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].

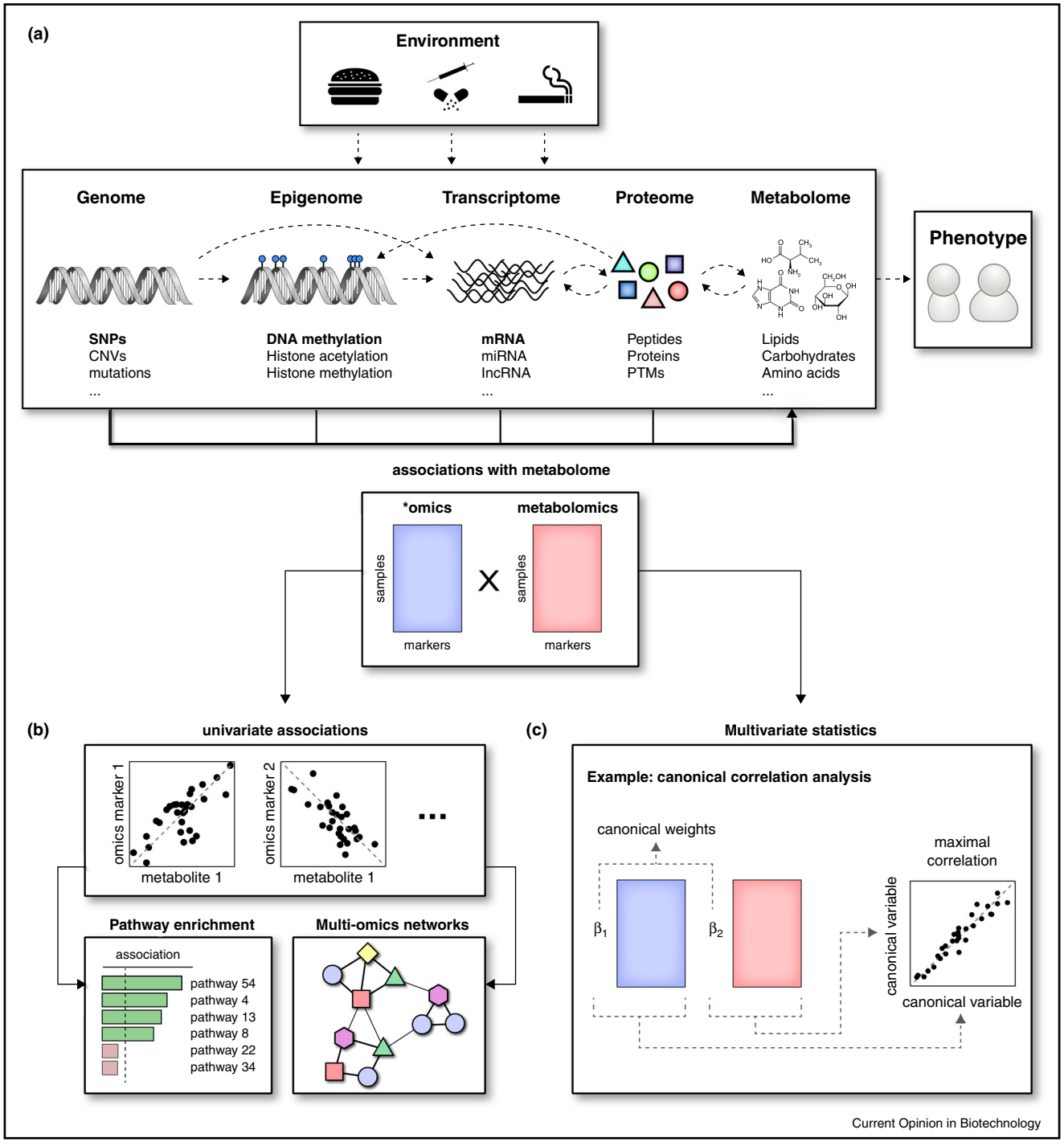
In this review, we summarize recent papers and developments in the analysis of metabolomics data with other molecular omics layers, with a special focus on studies in the human system. We particularly discuss statistical and computational methods, including pathway analysis, networks and multivariate 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.