Metabolomics: A Primer
Xiaojing Liu1 and Jason W. Locasale1,*

Metabolomics generates a profile of small molecules that are derived from cellular metabolism and can directly reflect the outcome of complex networks of biochemical reactions, thus providing insights into multiple aspects of cellular physiology. Technological advances have enabled rapid and increasingly expansive data acquisition with samples as small as single cells; however, substantial challenges in the field remain. In this primer we provide an overview of metabolomics, especially mass spectrometry (MS)-based metabolomics, which uses liquid chromatography (LC) for separation, and discuss its utilities and limitations. We identify and discuss several areas at the frontier of metabolomics. Our goal is to give the reader a sense of what might be accomplished when conducting a metabolomics experiment, now and in the near future.

Metabolomics: Overview
Metabolomics investigates the activity and status of cellular and organismal metabolism on a global or network scale to delineate the end points of physiology and pathophysiology [1–5]. It involves the measurement of small-molecule compounds, including endogenous and exogenous molecules, that are the products and substrates of chemical reactions within biological systems. A metabolomics experiment directly reflects the activity of the metabolic network that leads to the production of these metabolites and yields essential information about the underlying biological status of the system in question. Thus, metabolomics is not defined by any particular experiment but reflects the study of metabolism in a comprehensive way. It can involve an ‘untargeted’ screen where thousands of unknown features (see Glossary) are profiled and the relative differences in two conditions or across a population (semiquantitation) are measured (Figure 1). Such a screening experiment could be useful in identifying a new metabolite that may be present in a genetic condition or a newly engineered metabolic pathway [6–12]. However, a semi-targeted metabolomics experiment is often more useful; here, a large number of molecules are unambiguously identified and quantified [13,14]. This process allows the user to generate data that are otherwise obtained from hundreds of separate biochemical assays, to characterize the properties of a network or pathway [12,15–20]. These experiments are considered semi-targeted because, while the list of metabolites is defined, the hypothesis may not be. Targeted experiments often provide deeper insights by testing a specific hypothesis because the absolute concentrations of molecules are measured (absolute quantitation) [16,21–23] or the rates or fluxes of the conversion of one molecule to another can be obtained [22,24,25]. Thus, a targeted metabolomics analysis requires substantial preexisting knowledge and its success depends on the strength of the hypothesis being tested.

Both NMR and MS are effective tools that analyze the molecular composition of a sample (Table 1) [26]. NMR detects molecular features by measuring an intrinsic magnetic property of atomic nuclei (i.e., the ‘spin’) that encodes information about the chemical environment and thus its molecular structure. MS is more commonly used for semi-targeted or untargeted metabolomics because it is more sensitive, has a higher throughput, and can measure more molecules in a complex biological sample. However, one advantage of NMR is that it is...
Efforts have also been made to combine these different instruments (LC–metabolomics discussing its process, the types of data that are obtained, and its applications. Metabolomics is becoming smaller, which allows single-cell metabolite profiling or spatial resolution within cells or tissues. However, data collection is merely the first step of metabolomics, and metabolomics is ultimately an integration of instrumentation, chemistry, statistics, and computer science with a biological problem

From Experiment to MS Data: Sample Preparation and Instrumentation

There is no limitation on the sample types that are suitable for a metabolomics study. However, the sample type and metabolites of interest determine the appropriate sample preparation procedures [32–34]. Furthermore, data interpretation differs markedly depending on the biological system from which the metabolites originated. For example, to compare health and disease states or to study drug actions, measuring the level of a metabolite in the serum may be a reasonable proxy for a physiological function [35] since the serum metabolome is the net effect of diet, environment, and the whole-body response to a disease or a drug. However, measuring the same metabolite level in a tissue would require a different interpretation, because it more reflects the cell-autonomous effect. In sample preparation procedures, a general principle is to preserve the original state of the biological system as much as possible by minimizing the amount of enzyme activity and chemical reactivity that occurs during metabolite extraction – the process of isolating or purifying metabolites from the original biological matrix (e.g., cells, serum, tissue).

