Kelli M. Sas,[1] Alla Karnovsky,[2] George Michailidis,[3] and Subramaniam Pennathur[1]

Metabolomics and Diabetes: Analytical and Computational Approaches

Diabetes 2015;64:718–732 | DOI: 10.2337/db14-0509

METHODOLOGY REVIEW

Diabetes is characterized by altered metabolism of key molecules and regulatory pathways. The phenotypic expression of diabetes and associated complications encompasses complex interactions between genetic, environmental, and tissue-specific factors that require an integrated understanding of perturbations in the network of genes, proteins, and metabolites. Metabolomics attempts to systematically identify and quantitate small molecule metabolites from biological systems. The recent rapid development of a variety of analytical platforms based on mass spectrometry and nuclear magnetic resonance have enabled identification of complex metabolic phenotypes. Continued development of bioinformatics and analytical strategies has facilitated the discovery of causal links in understanding the pathophysiology of diabetes and its complications. Here, we summarize the metabolomics workflow, including analytical, statistical, and computational tools, highlight recent applications of metabolomics in diabetes research, and discuss the challenges in the field.

Diabetes is a metabolic disorder characterized by complex alterations in glucose and lipid metabolism in both type 1 (insulin deficiency due to autoimmune destruction of the pancreatic β-cells) and type 2 (insulin resistance and impaired insulin secretion due to islet cell dysfunction) diabetes. In congruence with the rise in obesity, diabetes is becoming increasingly prevalent. According to the Centers for Disease Control and Prevention, 8.3% of the U.S. population has diabetes and an estimated 35% have prediabetes (1). Metabolic diseases such as diabetes are often difficult for physicians to manage because they can be present for years before becoming clinically apparent. For example, significant β-cell dysfunction has already occurred by the time hyperglycemia becomes clinically evident. Conventional risk predictors of diabetes complications, such as degree of glycemic control, remain imperfect predictors of complications, mirroring our incomplete understanding of underlying pathophysiology. Metabolomics offers a new avenue for the identification of novel risk markers with the advent of high-throughput analytical platforms in which measurements of hundreds of analytes are now possible. Together with other omics data (genomics, transcriptomics, and proteomics) and bioinformatics pathway integration strategies, these technologies have the ability to illuminate the underlying biology and discover clinically relevant diagnostic and prognostic markers of disease risk. The purpose of this review is to highlight the role of metabolomics in diabetes research and discuss the tools for analyzing and integrating metabolomics data.

CHALLENGES OF METABOLOMICS IN HEALTH SCIENCES RESEARCH

Metabolomics attempts to comprehensively identify and quantify all or select groups of endogenous small molecule metabolites (<1,500 Da) in a biological system in a high-throughput manner. Although quantification of metabolites to study disease process is decades old (2–5), recent high-throughput methods have improved coverage of metabolites in biofluids (6). However, there are several technical challenges in broad-spectrum metabolomics studies. First, the metabolome is composed of a variety of chemically diverse compounds such as lipids, organic acids, carbohydrates,

[1]Division of Nephrology, Department of Internal Medicine, University of Michigan, Ann Arbor, MI
[2]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI
[3]Department of Statistics, University of Michigan, Ann Arbor, MI
Corresponding author: Subramaniam Pennathur, spennath@umich.edu.

Received 28 March 2014 and accepted 24 September 2014.

amino acids, nucleotides, and steroids, among others. In comparison, genes and proteins may perhaps be more chemically homogenous as each gene is a combination of only four basic nucleotides and each protein is composed of a mixture of 32 amino acids. Second, metabolites occur in a wide dynamic range of concentrations (nanomolar to millimolar) in biological systems. Third, not every metabolite is present in each tissue or biofluid. Finally, the metabolome can be altered by exogenous substances obtained from food or medications or endogenously by metabolism of gut microbiota, which may not be uniform in each subject. Therefore, comprehensive metabolomics is an analytical challenge. Indeed, no single metabolomics methodology is currently able to measure the entire metabolome accurately.

THE METABOLOMICS WORKFLOW

Metabolomics experiments follow a typical workflow consisting of experimental design, sample preparation, separation and detection of metabolites, data processing, and bioinformatics analysis (Fig. 1).

Experimental Design

When designing a metabolomics experiment, several aspects need to be considered. These include determining metabolites of interest (specific subset vs. all measurable), whether a snapshot of metabolite levels or determination of dynamic changes to the metabolome are required, and incorporation of biological and technical controls.

Targeted and Untargeted Approaches

Experiments can be designed with either a targeted or untargeted approach (Table 1). In targeted metabolomics, there is a predetermined list or class of metabolites that are being investigated. This approach is hypothesis driven, where a specific question is being addressed. One of the key features of targeted metabolomics is the use of isotope-labeled internal standards, which allows for the clear identification and quantification of analytes. Therefore, targeted metabolomics results in the high sensitivity and accurate detection and quantification of a relatively low number of metabolites at a given time. Conversely, untargeted metabolomics is hypothesis generating and aims to detect as many metabolites as possible, followed by identification of metabolites using software tools based on known or predicted spectral patterns. Various statistical tests, such as principal component analysis or random forest, can be used to classify phenotypes based on metabolite patterns (7). Untargeted metabolomics is particularly useful for identifying putative biomarkers, and experimental results can be confirmed by following untargeted experiments with a targeted approach. Although untargeted metabolomics can detect a large number of metabolites in a single run, quantification and high-quality precision is lost and the time required for accurate metabolite identification and quantification can be significant (8).

Steady-State or Metabolic Flux Analysis

Traditional metabolomics analyses assess steady-state metabolite levels or levels at a given time (time of cell/tissue harvest or time of biofluid collection) either in a targeted or untargeted manner. Steady-state detection will establish a difference in levels of a metabolite but will not provide information on why a difference occurs or through which metabolic pathway (e.g., glycolysis or gluconeogenesis). Therefore, it is sometimes necessary to determine the dynamic flow of metabolites through metabolic pathways. The influx into a pathway may not be equal to the efflux out of a pathway, resulting in a buildup or loss of a specific metabolite, pointing out critically regulated steps in metabolism. Metabolic flux analysis (MFA) allows for the time-dependent assessment of flux through pathways (9). For MFA, incorporation of heavy isotopes from individual substrates (e.g., U-[^{13}C]glucose) into specific metabolites (e.g., glyceraldehyde 3-phosphate) is used to determine the amount of a specific metabolite derived from a given pathway (Fig. 2) (10–13). The mass shift due to the heavy isotope is detected and the percent enrichment of the isotope in each metabolite, after correction for natural abundances, allows for determination of the percent or amount of metabolite present that was derived from a particular substance or pathway. MFA has been performed in cell culture (14), animal models (4), and humans (3). With MFA, isotope-labeled internal standards are not used as they can interfere with results. Although informative, MFA has its disadvantages over steady-state analysis; the isotope tracers are costly and analysis is time-consuming and complex. Due to data complexity, MFA has primarily been used with targeted analysis; however, recent strides have been made for the use of MFA with untargeted analysis (13).

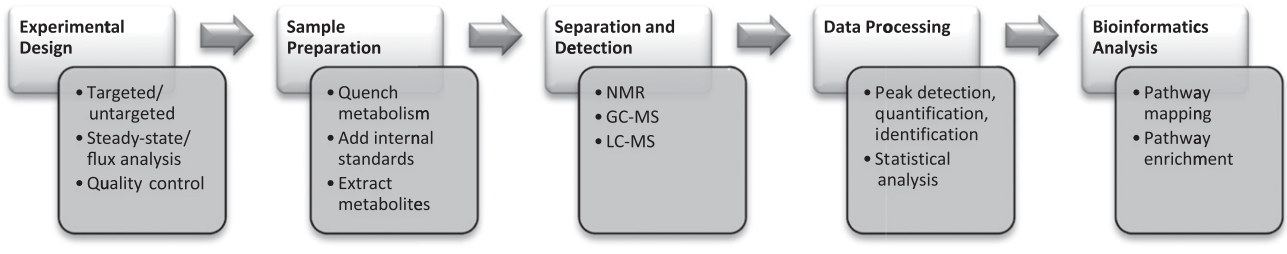

Figure 1—Summary of the metabolomics workflow.

Table 1—Comparison of targeted and untargeted metabolomics platforms

Feature	Targeted	Untargeted
Number of metabolites detected	Specific subset (usually <25 per run)	Typically ~500 reproducible known compounds and ~2,500 unknown compounds in human plasma
Identification	Individual isotope-labeled standards or authentic compounds	Library and software based
Quantitative	Yes	No
Data analysis time	Minimal (a typical experiment with two groups of n = 10/group and 20 analytes takes 2–3 days)	Significant (a typical experiment with two groups of n = 10/group takes 4–6 weeks)
Orthogonal technique required for confirmation	No	Yes

Experimental Quality Control

When planning any experiment, quality control needs to be considered. In the field of metabolomics, sample quality and technical reproducibility must be addressed (2,5). For sample quality, it is important to predetermine how samples will be collected and stored. For cell culture experiments, type of media, exposure to nutrients (glucose or amino acids), and processes to quench metabolism prior to sample analysis need to be carefully addressed. In human studies, factors such as diurnal variation, fasting versus fed state, type of anticoagulant used, diet, and medications need to be considered. Effects of anesthesia can lead to significant variation in animal studies. As a technical control, analytical pools should be interspersed among sample runs to allow for determination of instrumental variability and data quality. The reader is referred to an excellent review for a more in-depth discussion of quality control (2).

Sample Preparation

Due to the complexity of the metabolome, sample preparation varies depending upon experimental goals, sample matrix (tissue, biofluid, or cell culture), and analytical method to be used. Regardless of these factors, metabolism should be quenched as quickly as possible after sample collection and samples stored at −80°C. For a more in-depth review of sample preparation, please see previous literature on optimization techniques (5,6,15,16).

