, Sarin A-P, Ortega-Alonso A, Tikkanen E, Lyytikäinen L-P *et al.*: **Genome-wide association study identifies multiple loci influencing human serum metabolite levels**. *Nat Genet* 2012, **44**:269-276 http://dx.doi.org/10.1038/ng.1073.

15. Yu B, Zheng Y, Alexander D, Morrison AC, Coresh J, Boerwinkle E: **Genetic determinants influencing human serum metabolome among African Americans**. *PLoS Genet* 2014, **10**:e1004212 http://dx.doi.org/10.1371/journal.pgen.1004212.

16. • Kastenmüller G, Raffler J, Gieger C, Suhre K: **Genetics of human metabolism: an update**. *Hum Mol Genet* 2015 http://dx.doi.org/10.1093/hmg/ddv263. ddv263.
Review of metabolomics GWAS, including various measurement platforms, different populations, different body fluids and genetic mapping approaches.

17. Inouye M, Ripatti S, Kettunen J, Lyytikäinen L-P, Oksala N, Laurila P-P *et al.*: **Novel loci for metabolic networks and multi-tissue expression studies reveal genes for atherosclerosis**. *PLoS Genet* 2012, **8**:e1002907 http://dx.doi.org/10.1371/journal.pgen.1002907.

18. Ried JS, Shin S-Y, Krumsiek J, Illig T, Theis FJ, Spector TD *et al.*: **Novel genetic associations with serum level metabolites identified by phenotype set enrichment analyses**. *Hum Mol Genet* 2014 http://dx.doi.org/10.1093/hmg/ddu301.

19. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA *et al.*: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles**. *Proc Natl Acad Sci U S A* 2005, **102**:15545-15550 http://dx.doi.org/10.1073/pnas.0506580102.

20. Demirkan A, van Duijn CM, Ugocsai P, Isaacs A, Pramstaller PP, Liebisch G *et al.*: **Genome-wide association study identifies novel loci associated with circulating phospho- and sphingolipid concentrations**. *PLoS Genet* 2012, **8**:e1002490 http://dx.doi.org/10.1371/journal.pgen.1002490.

21. • Hong M-G, Karlsson R, Magnusson PKE, Lewis MR, Isaacs W, Zheng LS *et al.*: **A genome-wide assessment of variability in human serum metabolism**. *Hum Mutat* 2013, **34**:515-524 http://dx.doi.org/10.1002/humu.22267.
mQTL study in human serum, including an enrichment analysis to identify pathways which accumulate associating SNPs.

22. • Krumsiek J, Suhre K, Evans AM, Mitchell MW, Mohney RP, Milburn MV *et al.*: **Mining the unknown: a systems approach to metabolite identification combining genetic and metabolic information**. *PLoS Genet* 2012, **8**:e1003005 http://dx.doi.org/10.1371/journal.pgen.1003005.
Network-integration of mQTL hits, combined with partial correlation networks within the metabolome. This study gave detailed insights into pathway interaction in human blood.

23. Urbanczyk-Wochniak E, Luedemann A, Kopka J, Selbig J, Roessner-Tunali U, Willmitzer L *et al.*: **Parallel analysis of transcript and metabolic profiles: a new approach in systems biology**. *EMBO Rep* 2003, **4**:989-993 http://dx.doi.org/10.1038/sj.embor.embor944.

24. Hirai MY, Yano M, Goodenowe DB, Kanaya S, Kimura T, Awazuhara M *et al.*: **Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in *Arabidopsis thaliana***. *Proc Natl Acad Sci U S A* 2004, **101**:10205-10210 http://dx.doi.org/10.1073/pnas.0403218101.

25. • Cavill R, Jennen D, Kleinjans J, Briedé JJ: **Transcriptomic and metabolomic data integration**. *Brief Bioinform* 2015 http://dx.doi.org/10.1093/bib/bbv090. bbv090.
Comprehensive review of analysis techniques for combined metabolomics and transcriptome. This review also outlines general concepts and challenges for multi-omics integration.

26. Griffin JL, Bonney SA, Mann C, Hebbachi AM, Gibbons GF, Nicholson JK *et al.*: **An integrated reverse functional genomic and metabolic approach to understanding orotic acid-induced fatty liver**. *Physiol Genomics* 2004, **17**:140-149 http://dx.doi.org/10.1152/physiolgenomics.00158.2003.

27. Bylesjö M, Eriksson D, Kusano M, Moritz T, Trygg J: **Data integration in plant biology: the O2PLS method for combined modeling of transcript and metabolite data**. *Plant J Cell Mol Biol* 2007, **52**:1181-1191 http://dx.doi.org/10.1111/j.1365-313X.2007.03293.x.

28. Jozefczuk S, Klie S, Catchpole G, Szymanski J, Cuadros-Inostroza A, Steinhauser D *et al.*: **Metabolomic and transcriptomic stress response of *Escherichia coli***. *Mol Syst Biol* 2010, **6**:364 http://dx.doi.org/10.1038/msb.2010.18.