For small-molecule metabolites, protein precipitation or liquid–liquid extraction are the most commonly used methods. Polar organic solvents such as methanol, acetonitrile, or isopropanol are used to extract mostly polar metabolites, whereas relatively non-polar solvents such as hexane, chloroform, or methyl tertiary butyl ether, or combinations of polar and non-polar solvents, are used to extract lipids [36,37]. Occasionally, acid is added to the extraction solvent to preserve the stability of certain compounds, such as acyl-coenzyme A (CoA) compounds [38]. However, although acidic solvents may stabilize one class of metabolites they may also simultaneously cause degradation of other types of metabolites and possibly cause an overall reduction in the sensitivity of the experiment due to ion suppression [39,40]. Despite the diverse chemical properties of metabolites, protocols have been developed that allow broad

**Glossary**

**Absolute quantitation**: a measurement of the concentration of a metabolite.

**Chromatography**: techniques used to separate metabolites based on chemical properties.

**Features**: signal patterns recorded in metabolomics data (defined by the pair of retention time and m/z in LC- or GC-MS-based metabolomics data).

**In-source fragmentation**: fragmentation of molecules that occurs during ionization of the sample.

**Ion suppression**: the reduction of sensitivity due to the difference in the ability to ionize a molecule as a result of the surrounding composition of that molecule.

**Mass resolution**: the ability to distinguish a molecule due to its m/z from another molecule with a different m/z.

**Pooled quality control samples**: a mixture of samples to be analyzed with similar matrices to provide a representative ‘mean’ sample that is used to check the performance variation among individuals or over time.

**Retention time (RT)**: the time taken for a metabolite to be eluted from a column.

**Semi-quantitation**: the measurement of the relative level of a molecule in two conditions or across a population.

**Semi-targeted metabolomics**: a system for the quantitation of hundreds of known metabolites and simultaneous detection of thousands of unknown features that are tentatively identified or not identified.
For instance, these general protocols may not offer the best sensitivity for every metabolite and therefore unstable or low-abundance metabolites may require special care.

MS involves first ionizing the molecules (i.e., adding a positive or negative charge to them) and then moving these molecules through electric fields where they are eventually analyzed. At each time point, the data are recorded as mass spectra comprising the mass-to-charge ratio ($m/z$) for each intact ion and corresponding intensity. Each ion has a retention time (RT) and mass spectrum, the values of which are dependent on the instrumental setup.

Two commonly used approaches for MS analysis are multiple reaction monitoring (MRM) and high-resolution MS (HRMS). MRM experiments are usually conducted on a triple-quadrupole mass spectrometer [21,27,52]. The first quadrupole filters ions (parent ion) within a defined molecular weight (usually with a resolution of 1 atomic mass unit). The second quadrupole fragments the molecules that have been selected and the third quadrupole selects characteristic fragments. Therefore, before any data acquisition, the parent and fragment ions must be defined and the optimized energy for the fragmentation and RT for each metabolite is needed.

Another increasingly common method for MS is HRMS. These approaches rely on the high mass resolution of the mass analyzer. One commonly used mass analyzer is the Orbitrap™, which records the oscillation of ions, the frequency of which provides information on the molecular mass [44]. Another is a time-of-flight (TOF) instrument, which records the time it takes
for an ion to traverse an electric field [45,46]. These mass analyzers greatly simplify compound identification compared with lower-resolution methods. When high-resolution mass analyzers are coupled to collision cells, which can fragment ions before they are sent to the mass analyzer, the fragment pattern of ions provides additional structural information. Then, at the data-processing stage, metabolite assignments are made.

**From MS Data to Metabolite Profiling**

There are numerous open-source and commercial software packages available for raw MS data analysis (Table 2). The output data includes peaks with specific RT and m/z; peak area is usually the preferred parameter to represent the relative abundance of each metabolite in different samples. These software packages typically involve chromatographic alignment, peak selection, and compound identification by searching against metabolomics databases. This untargeted experiment is relatively thorough and unbiased but usually produces thousands of features and, unfortunately, these features do not directly reflect metabolite identity (Figure 1). Characterization of metabolites starts with the matching of selected features with those of known metabolites. These databases are publically available at several online servers [47–51]. For metabolomics data generated by HRMS, m/z is often used as the only criterion for feature identification and, therefore, many features often return multiple metabolite identities, which is due to the presence of isomers or in-source fragmentation [52]. A further complication is that chromatographic RTs are highly dependent on the LC or GC setup, are difficult to reproduce from external databases, and vary over time even within a given laboratory. Much effort has been made to advance untargeted analysis and feature identification, including new MS/MS workflows and network integration (Table 2) [53,54]. A more detailed discussion of unknown metabolite identification is contained in a recent article [4]. Overall, untargeted analysis is challenging and the results are often hard to interpret.