Analytical Approaches for Separation and Detection of Metabolites

The most frequently used analytical platforms in metabolomics are nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS), which is generally coupled to a chromatographic technique such as gas chromatography (GC) or liquid chromatography (MS). For a summary of MS separation and detection techniques for classes of metabolites common to diabetes research, see Table 2.

NMR Spectroscopy

NMR spectroscopy is highly reproducible and quantitative. NMR requires little sample preparation, as no separation or derivatization is required, and therefore does not destroy the sample. The basis for NMR spectroscopy revolves around the fact that the nuclei of many isotopes (e.g., ^{1}H, ^{13}C, etc.) have a characteristic spin and, when placed in a magnetic field, absorb radiation and resonate at a specific frequency. The primary limitation to NMR spectroscopy is its sensitivity, as concentrations can only be detected into the micromolar range, limiting its ability with low-abundance metabolites (17).

GC-MS

GC-MS is a highly sensitive and specific method for separation and detection of volatile metabolites such as organic acids. A carrier gas propels the sample through the separation column, after which it can be ionized by electron ionization or chemical ionization for detection by the mass spectrometer. As separation by GC occurs at high temperatures, samples need to be thermally stable as well as volatile. For samples to be readily volatile, chemical derivatization of samples may be necessary prior to analysis. Derivatization is one of the major drawbacks of GC-MS, as it can result in metabolite loss and can complicate analysis due to incomplete derivatization or artifact formation. For this reason, the proper derivatization method needs to be determined based upon the metabolite(s) of interest. Another drawback of GC-MS is the relatively limited mass range. However, GC-MS has some distinct advantages. Spectral patterns and retention times of compounds are highly reproducible by GC-MS, allowing compounds to be searched against existing libraries. Also, there is lower instrumental-based variability among results than with LC-MS.

LC-MS

LC-MS is the most commonly used platform for metabolomics studies. As opposed to GC-MS, there is no need for sample derivatization and there is greater coverage of mass ranges. LC-MS is versatile, allowing for the separation and detection of many different classes of metabolites. Part of the versatility of LC-MS is due to the various separation techniques and wide array of mass analyzers.

Selection of the appropriate chromatography column is an important step in LC-MS. Reverse-phase columns, such as C18 columns, provide good retention and separation of nonpolar compounds. Conversely, hydrophilic interaction chromatography (HILIC) columns have a high affinity for polar compounds. HILIC has increased sensitivity but less reproducibility of retention time, even within the same run. The introduction of ultraperformance LC, which uses smaller

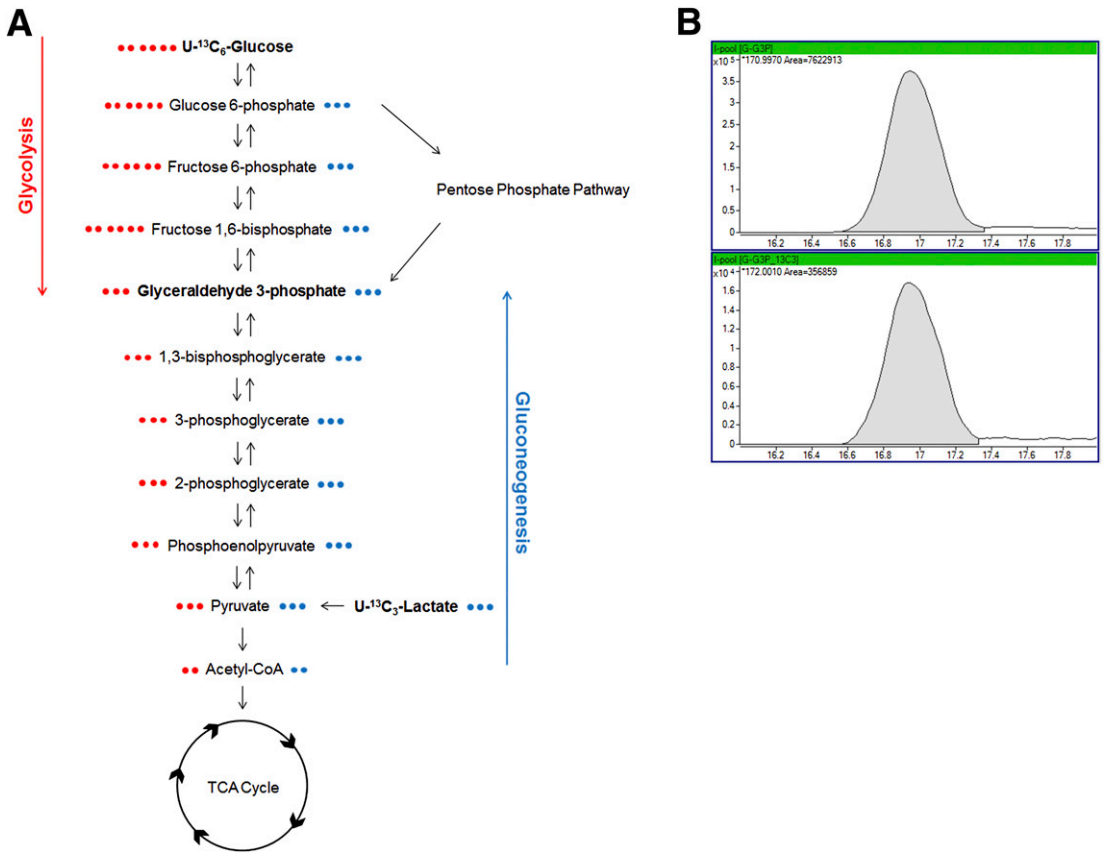

Figure 2—MFA of isotope tracers into glyceraldehyde 3-phosphate (G3P). *A*: Schematic depicting isotope incorporation into G3P using either U-[$^{13}C_6$]glucose or U-[$^{13}C_3$]lactate. With each isotope-labeled substance, G3P derived from glycolysis or gluconeogenesis, respectively, would have a mass shift of +3 due to all three carbons incorporating the ^{13}C label. Comparison of percent incorporation following addition of U-[$^{13}C_6$]glucose or U-[$^{13}C_3$]lactate would allow for the determination of how much G3P is derived from each pathway. Characterization of each metabolite in the pathway (G6P, F6P, FBP, TCA cycle metabolites, etc.) could help identify blockages in each metabolic pathway. *B*: MS/MS spectrum of [^{12}C]G3P (top panel) and [$^{13}C_3$]G3P (bottom panel) in the liver following treatment with U-[$^{13}C_6$]glucose. The area of G3P m+3 (bottom panel) divided by the sum of the total, following correction for naturally occurring ^{13}C isotopes, gives the percent of G3P derived from glycolysis following the addition of U-[$^{13}C_6$]glucose.

particle sizes, has led to enhanced peak capacity and allows for greater resolution and higher throughput due to reduced run times per sample. With LC, polarity of the solvent used to drive the sample through the column affects sample retention, as does solvent pH.

Sample ionization needs to occur, after which the mass of the analyte is determined by the mass analyzer as mass-to-charge ratio (*m*/*z*). Electrospray ionization is widely used as it works well with most metabolites and provides no matrix interference to the mass analyzer. Atmospheric pressure chemical ionization is slightly less sensitive but works well with nonpolar compounds such as lipids. Matrix-assisted laser desorption/ionization is very useful for complex samples and is highly sensitive. Matrix-assisted laser desorption/ionization is preferred for higher mass compounds, although this is generally not an issue for metabolites. The primary disadvantage is background interference, particularly with lower molecular weight compounds.

Several mass analyzers can be coupled to LC and optimized for the analytical strategy used. The most common mass analyzers are the quadrupole, time of flight (TOF), and ion trap analyzers. Due to their relatively low cost, quadrupole analyzers are widely used. Triple quadrupole (QQQ) analyzers, in which three quadropoles are combined in succession, allow for MS/MS, or further fragmentation, of ions during analysis. QQQs are capable of multiple reaction monitoring for specific detection and quantification of selected metabolites. TOF analyzers accelerate ions and then measure the velocity, or the time it takes to travel down a flight tube, to determine the *m*/*z*. TOF analyzers have high mass accuracy, are highly sensitive, and acquire data quickly. TOF analyzers can be coupled with a quadrupole (Q-TOF). Q-TOFs are well suited for metabolomics experiments. They have very high mass accuracy and sensitivity and can analyze a wide array of metabolites. Ion trap analyzers are similar to quadrupoles in that they can also focus on particular ions and are relatively low cost. They can trap ions of interest and accumulate them for better sensitivity, or they can trap and fragment a specific ion multiple times,

Table 2—MS methods of detection for metabolites of interest in diabetes research

Metabolites/pathways	Method	Conditions/comments	References
Acyl-carnitines	LC-MS	+ ion mode, C18 column, QQQ	(112,113)
Acyl-CoAs	LC-MS	+ ion mode, C18 column, QQQ	(114,115)
Acyl-glycerols	LC-MS	+ ion mode, silica column, QTrap	(116,117)
Amino acids	GC-MS	SIM, EZ:faast Kit (Phenomenex Inc., Torrance, CA)	(118)
Bile acids	LC-MS	− ion mode, C18 column, QQQ	(119)
Cholesterol esters	LC-MS	+ ion mode, silica column, QTrap	(120)
Eicosanoids	LC-MS	− ion mode, chiral column, QQQ	(121,122)
Fatty acids	GC-MS	Derivatize with FAME or PFB bromide	(123)
Glycerophospholipids	LC-MS	+ and – ion mode, silica column, QQQ or QTrap	(124)
Glycolysis	LC-MS	+ ion mode, C18 column, QQQ − ion mode, HILIC column, TOF	(125) (16)
Lipid profiling	LC-MS	+ and – ion mode, QQQ or TripleTOF QTrap	(126) (127) Lipidmaps.org/protocols
Nucleotides	LC-MS	− ion mode, ODS column, QQQ or Ion Trap	(128)
Organic acids	LC-MS GC-MS	+ ion mode, C18 column, QQQ − ion mode, HILIC column, TOF SIM, derivatize with MTBSTFA	(125) (16) (129)
Organic cofactors	LC-MS	+ ion mode, HILIC column, QQQ − ion mode, HILIC column, TOF	(130) (16)
Oxidized amino acids	LC-MS	+ ion mode, C18 column, QQQ	(131)
Oxidized lipids	LC-MS	− ion mode, C18 column, QQQ or QTrap	(132,133)
Pentose phosphate	LC-MS	+ ion mode, C18 column, QQQ	(130)
Sphingolipids	LC-MS	+ ion mode, C18, amino and silica columns, QQQ or QTrap	(134)
Steroid hormones	LC-MS	+ ion mode, C18 column, QQQ	(135)
Sterols	LC-MS	+ ion mode, C18 column, QTrap	(136)
Urea cycle	GC-MS GC-MS and LC-MS	Derivatize with BSTFA/TMCS Review of methods	(137) (138)
Uremic solutes	LC-MS	+ ion mode, C18 column, QQQ	(8)