29. Wahl S, Vogt S, Stückler F, Krumsiek J, Bartel J, Kacprowski T *et al.*: **Multi-omic signature of body weight change: results from a population-based cohort study**. *BMC Med* 2015, **13**:48 http://dx.doi.org/10.1186/s12916-015-0282-y.

30. Su G, Burant CF, Beecher CW, Athey BD, Meng F: **Integrated metabolome and transcriptome analysis of the NCI60 dataset**. *BMC Bioinform* 2011, **12**:S36.

31. Çakir T, Patil KR, Önsan ZI, Ülgen KÖ, Kirdar B, Nielsen J: **Integration of metabolome data with metabolic networks reveals reporter reactions**. *Mol Syst Biol* 2006, **2**:50 http://dx.doi.org/10.1038/msb4100085.

32. Kamburov A, Cavill R, Ebbels TMD, Herwig R, Keun HC: **Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA**. *Bioinform Oxf Engl* 2011, **27**:2917-2918 http://dx.doi.org/10.1093/bioinformatics/btr499.

33. •• Inouye M, Kettunen J, Soininen P, Silander K, Ripatti S, Kumpula LS *et al.*: **Metabonomic, transcriptomic, and genomic variation of a population cohort**. *Mol Syst Biol* 2010:6 http://dx.doi.org/10.1038/msb.2010.93.
First paper to compute a metabolome/transcriptome correlation network in a human population cohort.

34. •• Bartel J, Krumsiek J, Schramm K, Adamski J, Gieger C, Herder C *et al.*: **The human blood metabolome-transcriptome interface**. *PLoS Genet* 2015, **11**:e1005274 http://dx.doi.org/10.1371/journal.pgen.1005274.
Detailed bioinformatic analysis of metabolomics/transcriptome correlations, including pathway analysis, coregulation analysis and phenotype integration.

35. Breker M, Schuldiner M: **The emergence of proteome-wide technologies: systematic analysis of proteins comes of age**. *Nat Rev Mol Cell Biol* 2014, **15**:453-464 http://dx.doi.org/10.1038/nrm3821.

36. Vogel C, Marcotte EM: **Insights into the regulation of protein abundance from proteomic and transcriptomic analyses**. *Nat Rev Genet* 2012, **13**:227-232 http://dx.doi.org/10.1038/nrg3185.

37. Saghatelian A, Cravatt BF: **Global strategies to integrate the proteome and metabolome**. *Curr Opin Chem Biol* 2005, **9**:62-68 http://dx.doi.org/10.1016/j.cbpa.2004.12.004.

38. Weckwerth W, Wenzel K, Fiehn O: **Process for the integrated extraction, identification and quantification of metabolites, proteins and RNA to reveal their co-regulation in biochemical networks**. *Proteomics* 2004, **4**:78-83 http://dx.doi.org/10.1002/pmic.200200500.

39. Wienkoop S, Morgenthal K, Wolschin F, Scholz M, Selbig J, Weckwerth W: **Integration of metabolomic and proteomic phenotypes analysis of data covariance dissects starch and RFO metabolism from low and high temperature compensation response in *Arabidopsis thaliana***. *Mol Cell Proteomics* 2008, **7**:1725-1736 http://dx.doi.org/10.1074/mcp.M700273-MCP200.

40. • Oberbach A, Blüher M, Wirth H, Till H, Kovacs P, Kullnick Y *et al.*: **Combined proteomic and metabolomic profiling of serum reveals association of the complement system with obesity and identifies novel markers of body fat mass changes**. *J Proteome Res* 2011, **10**:4769-4788 http://dx.doi.org/10.1021/pr2005555.

Multivariate analysis of combined metabolome/proteome data. The authors use independent component analysis (ICA) and correlation networks to identify pathways related to obesity.

41. Yizhak K, Benyamini T, Liebermeister W, Ruppin E, Shlomi T: **Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model**. *Bioinform Oxf Engl* 2010, **26**:i255-i260 http://dx.doi.org/10.1093/bioinformatics/btq183.

42. • Baumann S, Rockstroh M, Bartel J, Krumsiek J, Otto W, Jungnickel H *et al.*: **Subtoxic concentrations of benzo[a]pyrene induce metabolic changes and oxidative stress in non-activated and affect the mTOR pathway in activated Jurkat T cells**. *J Integr OMICS* 2014:4 http://dx.doi.org/10.5584/jiomics.v4i1.157.

Integrated pathway analysis of metabolomics and proteomics data in stimulated Jurkat cells. Proposes an algorithm to identify highly regulated network modules.

43. • Wang Y, Liu S, Hu Y, Li P, Wan J-B: **Current state of the art of mass spectrometry-based metabolomics studies – a review focusing on wide coverage, high throughput and easy identification**. *RSC Adv* 2015, **5**:78728-78737 http://dx.doi.org/10.1039/C5RA14058G.

Review on current technological advances in mass-spectrometry-based metabolomics.

44. Lauc G, Krištić J, Zoldoš V: **Glycans – the third revolution in evolution**. *Front Genet* 2014, **5** http://dx.doi.org/10.3389/fgene.2014.00145.