Therefore, more targeted or semi-targeted data analysis is also performed simultaneously (Figure 1). An internal reference library including both m/z and RT, or even MS/MS spectra, is constructed in house from authentic chemical standards and additionally with other approaches such as the co-injection of an isotopically labeled extract from for example *E. coli* that contains metabolites with known assignments. Subsequent samples are then analyzed with respect to the reference library. Alternatively, sets of pure compounds or labeled *E. coli* can be spiked into the extract of interest. For a typical semi-targeted analysis, dozens to hundreds of metabolites can be assigned with high confidence from various biological samples and 3000–10000 other features are present in the spectra and remain unidentified [55,56]. Nevertheless, this semi-targeted analysis provides a time- and cost-effective yet informative metabolic profile and allows researchers to either test multiple hypotheses at once or investigate systems biology-level questions.

Converting ion intensity to metabolite concentration is complex, depending on variables such as the percentage of the compound recovered from the original material, the column binding capacity, the ionization efficiency, and the transmission efficiency through the mass spectrometer. However, MS-based metabolomics data are often semiquantitative, which means that although the signal itself (metabolite peak area) does not reflect the absolute concentration, differences in peak area do scale linearly with metabolite concentration. A differential analysis provides the relevant biological information.

Normalization may be required in certain cases. For example, inconsistent sample preparation or extraction from different sources can result in varying ion suppression and cause nonlinear shifts of MS intensities of metabolites in different samples. In these situations isotopically labeled standards (internal standards) can be added to each sample before the extraction for
normalization [41]. However, applying internal standards is challenging due to the wide diversity of chemical properties and wide concentration ranges of metabolites in biological samples. When conducting analyses in the absence of these standards, we therefore recommend making comparisons using material from similar origins (compare serum A with serum B or tumor A with tumor B) and as similar amounts of material as possible. Alternatively, pooled quality control samples can be used to reduce variation due to batch effects [57]. When applying these principles, most MRM and HRMS methods have yielded a linear range of quantitation for three to four orders of magnitude.

Analysis of a Metabolomics Experiment
Since metabolomics experiments typically contain information that could otherwise be obtained from hundreds of separate biochemical assays, some preexisting biological knowledge helps with the interpretation of a metabolomics experiment. Under this framework one can simply use the data to ask biologically relevant questions and make conclusions following the standard scientific method. Such questions could be as follows. Does the energy status change under this condition? What about the redox status? Are the nucleotide levels maintained when this gene is overexpressed? Treating a metabolomics experiment with this mindset often allows one to reach conclusions about a biological mechanism in a highly expedited fashion as opposed to pursuing these hypotheses one by one with separate assays.

Nevertheless, it is not humanly possible to process the entirety of the data from intuition alone. Computational tools are needed for further analysis. Software for feature extraction (Table 2) often include additional data analysis functions, such as principal components analysis and hierarchical clustering, and numerous statistical tests and data visualization plots to identify the largest-changing features and specific signatures in the data. Pathway-enrichment analyses,
which are commonly used in gene expression analysis [58], can also be used with metabolomics data to identify affected metabolic pathways. However, metabolite annotation and pathway demarcations are not as well developed as in the genomics field; thus, these analyses often produce results that are hard to interpret. Nevertheless, pathway-enrichment analyses still give useful insights into groups of metabolites of interest. For example, in a recent study pathway-enrichment analysis of metabolomics data showed that the methionine cycle was altered before other parts of the network were affected [20]. Further, network mappings can contribute to systems-level analysis on integration with other ‘omics’ data sets [59,60]. One commonly used network mapping tool is Cytoscape and the metabolism plugin for Cytoscape, called Metscape [61,62]. There are other software that employ the Cytoscape platform, such as MetaMapR and GAM, which may generate metabolic networks based on enzymatic transformations, metabolite structural similarity, mass spectral similarity, or empirical associations [59,63]. The recently developed software PIUMet facilitates unknown metabolite identification by network integration of untargeted metabolomics [54]. These tools become more powerful on integration with other omics data and allow the user to find regions of the metabolic network that correspond to a phenotype or are altered in a condition.