BSTFA/TMCS, N,O-bis(trimethylsilyl)trifluoroacetamide/trimethylchlorosilane; FAME, fatty acid methyl ester; MTBSTFA, N-(tert-butyldimethylsilyl)-N-methyltrifluoroacetamide; ODS, octadecylsilyl; PFB, pentafluorobenzyl; QTrap, quadrupole ion trap; SIM, selected ion monitoring.

which is referred to as MS^n. One main limitation of ion traps is their inability to do multiple reaction monitoring measurements. Newer techniques such as Fourier transform ion cyclotron resonance have the highest degree of mass accuracy of 100,000, the best accuracy (<1 ppm mass error), and have MS/MS and MS^n capabilities but are limited by expense.

Data Processing

Peak Processing and Inclusion

Most MS data must be initially processed with proprietary software from the manufacturer of the analyzer. Freely available programs are capable of peak detection and integration, although they may lack the ability to read all file types. Regardless of software, data processing depends upon the type of analysis used. For targeted analyses, processing is generally straightforward as there are often isotope-labeled or authentic standards used for validation. For untargeted metabolomics, however, the software needs to be capable of peak selection, evaluation, and relative quantification. For identifying peaks, several libraries exist for searching against generated MS and MS/MS spectra (18–21). Before any data preprocessing occurs, the instrument operator will need to examine the instrument-associated quality assurance/quality control (QA/QC) using data acquisition and visualization tools associated with the instrument. This assessment should include tuning parameters, evaluating the calibration curves both in matrix-containing and matrix-free samples, examining for retention time shifts, and comparing the ratio of quantifier to qualifier ions in the analyte with those obtained using standards. Ideally, the available laboratory information management system should integrate these QA/QC parameters and serve as an automated filter before passing the data for subsequent analysis.

The downstream data analysis should be further examined for reproducibility using different plots (boxplots, histograms, or heat maps) that will check for outlier

samples, samples with high degree of missingness, and data points with low signal-to-noise ratio. Only those samples that pass these stringent QA/QC criteria should be further processed.

For MS data, the mode of acquisition (targeted vs. untargeted) will impact the imputation procedure used. For targeted acquisition, it is not usually the case that the data will have missing values, but for untargeted acquisition, missing values are a common feature (8). In that case, a predefined threshold level is appropriate. The empirical threshold for missing values is mostly based on sample size, and the goal is to diminish false-positive identification for differentially expressed metabolites based on skewness caused by imputation. For example, for cell line studies, the number of replicates is usually small per experimental condition (<10). Hence, the threshold needs to be stringent (~20%). On the other hand, a large study could have >200 samples per condition of interest (e.g., clinical study with healthy vs. diabetic plasma). In this scenario, one could consider imputing ~50% given the large number of data that are available. For studies with dozens of samples but not hundreds, the threshold should be calibrated between 20 and 50%. Missing values can be imputed either at the minimum detection level or through imputation using a nearest neighbor (KNN) procedure, where the number of nearest neighbors should be selected judiciously depending upon the number of samples available in the study (pamr package in the R programming language) (22). Depending on the study design, several different approaches are available, ranging from simple median centering, to centering and scaling based on the values of internally spiked standards, to using more advanced fixed-effects ANOVA procedures that use factors, data platform, batch information, and ionization mode. The best strategy is to use different thresholds and different imputation strategies and assess the sensitivity of the results obtained; the downside is that this is a labor-intensive process. To demonstrate the importance of selecting the correct imputation method, four different imputation strategies were applied to a data set of control and diabetic urine samples for amino acid analysis, of which 5 out of 27 samples had missing values for methionine (K.M. Sas and S. Pennathur, unpublished data; Fig. 3). The first strategy is KNN with three nearest neighbors, the second with five nearest neighbors, the third imputation by the mean of the metabolite, and the fourth by the median of the metabolite. As shown, due to the presence of outliers, mean imputation alters the distribution of data, median imputation slightly compresses the variance, KNN 5 impacts the overall median of the metabolite, and KNN 3 preserves the distributional characteristics observed in the original data (with the missing values). Hence, KNN 3 is the one that does not "perturb" the data architecture in this example. If the missing values threshold is set low (e.g., 20%), different strategies may produce approximately similar results (Table 3). But as the threshold increases, the results would be different. Importantly, even if one resorts to robust nonparametric tests to assess differentials (e.g., rank sum test), the distribution of the data matters, as shown in Table 3. Therefore, careful assessment of all of these factors needs to be

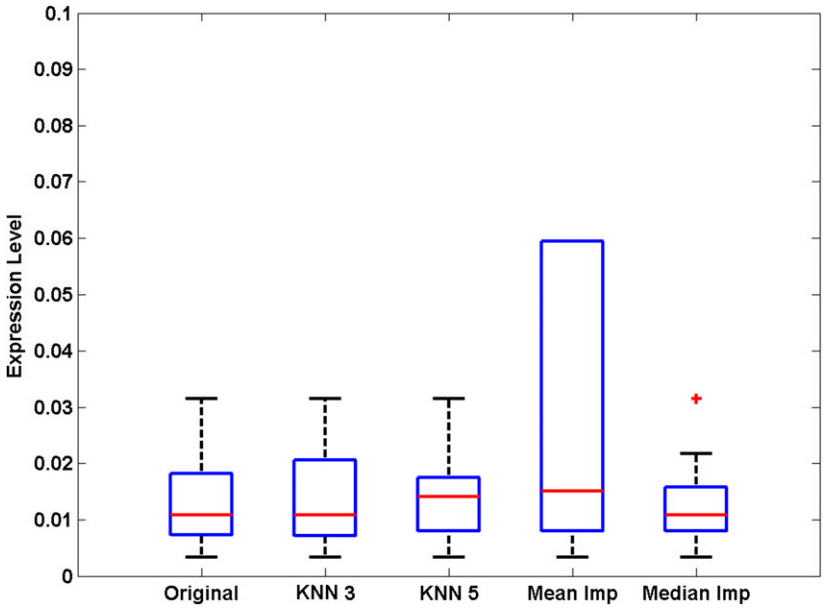

Figure 3—Comparison of imputation methods for missing values. Methionine concentrations in urine were determined by GC-MS, and 5 out of 27 control subjects had values below the limit of detection. Data were $\log_2$ transformed and analyzed using different imputation methods for the missing values (three nearest neighbors [KNN 3], five nearest neighbors [KNN 5], metabolite mean value [Mean Imp], or metabolite median value [Median Imp]). The median and variance for each imputation method is shown. KNN 3 had the smallest effect on data distribution. +, outlier.

Table 3—Comparison of imputation methods for missing values

Imputation method	Student *t* test FDR *P* value	Fold change	Rank sum FDR *P* value
KNN 3	0.2400	1.7702	0.0203
KNN 5	0.3164	1.7777	0.0241
Mean	0.9733	2.1206	0.4234
Median	0.2393	1.7348	0.0037

Comparison of imputation methods for $\log_2$-transformed methionine concentrations in urine of control and diabetic patients. Missing values present for 5 out of 27 control samples.

considered in consultation with the statistician with expertise in metabolomics before choosing the appropriate procedure for normalization.

Midlevel Analyses

These types of analysis involve identification of differentially expressed metabolites, model building for classificatory or survival analysis purposes, dimension reduction for extracting broad patterns from the data, and identification of groups of samples and/or metabolites.

Specifically, differentially abundant compounds across two classes can be identified using parametric (Student *t* tests) and nonparametric (rank sum) tests, whereas for multiple classes, ANOVA models can be used. The latter models in addition to the key treatment factors being tested also allow for the incorporation of key covariate information, such as clinical (stage of the disease and indices of physiological impairments), as well as demographic and health habits of the subjects (e.g., age, race, sex, education, smoking, and alcohol consumption in the case of humans, and strain, sex, and housing conditions in the case of mice).

Given the large number of markers that are likely to be identified as significantly different between groups, as well as the number of conditions and differences in any given experiment, the possibility of type I error (false positives) can occur due to multiple comparisons. Hence, family-wise error rate methods (23,24) and false discovery rate (FDR) methods (23–25) should be used as a first filter to reduce or eliminate false positives. Following this, additional filters (e.g., fold change) can be used based on the platform (targeted vs. untargeted) and other biological considerations. As we have previously shown, untargeted data are more variable and therefore less reliable, owing to a greater need for additional filters, and ultimately, follow-up using a targeted approach is required (8). It is important to keep in mind that as the sample size increases, metabolites with even small fold changes may be considered differentially expressed. Therefore, it would be important to adjust the FDR threshold before other criteria are introduced.