45. Roux PP, Thibault P: **The coming of age of phosphoproteomics; from large data sets to inference of protein functions**. *Mol Cell Proteomics* 2013 http://dx.doi.org/10.1074/mcp.R113.032862.

46. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M *et al.*: **NCBI GEO: archive for functional genomics data sets—update**. *Nucleic Acids Res* 2013, **41**:D991-D995 http://dx.doi.org/10.1093/nar/gks1193.

47. • Haug K, Salek RM, Conesa P, Hastings J, Matos P, de Rijnbeek M *et al.*: **MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data**. *Nucleic Acids Res* 2013, **41**:D781-D786 http://dx.doi.org/10.1093/nar/gks1004.

EBI initiative to make metabolomics datasets publically available.

48. • Hou L, Zhao H: **A review of post-GWAS prioritization approaches**. *Front Genet* 2013:4 http://dx.doi.org/10.3389/fgene.2013.00280.

Review covering the functional annotation of GWAS hits, for example, using large-scale ENCODE project information.

49. Network and Pathway Analysis Subgroup of Psychiatric Genomics Consortium: **Psychiatric genome-wide association study analyses implicate neuronal, immune and histone pathways**. *Nat Neurosci* 2015, **18**:199-209 http://dx.doi.org/10.1038/nn.3922.

50. Zhu J, Sova P, Xu Q, Dombek KM, Xu EY, Vu H *et al.*: **Stitching together multiple data dimensions reveals interacting metabolomic and transcriptomic networks that modulate cell regulation**. *PLoS Biol* 2012, **10**:e1001301 http://dx.doi.org/10.1371/journal.pbio.1001301.

51. Bock C: **Analysing and interpreting DNA methylation data**. *Nat Rev Genet* 2012, **13**:705-719 http://dx.doi.org/10.1038/nrg3273.

52. • Petersen A-K, Zeilinger S, Kastenmüller G, Römisch-Margl W, Brugger M, Peters A *et al.*: **Epigenetics meets metabolomics: an epigenome-wide association study with blood serum metabolic traits**. *Hum Mol Genet* 2014, **23**:534-545 http://dx.doi.org/10.1093/hmg/ddt430.

First and only epigenome-wide association study (EWAS) with systematically measured metabolic traits.

53. McHardy IH, Goudarzi M, Tong M, Ruegger PM, Schwager E, Weger JR *et al.*: **Integrative analysis of the microbiome and metabolome of the human intestinal mucosal surface reveals exquisite inter-relationships**. *Microbiome* 2013, **1**:17 http://dx.doi.org/10.1186/2049-2618-1-17.

54. Feng Q, Liu Z, Zhong S, Li R, Xia H, Jie Z *et al.*: **Integrated metabolomics and metagenomics analysis of plasma and urine identified microbial metabolites associated with coronary heart disease**. *Sci Rep* 2016, **6**:22525 http://dx.doi.org/10.1038/srep22525.

55. Inouye M, Silander K, Hamalainen E, Salomaa V, Harald K, Jousilahti P *et al.*: **An immune response network associated with blood lipid levels**. *PLoS Genet* 2010, **6**:e1001113 http://dx.doi.org/10.1371/journal.pgen.1001113.

56. Shin S-Y, Petersen A-K, Wahl S, Zhai G, Römisch-Margl W, Small KS *et al.*: **Interrogating causal pathways linking genetic variants, small molecule metabolites, and circulating lipids**. *Genome Med* 2014, **6**:25 http://dx.doi.org/10.1186/gm542.

57. Smith GD, Hemani G: **Mendelian randomization: genetic anchors for causal inference in epidemiological studies**. *Hum Mol Genet* 2014, **23**:R89-R98 http://dx.doi.org/10.1093/hmg/ddu328.

58. Fendt S-M, Buescher JM, Rudroff F, Picotti P, Zamboni N, Sauer U: **Tradeoff between enzyme and metabolite efficiency maintains metabolic homeostasis upon perturbations in enzyme capacity**. *Mol Syst Biol* 2010, **6** http://dx.doi.org/10.1038/msb.2010.11.

59. Bordbar A, McCloskey D, Zielinski DC, Sonnenschein N, Jamshidi N, Palsson BO: **Personalized whole-cell kinetic models of metabolism for discovery in genomics and pharmacodynamics**. *Cell Syst* 2015, **1**:283-292 http://dx.doi.org/10.1016/j.cels.2015.10.003.

60. Houle D, Govindaraju DR, Omholt S: **Phenomics: the next challenge**. *Nat Rev Genet* 2010, **11**:855-866 http://dx.doi.org/10.1038/nrg2897.

61. • Auffray C, Chen Z, Hood L: **Systems medicine: the future of medical genomics and healthcare**. *Genome Med* 2009, **1**:2 http://dx.doi.org/10.1186/gm2.

Review introducing the new field of 'systems medicine', the patient-centric approach to systems genetics.

62. Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HYK, Chen R *et al.*: **Personal omics profiling reveals dynamic molecular and medical phenotypes**. *Cell* 2012, **148**:1293-1307 http://dx.doi.org/10.1016/j.cell.2012.02.009.