Isotope Tracing, Flux Analysis, and Computational Modeling
The overall metabolite profile is informative in many cases as discussed above, but for metabolites involved in multiple catabolic and anabolic pathways, metabolite levels reflect some complicated conglomeration of each individual pathway contribution. An appropriately designed isotopic tracing study is then used to identify the activity within each pathway of that metabolite and is the phenotypic readout of metabolism. The isotopic labeling pattern of downstream metabolites is used to represent the metabolic flux from different sources. Typical protocols involve incubating a stable isotope-labeled nutrient (such as glucose, an amino acid, a lipid, or other molecules) at the same concentration as the original experiment (different concentrations will induce changes to metabolism from the original condition of interest) and waiting for the metabolic flux to reach steady state [64,65]. Note that ‘steady state’ means that the isotopic labeling pattern of metabolites of interest is no longer time dependent.

Much of the time, comparing the labeling patterns of a certain metabolite under different experimental conditions or performing a simple calculation of the labeled nutrient’s contribution to downstream metabolites is sufficient to interpret a metabolic flux qualitatively. However, model-based flux calculations often provide more accurate and sometimes more comprehensive views of metabolism that cannot be obtained from intuition alone [64,65–68]. These techniques are collectively termed ‘metabolic flux analysis’ and involve taking a series of isotope measurements of different metabolites, overlaying them onto a metabolic network, and fitting a mathematical model of the fluxes onto the network that best fits the data. This approach is powerful but at present the implementation of the technology has been limited and most computations involving flux estimation and statistical analysis of the results are typically conducted on an ad hoc basis. Future directions will involve successfully implementing these approaches to serve a broader audience. More detailed discussions about flux calculations have been covered by previous reviews and we refer the reader to these [64,65,77,79].

Recent Biological Insights Obtained From Metabolomics
In recent years metabolomics has been applied in multiple fields to make new discoveries and confirm hypotheses (Figure 2). A seminal advance in cancer biology used untargeted LC-HRMS to discover that cancer cells with mutant $IDH1$ produce 2-hydroxyglutarate, a metabolite that was found in the spectra of the mutants but not the wild-type cells after untargeted metabolomics using LC-HRMS. This metabolite was later shown to provide a link between metabolism and epigenetics [69,70].
Metabolomics has also been widely used in drug discovery and drug action [3, 71]. An extensive drug database, DrugBank (http://www.drugbank.ca/), now exists [72]. Over 8000 drug entries are recorded in this database and related information, such as known drug metabolism and known or proposed drug targets, is also provided. Meanwhile, metabolomics is helping researchers gain new knowledge of drug action. For example, through a metabolomics study Ser et al. showed that 5-fluorouracil (5-FU), a commonly used chemotherapy drug, caused overproduction of the nucleotide deoxyuridine, which was also measured in 5-FU-treated mouse serum. The overflow of this nucleotide could potentially be used as a biomarker for positive response to 5-FU treatment [73].

Semi-targeted metabolomic analyses have led to a deeper understanding of the etiology of aging, cardiometabolic disease, and cancer, where metabolomics data not only test existing hypotheses, which can come from transcriptional or proteomics data, but also provides insights when there is a lack of hypotheses or when the original hypothesis is rejected. For example, a study on Drosophila employed HRMS and demonstrated a critical role for specific clock genes in modulating the effects of nutrient manipulation on fat metabolism and aging [74]. In a separate study, MRM-based analyses have identified signatures of the onset of pancreatic cancer and diabetes [75]. Elevated plasma levels of branched-chain amino acids (BCAAs) were found to be associated with increased risk of diabetes and other cardiometabolic diseases [76, 77]. These studies spawned further mechanistic analyses, which have pointed to the role of amino acid oxidation in the mitochondria as a key player in the etiology of these diseases. The same signature was also associated with increased risk of future pancreatic cancer diagnosis [77]. These findings led to the mechanistic insight that the cause of this elevation was an increase in tissue breakdown, at least in the case of tumor development. Semi-targeted metabolomic analyses have also provided insights into human population genetics and the origins of metabolic traits. Genome-wide association studies (GWASs) for variation in metabolite levels were conducted on human blood from 2820 individuals [78]. Surprisingly, numerous genetic loci associated with blood metabolite concentrations were discovered and the locations of these loci, for the most part, were in proximity to a gene encoding a metabolic enzyme involved in the production or consumption of the given metabolite. This analysis provided evidence that some of the metabolome variation observed across individuals may be directly encoded in the genome. Such studies may be valuable for the understanding of personalized nutrition, which is still in its infant stages [79].