In many studies, classificatory and/or prognostic models have to be built. Such models are important for delineating metabolomic signatures associated with clinical outcomes, including disease/normal status, and clinical characteristics. For categorical outcomes (e.g., disease/normal status), there are several standard models in the machine learning literature that can be used, including logistic regression, random forests, and support vector machines (26), whereas for outcomes capturing event times (e.g., disease recurrence or survival), Cox proportional hazards models can be used (27). An important aspect of this modeling is to enforce sparsity through penalization (e.g., lasso or group lasso penalties) that leads to more parsimonious models that exhibit good theoretical properties in terms of inference and predictive ability (26). In the case of structured penalties (e.g., group lasso), one can impose a priori biological information, such as pathway structure. The performance of these classificatory and prognostic models can be assessed through K-fold cross-validated error rates, and through receiver-operator characteristic curve, and the area under the curve can be used as an overall measure of model fit. The significance of the area under the curve metric for each fitted model can be assessed through the Mann-Whitney *U* test and it can also be used to select between competing models (24).

Finally, depending on project needs, other analyses to gain insight into global properties of the available data need to be undertaken. These include dimension reduction techniques, such as principal components analysis and penalized (for sparsity) variants for obtaining more robust low-dimensional representations of the samples and/or the metabolites (see Guo et al. [28] and references therein), clustering of samples and/or metabolites into groups using a wide range of algorithms (hierarchical, model-based, partition such as k-means and robust variants and graph-based ones such as normalized cuts) (26). In addition, enhanced visualization capabilities by mapping results into pathways have proved a useful task (see PATHWAYS MAPPING AND ENRICHMENT-BASED METHODS).

Methods and Tools for Bioinformatics Analysis of Metabolomics Data

Pathways Mapping and Enrichment-Based Methods

As the metabolomics data sets generated by the analytical methods described above are becoming increasingly large and complex, there is a growing need for the computational data analysis and visualization tools that would help interpret experimentally observed changes and put them in relevant biological or disease context. One widely used approach to interpreting metabolomics data relies on mapping them onto metabolic pathways. Kyoto Encyclopedia of Genes and Genomes (KEGG) (29) and BioCyc (30,31) are the most widely used databases of this kind and contain information about metabolic pathways, metabolites, metabolic reactions, and enzymes and the genes that encode them. The data contained in these databases were generated via genome-based metabolic reconstructions combined with extensive literature searches and expert curation. Subsequently, a number of more detailed organism-specific

metabolic reconstructions have been developed (32–36). In addition to detailed information about metabolic pathway topology and individual components of pathways, some of these include information about subcellular compartments where the metabolic reactions occur (36) and describe metabolic enzyme complexes and transporters (33).

There are a number of bioinformatics tools for pathway mapping and visualization that make use of these data sets. Some of these tools use the static pathways charts (37,38), whereas others make them interactive (39,40). One such tool, Paintomics, can load metabolite and gene expression measurements and visualize them over KEGG pathway maps (38). A more interactive tool, Visualization and Analysis of Networks containing Experimental Data (VANTED), has been developed for exploration of experimental metabolomics data in the context of metabolic pathways, originally from plants (39,40). However, it can be used for any data set; users can load KEGG maps or build their own pathways. Another metabolomics pathways analysis tool, Metabolomics Pathway Analysis (MetPA) (41), that is now part of the comprehensive data analysis package MetaboAnalyst (42), in addition to pathway mapping, calculates pathway impact based on a normalized centrality measure of a given compound relative to the other compounds.

One of the limitations of visualizing data over pathway charts stems from the fact that metabolites are often involved in multiple pathways. In order to understand the overall effect of the altered level of a given metabolite, the user has to go through multiple pathways and understand the connections between them. An alternative to this approach is building a network of genes/metabolites where each node is unique and nodes from multiple pathways can be linked together. Such networks provide an easy way to connect multiple pathways and build gene/compound centric maps enabling quick data exploration and logical, well-informed hypothesis generation. MetScape (43) is an example of a tool that uses this approach. MetScape is a plugin for the widely used network visualization program Cytoscape (44). It allows users to upload a list of metabolites with experimentally determined concentrations and map them to reactions, genes, and pathways. It also supports identification of enriched biological pathways from expression profiling data, building the networks of genes and metabolites involved in these pathways, and allows users to visualize the changes in the gene/metabolite data over time/experimental conditions. MetScape uses human metabolic pathways, although it can also map mouse and rat genes to their human homologs.

To illustrate the utility of MetScape for mapping metabolomics data and merging them with other omics data, we loaded the list of metabolites detected in plasma samples of individuals with and without incident type 2 diabetes from the Framingham Heart Study, reported by Wang et al. (45). Among the most significant metabolites that had higher concentrations at baseline between case and control subjects were three branched-chain amino acids (BCAAs), leucine ($P = 0.0005$), isoleucine ($P = 0.0001$), and valine ($P = 0.001$), and three aromatic amino acids, phenylalanine ($P < 0.0001$), tyrosine ($P < 0.0001$), and tryptophan ($P = 0.003$). Figure 4 shows the MetScape network for the valine, leucine, and isoleucine degradation pathway. To complement the metabolomics data, we also loaded gene expression data and the list of pathways that are differentially expressed in human diabetic muscle compared with healthy controls (46). This tool supports simultaneous

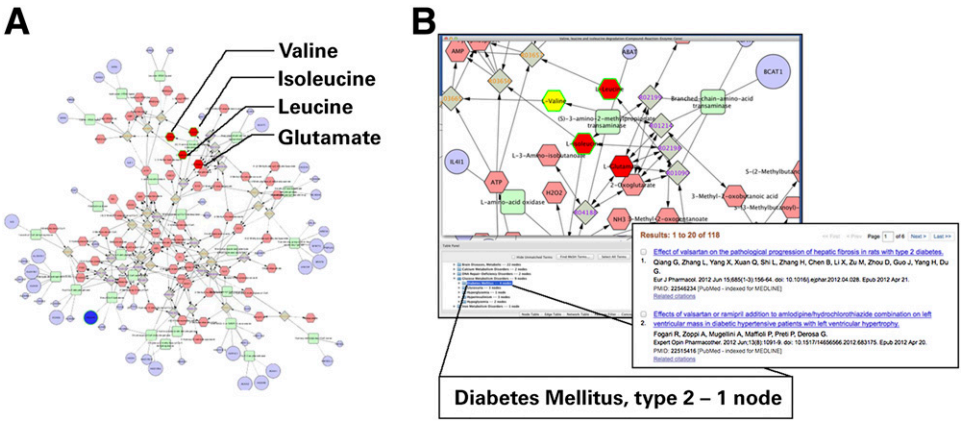

Figure 4— MetScape network for valine, leucine, and isoleucine degradation pathway. *A*: The metabolites are shown as pink hexagons. The metabolites that were experimentally measured by Wang et al. (45) are shown in red. Green border shows significant metabolites. Gene expression data from Mootha et al. (46) were superimposed on the metabolic network. Gene nodes are blue; the size of the node represents the direction of the change. Dark blue color is reserved for genes from enriched pathways. Gray nodes represent the reactions, and green nodes are enzymes. *B*: A zoomed in view of the same network, where MetDisease was used to annotate the metabolites with MeSH disease terms. The lower part of the figure shows the portion of MeSH tree. When diabetes mellitus is selected, the related metabolites (in this case valine) are selected. Additional information can be obtained by right-clicking on metabolite node. The insert on the right shows the list of publications that support the connection between the metabolite and the MeSH term.

analysis of gene expression and metabolomics data that can facilitate generating new hypotheses and prioritization of genes/compounds for targeted follow-up studies. It is worth pointing out that MetScape networks tend to get large, especially when gene expression data are included, introducing the so called "hairball" effect and making it difficult to comprehend the underlying perturbed pathways. MetScape has built in powerful filtering mechanisms that allow users to focus on the relevant parts of networks. When a network is built, the tool displays the list of the pathways that involve the nodes in that network. Users can select one or more pathways and respective nodes will be highlighted. MetScape also provides an easy way to create manageable subnetworks that can be interrogated further. Additional options for interrogating large networks include Concept filter (if a file containing the list of concepts enriched based on gene expression data have been supplied by the user) as well as built-in Cytoscape features, such as Group Attributes layout.

Thus pathway mapping and network analysis tools can help shed light on the molecular mechanisms underlying many complex diseases, including diabetes, especially when linked with other omics data. Experience with other omics data (especially expression profiling) shows that it is also very important to have a measure of significance for the pathways that are involved in the observed changes. A large number of methods and tools have been developed for performing what is often called an enrichment analysis on gene expression data (recently reviewed in Khatri et al. [47]). The goal of enrichment analysis is to evaluate what predefined sets of genes (e.g., pathways) are enriched with differentially expressed genes from a given experiment (e.g., microarray). Several recently published tools (Metabolites Biological Role [MBRole] [48], Metabolite Set Enrichment Analysis [MSEA] [49,50], and 3Omics [51]) attempted to extend this approach to metabolomics data. Although these tools can certainly be quite useful in some situations, it is important to keep in mind that one of the reasons why these methods perform well in analysis of gene expression data is that each transcriptomic experiment can measure tens of thousands of genes. In contrast, current metabolomics techniques at best can identify a few hundred metabolites. As a result, metabolite set enrichment testing has considerably lower statistical power, which is further complicated by metabolites appearing in multiple metabolic pathways. One way to address this issue is to incorporate network topology information (e.g., from KEGG) into the pathway enrichment procedure that leads to higher statistical power (52,53). Further, the advantages of network-based methods pathway enrichment methods are discussed in Mitrea et al. (54). Scarcity of metabolite annotations further compounds this problem. More recent efforts include attempts to incorporate the unknown spectral features into enrichment analysis and network building (55). The advantage of this approach is twofold: it has the potential to identify the unknown compounds and boost the statistical power at the same time.

Going Beyond Pathways

Pathways analysis and visualization have become an integral part of biological interpretation of metabolomics experiments. Although the pathway databases provide carefully curated, high-quality data that cover the majority of primary metabolites, the coverage of lipids, secondary, and volatile metabolites is significantly lower (56), resulting in relatively low overall coverage of experimentally identified metabolites. Additional factors contributing to this problem include the presence of metabolites from different organisms (e.g., presence of bacterial metabolites in human samples originating from microbiome), drug metabolites, and compounds of environmental origin.

MetaMapp tool attempts to overcome this problem by combining the biochemical reactions from KEGG with Tanimoto chemical and National Institute of Standards and Technology mass spectral similarity scores (56). Efforts have been made to extend metabolite annotation coverage beyond pathways using Medical Subject Headings (MeSH) to link them to publications (http://metab2mesh.ncibi.org) (57). Figure 4*B* shows the annotations generated by the Cytoscape plugin MetDisease (http://apps.cytoscape.org/apps/metdisease) that uses the Metab2MeSH data set for the BCAA degradation network (58).

In summary, the development of methods and tools for analysis and visualization of metabolomics data remains an active area of research.

METABOLOMICS AND DIABETES RESEARCH

Identification of Putative Biomarkers

Lifestyle alterations such as diet and exercise can reduce the incidence of diabetes (59,60). Therefore, it is increasingly important to identify early biomarkers that predict risk of development. Recent metabolomics studies have identified two main classes of metabolites that have shown promise as biomarkers of diabetes risk, namely, amino acids and lipids.

Alterations in amino acid levels with obesity have been known for decades (61). However, recent studies have identified amino acids as potent predictors of diabetes and validated them in large, well-characterized cohorts. Using the Framingham Heart Study Offspring Cohort, Wang et al. (45) used a targeted LC-MS/MS approach to examine small metabolites such as amino acids, urea cycle metabolites, and nucleotide metabolites. Elevated levels of the BCAAs (isoleucine, leucine, and valine) as well as some aromatic amino acids (tyrosine and phenylalanine) were able to predict risk up to 12 years prior to onset of diabetes, particularly when three of the metabolites (isoleucine, phenylalanine, and tyrosine) were incorporated into a model together. Other groups have also identified BCAAs and aromatic amino acids as predictors of type 2 diabetes in both humans and animal models (62–65). Further work with the Framingham cohort identified 2-aminoadipic acid (2-AAA) as an independent biomarker for risk development and highlighted the role of 2-AAA as an insulin secretagogue (66). 2-AAA is an intermediary metabolite of lysine degradation and has previously been

shown to be increased by diabetes and renal failure (67,68) and has been suggested to be a biomarker of oxidative stress (69,70). In addition to the Framingham Study, metabolites were identified in plasma from patients in the Cooperative Health Research in the Region of Augsburg (KORA) cohort, followed by validation in the European Prospective Investigation into Cancer and Nutrition (EPIC)-Potsdam cohort in an attempt to identify biomarkers of prediabetes (71). Using a targeted LC-MS/MS approach with a commercially available kit that measures metabolites across five compound classes, two metabolites (glycine and lysophosphatidylcholine [18:2]) were identified as biomarkers of impaired glucose tolerance and type 2 diabetes. Reduced concentrations of glycine as a predictor of type 2 diabetes has been found in additional studies as well (62,72).

Zhang et al. (73) used untargeted metabolomics to explore underlying mechanisms of disease progression and treatment response in order to identify novel metabolite biomarkers of progressive murine diabetic nephropathy. In total, 56 features showed up- or downregulation by more than twofold in the diabetic animals. Of the 56 molecular features, 32 were identified by database searching. Rosiglitazone treatment reversed 9 of these 32 compounds (including indoxyl sulfate) back to baseline, and these may therefore serve as potential biomarkers for response to treatment and reversal of rodent diabetic nephropathy phenotype. Interestingly, a study from Barreto et al. (74) reported that serum indoxyl sulfate correlates inversely with renal function and might have a direct relationship with aortic calcification and pulse wave velocity in patients with chronic kidney disease. Niewczas et al. (8) studied plasma metabolomics profiles as determinants of progression to end-stage renal disease (ESRD) in patients with type 2 diabetes. This nested case-control study evaluated 40 case subjects that progressed to ESRD during 8–12 years of follow-up and 40 control subjects who remained alive without ESRD from the Joslin Kidney Study cohort. The metabolomics platform identified 16 uremic solutes that were already elevated in the baseline plasma of case subjects years before ESRD developed. Essential amino acids and their derivatives were significantly depleted in the case subjects, whereas certain amino acid–derived acylcarnitines were increased.

Dyslipidemia is an independent risk factor for type 2 diabetes (75,76). However, this includes total lipid or lipid class (i.e., triacylglycerols or HDL) levels. Several recent studies have identified signatures of particular lipids or patterns in lipid classes to be predictive of diabetes onset. Rhee et al. (77) used a targeted LC-MS/MS approach with plasma from the Framingham Heart Study cohort to identify that saturated or monounsaturated fatty acids of lower carbon number were associated with an increased risk of type 2 diabetes, whereas longer carbon chains with increased double bond content (polyunsaturated fatty acids) conveyed a decreased risk of type 2 diabetes. Although concerned primarily with triacylglycerols, this association was true across several lipid classes. Exercise- and diet-induced weight loss (78), which are known to decrease risk of type 2 diabetes, resulted in a change in triacylglycerol pattern to support this finding; that is, triacylglycerol composition changed to be enriched in unsaturated, long-carbon side chains. Findings from targeted LC-MS/MS studies with the KORA and EPIC-Potsdam cohorts also support this result, particularly in regards to degree of saturation (62,71). In addition to predicting risk of developing diabetes, the degree of lipid saturation has been linked to diabetes complications. NMR metabonomics with baseline serum from subjects in the Finnish Diabetic Nephropathy (FinnDiane) Study linked high levels of saturated fatty acids in serum to accelerated progression of kidney disease in patients with type 1 diabetes (79).

One important limitation of many of the current studies is the depiction of metabolite levels by quartiles. Although appropriate for categorizing patients for biomarker analysis, actual metabolite concentrations or thresholds need to be set that predict risk before these metabolites can be used clinically. An alternative to setting threshold values may be to determine metabolic phenotype by assessing metabolic responses in an individual before and after an oral glucose tolerance test (80). Although this is an exciting, personalized alternative, the potential use of this concept needs further testing.

Determining Pathogenesis

The biomarker identification studies discussed earlier provide insight into the pathogenesis of diabetes, as these early changes highlight pathways such as amino acid metabolism, specifically catabolism of BCAAs. The increase in BCAAs has been suggested to impact insulin sensitivity through the mammalian target of rapamycin complex (mTORC), as BCAAs activate mTORC1 and the downstream target ribosomal protein S6 kinase 1 (S6K1) (81). S6K1 can then impact insulin sensitivity through its repression of signaling through insulin receptor substrate 1 (82). Additionally, catabolism of BCAAs can provide intermediates for the TCA cycle, potentially driving energy production (81). The idea that TCA cycle flux is altered in diabetes has been supported in other metabolomics studies in rats and mice (83,84).

To determine the effect of insulin treatment on diabetes-associated metabolic changes, Dutta et al. (85) examined differences in the plasma metabolome of controls, type 1 diabetic patients treated with insulin, and the same type 1 diabetic patients following 8-h insulin withdrawal. Untargeted metabolomics identified that whereas many of the metabolites associated with insulin deficiency were normalized with insulin treatment, not all metabolites were restored to control levels. This suggests that the diabetes-mediated metabolic alterations are not due to substrate availability alone but an underlying mechanism such as metabolic reprogramming. Additionally, pathway enrichment analysis and integration of metabolomics data with transcriptomics data in this study identified

new pathways affected by insulin secretion, including several that lead to vascular complications. Although still in the early stages, identification of these pathways will allow for further investigation into the pathogenesis of one of the most life-threatening complications of diabetes.

A recent study by Sharma et al. (86) established that diabetic kidney disease is characterized by mitochondrial dysfunction. In this study, the urine metabolome from healthy and diabetic patients, with and without diabetic kidney disease, was assessed using targeted GC-MS. The comparison of results between subsets of diabetic patients based upon progression to kidney disease allows for greater in-depth determination of pathways involved in disease progression. Several of the metabolites linked to kidney disease were water-soluble organic anions, leading the authors to investigate expression of several organic anion transporters (OATs). OAT1 and OAT3 were reduced in renal biopsies from patients with diabetic kidney disease. Improper transport due to diminished OAT expression could result in increased intermediates of the TCA cycle, a finding that has previously been reported (83,84). Additionally, these transporters are important for energy metabolism (87,88). Nearly all of the metabolites (12 of 13) separating the diabetic groups based on kidney disease were associated with mitochondria, leading the authors to further examine mitochondrial function. Using kidney biopsies and urinary exosomes, the authors found evidence of reduced mitochondrial biogenesis. Confirming this finding, reduced mitochondrial biogenesis has been shown in a mouse model of diabetes, and this dysfunction was rescued by augmentation of AMPK activity (89). These studies provide insight into the mechanisms driving one of the primary complications of diabetes.

Using metabolomics techniques, recent studies have investigated the effects of diabetes on atherosclerotic lesion cells such as macrophages, which may account for increased atherosclerotic risk in diabetes. Fatty acids can exert inflammatory effects in macrophages, which could contribute to inflammation in the setting of diabetes-accelerated atherosclerosis, and possibly other complications (90). After entering the cell, fatty acids are thio-esterified into their acyl-CoA derivatives, catalyzed by long-chain acyl-CoA synthetases (ACSLs). Kanter et al. (91) demonstrated that monocytes from humans and mice with type 1 diabetes also exhibit increased ACSL1. Furthermore, myeloid-selective deletion of ACSL1 protected monocytes and macrophages from the inflammatory effects of diabetes. Strikingly, myeloid-selective deletion of ACSL1, but not overexpression of GLUT-1 (92), prevented accelerated atherosclerosis in diabetic mice without affecting lesions in nondiabetic mice (91). These observations indicate that ACSL1-derived lipids, but not glucose, play a critical role by promoting the inflammatory phenotype of macrophages associated with diabetes.