Flux analyses have also led to important advances in cancer biology. One example is the discovery of the diversion of glycolytic flux into de novo serine metabolism as an important process in some cancers. Further efforts to trace the fate of serine using stable isotopes,
HRMS, and mathematical modeling revealed how the network downstream of serine is coordinated, leading to the identification of a potential cancer therapeutic [80]. These studies have also identified systems-level properties that show how gene expression levels can sometimes be used to predict metabolic flux, in which case biomarkers for metabolic flux could then be obtained from human specimens, from which mRNA is often more readily available [81]. Other flux analyses have identified previously unappreciated pathways that are important in cancer, leading to new drug targets [82–84]. As another example, acetate has been identified as an alternative fuel source for cancer metabolism [85–87]. Mashimo et al. first performed in vivo tracing in patients and showed increased oxidation of acetate in the brains of patients with glioblastoma and brain metastases using $^{13}$C-NMR; furthermore, Comerford and Schug et al. demonstrated that the utilization of acetate in tumors requires the enzyme acetyl-CoA synthetase 2, which is now an important drug target in cancer. Thus, in these cases metabolomics has led to novel findings that have advanced key areas of biomedical research.

Concluding Remarks: Challenges and Future Directions in Metabolomics

In summary, much progress in metabolomics has been made and obtaining a metabolite profile or measuring metabolic flux is now standard practice. Efforts are also being made to advance the field by covering more metabolites with less material or effort, achieving spatial resolution, and the integration of multiomics (Figure 3). Nevertheless, challenges remain in multiple areas.

---

**Figure 3. Trends in Metabolomics.** Trends include broader metabolite coverage from smaller sample sizes, achieving spatial resolution, and the integration of multiomics data.
The greatest challenge for metabolomics is how to best obtain biological insight with the appropriate experimental design. Success often achieves a balance of biological intuition that is supplemented with computation. Choice of experimental model is an essential component in this process. For example, ex vivo systems that can successfully model in vivo metabolism are important [88–90]. Conclusions drawn from cell culture have substantially advanced biomedical knowledge but must be assessed with respect to the assumptions in the model, such as the nutrient-rich environments that typical culture conditions constitute. Another major challenge is subcellular compartmentalization, as metabolomics data reflect the sum of metabolites in various cellular organelles. Although it would be preferable to assay metabolites from specific organelles, it is very challenging, if currently impossible, to separate organelles while preserving the metabolic state of these structures. Recent studies have used clever isotope tracing strategies to probe cofactors that occur in a specific compartment such as the cytosol [91]. Metabolic flux analysis can also be used to estimate fluxes involving metabolites that are shared across compartments [80,95].

Metabolic heterogeneity is another issue in understanding metabolism and efforts to conduct metabolomics at the single-cell level are under way [92–95]. This particularly exciting frontier in metabolomics research involves the analysis of single cells to obtain information that is masked in bulk studies. Various approaches have been developed and successfully applied to plant, neuronal, yeast, bacterial, and animal cells in the past decade [105,96]. Single cells could be isolated with a needle or through microfluidic devices, followed by metabolite extraction and further analysis with HRMS or MS/MS [97–99]. Intact cells or tissue samples can also directly interact with an ionization source to generate ions, which are then analyzed by MS [100–102]. These technologies in principle can achieve submicrometer spatial resolution and therefore metabolomics imaging and analysis of subcellular organelles can be achieved. However, with this extremely fine spatial resolution the data acquisition time and data size are also dramatically increased, which makes it difficult to perform single-cell metabolomics in a high-throughput fashion. In addition, efforts are also required to improve the robustness, number of metabolites covered, and accuracy. Therefore, standard metabolomics techniques that average cells remain the mainstay.

Better approaches to standardize metabolomics data are also essential to advance the field. Thus, efforts are under way to create better normalization procedures and better protocols to rapidly obtain absolute metabolite concentration values [103,104]. For flux analysis the mathematical frameworks and algorithms are well beyond what is typically implemented and accessible by the larger community. Efforts to further disseminate these capabilities will prove valuable. Nevertheless, current metabolomics technologies coupled with interesting questions already allow rapid inroads to be made.

Acknowledgments
This work was funded by R00CA168997 and R01CA193256 to J.W.L. from the National Cancer Institute at the National Institutes of Health. The authors thank the anonymous reviewers for helpful suggestions on improving the manuscript.

Outstanding Questions
How does one best extract meaningful biological information from massive amounts of metabolomics data?
How can in vitro models more accurately reflect in vivo metabolism?
How can intracellular compartmentalization be resolved in metabolomics experiments?
What are the best ways forward in conducting single-cell metabolomics experiments?
What are the best ways to enable metabolic flux analysis to be more broadly used in the biomedical community?

References