Discoveries made from biomarker studies and mechanistic pathway analyses can provide new treatment targets for therapeutic intervention (93). This idea is highlighted by a recent study testing the AMPK activator COH-SR4 (94). Although not using metabolomics directly, Figarola et al. (94) identified that COH-SR4 was able to rescue many of the abnormalities associated with metabolic syndrome in an animal model of obesity, including reduction of several metabolic enzymes in pathways previously found to be altered by metabolomics. Additionally, recent exciting discoveries link intestinal microbiota metabolism of dietary-derived saturated fats to cardiovascular disease risk, highlighting these as attractive potential therapeutic targets for complications of obesity/diabetes (95–97).

Systems Genetics: Linking Genetic Variance to Clinical Outcomes via Metabolomics

Genetics of complex traits has been very effective in defining statistically significant risk loci by mapping genome-wide genetic variance in large cohorts onto clinical traits. However, this approach is agnostic concerning the molecular mechanism and intermediate regulatory cascades responsible for clinical disease manifestation. To close the molecular knowledge gap, systems genetics uses as a key tool linkage and association methods of genotype information with an intermediary molecular trait of interest, e.g., metabolite level (metabolic quantitative trait locus), to determine the impact of genetic variance on the trait in question (98–100). Subsequent mapping of metabolic quantitative trait loci onto association studies with clinical outcomes aids to identify the molecular impact of genetic variance associated with a disease phenotype. Recent studies identified gene-metabolite dependencies by integrating genome-wide association studies (GWAS) with metabolomics data sets (mGWAS) supporting the cross-omics strategy (101–104). Combining these studies in a systems approach (mGWAS reviewed in Adamski [105] and Adamski and Suhre [106]) synergistically expands insight into disease pathogenesis and strengthens the associations with disease phenotype. One such study applied targeted metabolomics, transcriptome analysis, and whole genome sequencing to liver samples from a diabetes-resistant C57BL/6 *leptin*$^{ob/ob}$ and diabetes-susceptible BTBR *leptin*$^{ob/ob}$ mouse strain (101). The authors showed that groups of liver metabolites significantly associate with distinct chromosomal regions. Suhre et al. (107) reported on the genetic association of urinary metabolites of 862 male participants from the epidemiological Study of Health in Pomerania (SHIP). Independent validation in an additional 2,031 samples (1,039 independent SHIP and 992 samples from the KORA study) revealed consistent genome-wide significant loci tagging SLC7A9 and NAT2, which have been already associated with CKD and drug-induced liver toxicity, respectively (108–110). Kettunen et al. (111) used NMR spectroscopy–based detection of serum metabolites of over 8,000 genotyped Finnish individuals and were able to ascertain a high degree of heritability for metabolic phenotypes, ranging from 40% up to 60%. Although these represent early-stage discovery, further

cross-omics data integration would be even more informative and a rich discovery platform for future research.

CONCLUSIONS AND FUTURE PERSPECTIVES

Metabolomics is an integral part for understanding disease processes as it measures functional outputs of a cell, tissue, or organ. Although still a relatively new field, significant strides in data collection and interpretation tools have allowed for a rapid expansion of metabolomics in the past few years. Several limitations in these areas still exist, such as lack of a platform that detects all metabolites simultaneously, an incomplete metabolome, lack of metabolite annotations in search databases, and low statistical power for enrichment analyses. Despite these limitations, metabolomics is being widely used in the field of diabetes and its complications, particularly in the identification of disease biomarkers and novel therapeutic interventions. Using the information garnered in the biomarker investigations, future research should shed more light on disease pathogenesis and explore new treatment options. As the analytical and bioinformatics tools continue to become more developed, integration of metabolomics data with other omics data sets will allow for a greater understanding of disease processes and ultimately allow for personalized medicine to become the mainstream standard of care.

Funding. This work is supported in part by grants from the National Institutes of Health (DK-094292, DK-089503, DK-082841, DK-081943, and DK-097153 to S.P.) and from the National Center for Advancing Translational Sciences (2UL1-TR-000433).

Duality of Interest. No potential conflicts of interest relevant to this article were reported.

Author Contributions. K.M.S., A.K., G.M., and S.P. wrote, reviewed, and edited the manuscript. S.P. is the guarantor of this work and, as such, had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

References

1. Centers for Disease Control and Prevention. *National Diabetes Fact Sheet: National Estimates and General Information on Diabetes and Prediabetes in the United States*. Atlanta, GA, U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2011

2. Dunn WB, Wilson ID, Nicholls AW, Broadhurst D. The importance of experimental design and QC samples in large-scale and MS-driven untargeted metabolomic studies of humans. Bioanalysis 2012;4:2249–2264

3. Fan TW, Lane AN, Higashi RM, et al. Altered regulation of metabolic pathways in human lung cancer discerned by (13)C stable isotope-resolved metabolomics (SIRM). Mol Cancer 2009;8:41

4. Lane AN, Fan TW, Bousamra M 2nd, Higashi RM, Yan J, Miller DM. Stable isotope-resolved metabolomics (SIRM) in cancer research with clinical application to nonsmall cell lung cancer. OMICS 2011;15:173–182

5. Want EJ, Masson P, Michopoulos F, et al. Global metabolic profiling of animal and human tissues via UPLC-MS. Nat Protoc 2013;8:17–32

6. Keun HC, Athersuch TJ. Nuclear magnetic resonance (NMR)-based metabolomics. Methods Mol Biol 2011;708:321–334

7. Chen T, Cao Y, Zhang Y, et al. Random forest in clinical metabolomics for phenotypic discrimination and biomarker selection. Evid Based Complement Alternat Med 2013;2013:298183

8. Niewczas MA, Sirich TL, Mathew AV, et al. Uremic solutes and risk of end-stage renal disease in type 2 diabetes: metabolomic study. Kidney Int 2014;85:1214–1224

9. Salmons S, Jarvis JC, Mayne CN, et al. Changes in ATP, phosphocreatine, and 16 metabolites in muscle stimulated for up to 96 hours. Am J Physiol 1996;271:C1167–C1171

10. Antoniewicz MR. ^{13}C metabolic flux analysis: optimal design of isotopic labeling experiments. Curr Opin Biotechnol 2013;24:1116–1121

11. Choi J, Antoniewicz MR. Tandem mass spectrometry: a novel approach for metabolic flux analysis. Metab Eng 2011;13:225–233

12. Crown SB, Ahn WS, Antoniewicz MR. Rational design of ^{13}C-labeling experiments for metabolic flux analysis in mammalian cells. BMC Syst Biol 2012;6:43

13. Huang X, Chen YJ, Cho K, Nikolskiy I, Crawford PA, Patti GJ. X13CMS: global tracking of isotopic labels in untargeted metabolomics. Anal Chem 2014;86:1632–1639

14. Lorenz MA, El Azzouny MA, Kennedy RT, Burant CF. Metabolome response to glucose in the β-cell line INS-1 832/13. J Biol Chem 2013;288:10923–10935

15. Dunn WB, Broadhurst D, Begley P, et al.; Human Serum Metabolome (HUSERMET) Consortium. Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. Nat Protoc 2011;6:1060–1083

16. Lorenz MA, Burant CF, Kennedy RT. Reducing time and increasing sensitivity in sample preparation for adherent mammalian cell metabolomics. Anal Chem 2011;83:3406–3414

17. Smolinska A, Blanchet L, Buydens LM, Wijmenga SS. NMR and pattern recognition methods in metabolomics: from data acquisition to biomarker discovery: a review. Anal Chim Acta 2012;750:82–97

18. Horai H, Arita M, Kanaya S, et al. MassBank: a public repository for sharing mass spectral data for life sciences. J Mass Spectrom 2010;45:703–714

19. Smith CA, O'Maille G, Want EJ, et al. METLIN: a metabolite mass spectral database. Ther Drug Monit 2005;27:747–751

20. Sud M, Fahy E, Cotter D, et al. LMSD: LIPID MAPS structure database. Nucleic Acids Res 2007;35:D527–D532

21. Wishart DS, Knox C, Guo AC, et al. HMDB: a knowledgebase for the human metabolome. Nucleic Acids Res 2009;37:D603–D610

22. Putluri N, Shojaie A, Vasu VT, et al. Metabolomic profiling reveals potential markers and bioprocesses altered in bladder cancer progression. Cancer Res 2011;71:7376–7386

23. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B 1995;57:289–300

24. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983;148:839–843

25. Klipper-Aurbach Y, Wasserman M, Braunspiegel-Weintrob N, et al. Mathematical formulae for the prediction of the residual beta cell function during the first two years of disease in children and adolescents with insulin-dependent diabetes mellitus. Med Hypotheses 1995;45:486–490

26. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer series in statistics. New York, Springer-Verlag, 2009

27. Collett D. *Modelling Survival Data in Medical Research*. 2nd ed. Chapman & Hall/CRC texts in statistical science series. Boca Raton, Florida, Chapman & Hall/CRC, 2003

28. Guo J, James G, Levina E, Michailidis G, Zhu J. Principal component analysis with sparse fused loadings. J Comput Graph Stat 2010;19:930–946

29. Kanehisa, M., Goto, S., Hattori, M., et al. From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res 2006;34:D354–D357

30. Romero P, Wagg J, Green ML, Kaiser D, Krummenacker M, Karp PD. Computational prediction of human metabolic pathways from the complete human genome. Genome Biol 2005;6:R2

31. Caspi R, Altman T, Dreher K, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res 2012;40:D742–D753

32. Ma H, Sorokin A, Mazein A, et al. The Edinburgh human metabolic network reconstruction and its functional analysis. Mol Syst Biol 2007;3:135

33. Duarte NC, Becker SA, Jamshidi N, et al. Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proc Natl Acad Sci U S A 2007;104:1777–1782

34. Sigurdsson MI, Jamshidi N, Steingrimsson E, Thiele I, Palsson BØ. A detailed genome-wide reconstruction of mouse metabolism based on human Recon 1. BMC Syst Biol 2010;4:140

35. Thiele I, Swainston N, Fleming RM, et al. A community-driven global reconstruction of human metabolism. Nat Biotechnol 2013;31:419–425

36. Hao T, Ma HW, Zhao XM, Goryanin I. Compartmentalization of the Edinburgh Human Metabolic Network. BMC Bioinformatics 2010;11:393

37. Paley SM, Karp PD. The Pathway Tools cellular overview diagram and Omics Viewer. Nucleic Acids Res 2006;34:3771–3778

38. García-Alcalde F, García-López F, Dopazo J, Conesa A. Paintomics: a web based tool for the joint visualization of transcriptomics and metabolomics data. Bioinformatics 2011;27:137–139

39. Junker BH, Klukas C, Schreiber F. VANTED: a system for advanced data analysis and visualization in the context of biological networks. BMC Bioinformatics 2006;7:109

40. Klukas, C. and F. Schreiber. Integration of -omics data and networks for biomedical research with VANTED. J Integr Bioinform 2010;7:112

41. Xia J, Wishart DS. MetPA: a web-based metabolomics tool for pathway analysis and visualization. Bioinformatics 2010;26:2342–2344

42. Xia, J., Psychogios N, Young N, Wishart DS. MetaboAnalyst: a web server for metabolomic data analysis and interpretation. Nucleic Acids Res 2009;37: W652–W660

43. Karnovsky A, Weymouth T, Hull T, et al. Metscape 2 bioinformatics tool for the analysis and visualization of metabolomics and gene expression data. Bioinformatics 2012;28:373–380

44. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003; 13:2498–2504

45. Wang TJ, Larson MG, Vasan RS, et al. Metabolite profiles and the risk of developing diabetes. Nat Med 2011;17:448–453

46. Mootha VK, Lindgren CM, Eriksson KF, et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet 2003;34:267–273

47. Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. PLOS Comput Biol 2012;8:e1002375

48. Chagoyen M, Pazos F. MBRole: enrichment analysis of metabolomic data. Bioinformatics 2011;27:730–731

49. Xia J, Wishart DS. Metabolomic data processing, analysis, and interpretation using MetaboAnalyst. Curr Protoc Bioinformatics 2011;Chapter 14: Unit 14.10

50. Xia, J. and D.S. Wishart, MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. Nucleic Acids Res 2010; 38:W71–W77

51. Kuo TC, Tian TF, Tseng YJ. 3Omics: a web-based systems biology tool for analysis, integration and visualization of human transcriptomic, proteomic and metabolomic data. BMC Syst Biol 2013;7:64

52. Shojaie A, Michailidis G. Analysis of gene sets based on the underlying regulatory network. J Comput Biol 2009;16:407–426

53. Shojaie A, Michailidis G. Network enrichment analysis in complex experiments. Stat Appl Genet Mol Biol 2010;9:Article 22

54. Mitrea C, Taghavi Z, Bokanizad B, et al. Methods and approaches in the topology-based analysis of biological pathways. Front Physiol 2013;4:278

55. Li S, Park Y, Duraisingham S, et al. Predicting network activity from high throughput metabolomics. PLOS Comput Biol 2013;9:e1003123

56. Barupal DK, Haldiya PK, Wohlgemuth G, et al. MetaMapp: mapping and visualizing metabolomic data by integrating information from biochemical pathways and chemical and mass spectral similarity. BMC Bioinformatics 2012; 13:99

57. Sartor MA, Ade A, Wright Z, et al. Metab2MeSH: annotating compounds with medical subject headings. Bioinformatics 2012;28:1408–1410

58. Duren W, Weymouth T, Hull T, et al. MetDisease–connecting metabolites to diseases via literature. Bioinformatics 2014;30:2239–2241

59. Knowler WC, Barrett-Connor E, Fowler SE, et al.; Diabetes Prevention Program Research Group. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. N Engl J Med 2002;346:393–403

60. Tuomilehto J, Lindström J, Eriksson JG, et al.; Finnish Diabetes Prevention Study Group. Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance. N Engl J Med 2001; 344:1343–1350

61. Felig P, Marliss E, Cahill GF Jr. Plasma amino acid levels and insulin secretion in obesity. N Engl J Med 1969;281:811–816

62. Floegel A, Stefan N, Yu Z, et al. Identification of serum metabolites associated with risk of type 2 diabetes using a targeted metabolomic approach. Diabetes 2013;62:639–648

63. Menni C, Fauman E, Erte I, et al. Biomarkers for type 2 diabetes and impaired fasting glucose using a nontargeted metabolomics approach. Diabetes 2013;62:4270–4276

64. Renner S, Römisch-Margl W, Prehn C, et al. Changing metabolic signatures of amino acids and lipids during the prediabetic period in a pig model with impaired incretin function and reduced β-cell mass. Diabetes 2012;61: 2166–2175

65. Suhre K, Meisinger C, Döring A, et al. Metabolic footprint of diabetes: a multiplatform metabolomics study in an epidemiological setting. PLoS ONE 2010;5:e13953

66. Wang TJ, Ngo D, Psychogios N, et al. 2-Aminoadipic acid is a biomarker for diabetes risk. J Clin Invest 2013;123:4309–4317

67. Sell DR, Strauch CM, Shen W, Monnier VM. Aging, diabetes, and renal failure catalyze the oxidation of lysyl residues to 2-aminoadipic acid in human skin collagen: evidence for metal-catalyzed oxidation mediated by alpha-dicarbonyls. Ann N Y Acad Sci 2008;1126:205–209

68. Wijekoon EP, Skinner C, Brosnan ME, Brosnan JT. Amino acid metabolism in the Zucker diabetic fatty rat: effects of insulin resistance and of type 2 diabetes. Can J Physiol Pharmacol 2004;82:506–514

69. Yuan W, Zhang J, Li S, Edwards JL. Amine metabolomics of hyperglycemic endothelial cells using capillary LC-MS with isobaric tagging. J Proteome Res 2011;10:5242–5250

70. Zeitoun-Ghandour S, Leszczyszyn OI, Blindauer CA, Geier FM, Bundy JG, Stürzenbaum SR. *C. elegans* metallothioneins: response to and defence against ROS toxicity. Mol Biosyst 2011;7:2397–2406

71. Wang-Sattler R, Yu Z, Herder C, et al. Novel biomarkers for pre-diabetes identified by metabolomics. Mol Syst Biol 2012;8:615

72. Fiehn O, Garvey WT, Newman JW, Lok KH, Hoppel CL, Adams SH. Plasma metabolomic profiles reflective of glucose homeostasis in non-diabetic and type 2 diabetic obese African-American women. PLoS ONE 2010;5:e15234

73. Zhang H, Saha J, Byun J, et al. Rosiglitazone reduces renal and plasma markers of oxidative injury and reverses urinary metabolite abnormalities in the amelioration of diabetic nephropathy. Am J Physiol Renal Physiol 2008;295: F1071–F1081

74. Barreto FC, Barreto DV, Liabeuf S, et al.; European Uremic Toxin Work Group (EUTox). Serum indoxyl sulfate is associated with vascular disease and mortality in chronic kidney disease patients. Clin J Am Soc Nephrol 2009;4: 1551–1558

75. D'Agostino RB Jr, Hamman RF, Karter AJ, Mykkanen L, Wagenknecht LE, Haffner SM; Insulin Resistance Atherosclerosis Study Investigators. Cardiovascular disease risk factors predict the development of type 2 diabetes: the insulin resistance atherosclerosis study. Diabetes Care 2004;27:2234–2240

76. Schulze MB, Weikert C, Pischon T, et al. Use of multiple metabolic and genetic markers to improve the prediction of type 2 diabetes: the EPIC-Potsdam Study. Diabetes Care 2009;32:2116–2119
77. Rhee EP, Cheng S, Larson MG, et al. Lipid profiling identifies a triacylglycerol signature of insulin resistance and improves diabetes prediction in humans. J Clin Invest 2011;121:1402–1411
78. Schwab U, Seppänen-Laakso T, Yetukuri L, et al.; GENOBIN Study Group. Triacylglycerol fatty acid composition in diet-induced weight loss in subjects with abnormal glucose metabolism—the GENOBIN study. PLoS ONE 2008;3:e2630
79. Mäkinen VP, Tynkkynen T, Soininen P, et al. Metabolic diversity of progressive kidney disease in 325 patients with type 1 diabetes (the FinnDiane Study). J Proteome Res 2012;11:1782–1790
80. Ho JE, Larson MG, Vasan RS, et al. Metabolite profiles during oral glucose challenge. Diabetes 2013;62:2689–2698
81. O'Connell TM. The complex role of branched chain amino acids in diabetes and cancer. Metabolites 2013;3:931–945
82. Zick Y. Ser/Thr phosphorylation of IRS proteins: a molecular basis for insulin resistance. Sci STKE 2005;2005:pe4
83. Guan M, Xie L, Diao C, et al. Systemic perturbations of key metabolites in diabetic rats during the evolution of diabetes studied by urine metabonomics. PLoS ONE 2013;8:e60409
84. Li M, Wang X, Aa J, et al. GC/TOFMS analysis of metabolites in serum and urine reveals metabolic perturbation of TCA cycle in db/db mice involved in diabetic nephropathy. Am J Physiol Renal Physiol 2013;304:F1317–F1324
85. Dutta T, Chai HS, Ward LE, et al. Concordance of changes in metabolic pathways based on plasma metabolomics and skeletal muscle transcriptomics in type 1 diabetes. Diabetes 2012;61:1004–1016
86. Sharma K, Karl B, Mathew AV, et al. Metabolomics reveals signature of mitochondrial dysfunction in diabetic kidney disease. J Am Soc Nephrol 2013;24:1901–1912
87. Ahn SY, Nigam SK. Toward a systems level understanding of organic anion and other multispecific drug transporters: a remote sensing and signaling hypothesis. Mol Pharmacol 2009;76:481–490
88. Wu W, Dnyanmote AV, Nigam SK. Remote communication through solute carriers and ATP binding cassette drug transporter pathways: an update on the remote sensing and signaling hypothesis. Mol Pharmacol 2011;79:795–805
89. Dugan LL, You YH, Ali SS, et al. AMPK dysregulation promotes diabetes-related reduction of superoxide and mitochondrial function. J Clin Invest 2013;123:4888–4899
90. Hummasti S, Hotamisligil GS. Endoplasmic reticulum stress and inflammation in obesity and diabetes. Circ Res 2010;107:579–591
91. Kanter JE, Kramer F, Barnhart S, et al. Diabetes promotes an inflammatory macrophage phenotype and atherosclerosis through acyl-CoA synthetase 1. Proc Natl Acad Sci U S A 2012;109:E715–E724
92. Nishizawa T, Kanter JE, Kramer F, et al. Testing the role of myeloid cell glucose flux in inflammation and atherosclerosis. Cell Reports 2014;7:356–365
93. Rabinowitz JD, Purdy JG, Vastag L, Shenk T, Koyuncu E. Metabolomics in drug target discovery. Cold Spring Harb Symp Quant Biol 2011;76:235–246
94. Figarola JL, Singhal P, Rahbar S, Gugiu BG, Awasthi S, Singhal SS. COH-SR4 reduces body weight, improves glycemic control and prevents hepatic steatosis in high fat diet-induced obese mice. PLoS ONE 2013;8:e83801
95. Koeth RA, Wang Z, Levison BS, et al. Intestinal microbiota metabolism of L-carnitine, a nutrient in red meat, promotes atherosclerosis. Nat Med 2013;19:576–585
96. Tang WH, Wang Z, Levison BS, et al. Intestinal microbial metabolism of phosphatidylcholine and cardiovascular risk. N Engl J Med 2013;368:1575–1584
97. Wang Z, Klipfell E, Bennett BJ, et al. Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. Nature 2011;472:57–63
98. Emilsson V, Thorleifsson G, Zhang B, et al. Genetics of gene expression and its effect on disease. Nature 2008;452:423–428
99. Ioannidis JP, Thomas G, Daly MJ. Validating, augmenting and refining genome-wide association signals. Nat Rev Genet 2009;10:318–329
100. Sieberts SK, Schadt EE. Moving toward a system genetics view of disease. Mamm Genome 2007;18:389–401
101. Ferrara CT, Wang P, Neto EC, et al. Genetic networks of liver metabolism revealed by integration of metabolic and transcriptional profiling. PLoS Genet 2008;4:e1000034
102. Gieger C, Geistlinger L, Altmaier E, et al. Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. PLoS Genet 2008;4:e1000282
103. Keurentjes JJ, Fu J, de Vos CH, et al. The genetics of plant metabolism. Nat Genet 2006;38:842–849
104. Shah SH, Hauser ER, Bain JR, et al. High heritability of metabolomic profiles in families burdened with premature cardiovascular disease. Mol Syst Biol 2009;5:258
105. Adamski J. Genome-wide association studies with metabolomics. Genome Med 2012;4:34
106. Adamski J, Suhre K. Metabolomics platforms for genome wide association studies—linking the genome to the metabolome. Curr Opin Biotechnol 2013;24:39–47
107. Suhre K, Wallaschofski H, Raffler J, et al. A genome-wide association study of metabolic traits in human urine. Nat Genet 2011;43:565–569
108. Daly AK. Drug-induced liver injury: past, present and future. Pharmacogenomics 2010;11:607–611
109. Köttgen A, Pattaro C, Böger CA, et al. New loci associated with kidney function and chronic kidney disease. Nat Genet 2010;42:376–384
110. Teslovich TM, Musunuru K, Smith AV, et al. Biological, clinical and population relevance of 95 loci for blood lipids. Nature 2010;466:707–713
111. Kettunen J, Tukiainen T, Sarin AP, et al. Genome-wide association study identifies multiple loci influencing human serum metabolite levels. Nat Genet 2012;44:269–276
112. An J, Muoio DM, Shiota M, et al. Hepatic expression of malonyl-CoA decarboxylase reverses muscle, liver and whole-animal insulin resistance. Nat Med 2004;10:268–274
113. Han CY, Umemoto T, Omer M, et al. NADPH oxidase-derived reactive oxygen species increases expression of monocyte chemotactic factor genes in cultured adipocytes. J Biol Chem 2012;287:10379–10393
114. Golej DL, Askari B, Kramer F, et al. Long-chain acyl-CoA synthetase 4 modulates prostaglandin E_2 release from human arterial smooth muscle cells. J Lipid Res 2011;52:782–793
115. Magnes C, Sinner FM, Regittnig W, Pieber TR. LC/MS/MS method for quantitative determination of long-chain fatty acyl-CoAs. Anal Chem 2005;77:2889–2894
116. Krank J, Murphy RC, Barkley RM, Duchoslav E, McAnoy A. Qualitative analysis and quantitative assessment of changes in neutral glycerol lipid molecular species within cells. Methods Enzymol 2007;432:1–20
117. Leiker TJ, Barkley RM, Murphy RC. Analysis of diacylglycerol molecular species in cellular lipid extracts by normal-phase LC-electrospray mass spectrometry. Int J Mass Spectrom 2011;305:103–109
118. Kugler F, Graneis S, Schreiter PP, Stintzing FC, Carle R. Determination of free amino compounds in betalainic fruits and vegetables by gas chromatography with flame ionization and mass spectrometric detection. J Agric Food Chem 2006;54:4311–4318
119. Xie G, Zhong W, Li H, et al. Alteration of bile acid metabolism in the rat induced by chronic ethanol consumption. FASEB J 2013;27:3583–3593
120. Murphy RC, Leiker TJ, Barkley RM. Glycerolipid and cholesterol ester analyses in biological samples by mass spectrometry. Biochim Biophys Acta 2011;1811:776–783
121. Lee SH, Williams MV, DuBois RN, Blair IA. Targeted lipidomics using electron capture atmospheric pressure chemical ionization mass spectrometry. Rapid Commun Mass Spectrom 2003;17:2168–2176
122. Schneider C, Yu Z, Boeglin WE, Zheng Y, Brash AR. Enantiomeric separation of hydroxy and hydroperoxy eicosanoids by chiral column chromatography. Methods Enzymol 2007;433:145–157

123. Quehenberger O, Armando AM, Dennis EA. High sensitivity quantitative lipidomics analysis of fatty acids in biological samples by gas chromatography-mass spectrometry. Biochim Biophys Acta 2011;1811:648–656
124. Ivanova PT, Milne SB, Byrne MO, Xiang Y, Brown HA. Glycerophospholipid identification and quantitation by electrospray ionization mass spectrometry. Methods Enzymol 2007;432:21–57
125. Lemons JM, Feng XJ, Bennett BD, et al. Quiescent fibroblasts exhibit high metabolic activity. PLoS Biol 2010;8:e1000514
126. Han X, Yang K, Gross RW. Multi-dimensional mass spectrometry-based shotgun lipidomics and novel strategies for lipidomic analyses. Mass Spectrom Rev 2012;31:134–178
127. Quehenberger O, Armando AM, Brown AH, et al. Lipidomics reveals a remarkable diversity of lipids in human plasma. J Lipid Res 2010;51:3299–3305
128. Coulier L, van Kampen JJ, de Groot R, et al. Simultaneous determination of endogenous deoxynucleotides and phosphorylated nucleoside reverse transcriptase inhibitors in peripheral blood mononuclear cells using ion-pair liquid chromatography coupled to mass spectrometry. Proteomics Clin Appl 2008;2:1557–1562
129. Mamer O, Gravel SP, Choinière L, Chénard V, St-Pierre J, Avizonis D. The complete targeted profile of the organic acid intermediates of the citric acid cycle using a single stable isotope dilution analysis, sodium borodeuteride reduction and selected ion monitoring GC/MS. Metabolomics 2013;9:1019–1030
130. Evans C, Bogan KL, Song P, Burant CF, Kennedy RT, Brenner C. NAD+ metabolite levels as a function of vitamins and calorie restriction: evidence for different mechanisms of longevity. BMC Chem Biol 2010;10:2
131. Vivekanandan-Giri A, Byun J, Pennathur S. Quantitative analysis of amino acid oxidation markers by tandem mass spectrometry. Methods Enzymol 2011;491:73–89
132. Cui T, Schopfer FJ, Zhang J, et al. Nitrated fatty acids: endogenous anti-inflammatory signaling mediators. J Biol Chem 2006;281:35686–35698
133. Levison BS, Zhang R, Wang Z, Fu X, DiDonato JA, Hazen SL. Quantification of fatty acid oxidation products using online high-performance liquid chromatography tandem mass spectrometry. Free Radic Biol Med 2013;59:2–13
134. Sullards MC, Allegood JC, Kelly S, et al. Structure-specific, quantitative methods for analysis of sphingolipids by liquid chromatography-tandem mass spectrometry: "inside-out" sphingolipidomics. Methods Enzymol 2007;432:83–115
135. Wooding KM, Auchus RJ. Mass spectrometry theory and application to adrenal diseases. Mol Cell Endocrinol 2013;371:201–207
136. McDonald JG, Smith DD, Stiles AR, Russell DW. A comprehensive method for extraction and quantitative analysis of sterols and secosteroids from human plasma. J Lipid Res 2012;53:1399–1409
137. Kuhara T, Ohse M, Inoue Y, Cooper AJ. A GC/MS-based metabolomic approach for diagnosing citrin deficiency. Anal Bioanal Chem 2011;400:1881–1894
138. Martens-Lobenhoffer J, Bode-Böger SM. Mass spectrometric quantification of L-arginine and its pathway related substances in biofluids: the road to maturity. J Chromatogr B Analyt Technol Biomed Life Sci 2014;964:89–102