Metabolomics Tier 2 Data Analysis:

Pathway Mapping and Analysis: Advances in high throughput technologies have enabled biomedical scientists to interrogate the entire transcriptome, as well as significant portions of the proteome and metabolome, to obtain insights into biological mechanisms associated with cellular function. Until recently, the analysis focused on identifying a list of differentially expressed molecules that may play a significant role in a biological process or determine phenotype. However, most often such a list fails to provide mechanistic insights into the underlying biology of the condition under study. Hence, over the years, attention has shifted from lists of molecules to sets of functionally related coordinated alterations, that constitute pathways or bioprocesses, that in concert orchestrate the underlying biology at cellular or organismal level. In other words, this so called pathway-centric approach results in reduction of data dimensionality while preserving the interaction between the components within an experiment [Glazko G, Emmert-Streib F Unite and conquer: univariate and multivariate approaches for finding differentially expressed gene sets. There are different approaches to define pathways using metabolomics data, some of which involve mapping metabolites to existing pathway maps and rely on enrichment methods, while others are much more sophisticated in that they explore enrichments across a larger compendia of molecular processes assembled within databases without the prerequisite for pre-defined pathway maps. Irrespective of the method employed, the initial step to define pathways requires mapping the metabolic data to unique identifiers. The latter could either be compound IDs or in more sophisticated methods, these could be gene IDs associated with the reactomes involving the metabolites.

Mapping of metabolites to unique identifiers: The metabolites measured in the samples will be related to their compound or CAS IDs (KEGG database), enzyme/gene ID using Python scripts and used for all analyses described below. Similarly gene and protein expression data will be linked to their respective Unigene and IPI IDs.

Pathway Enrichment Analysis Strategies: Once mapped to the identifiers, we will carry out pathway mapping using methods that rely on a priori pathway maps or those that work independent of such maps. Overall, experiments aimed at verification of a pre-
defined hypothesis or that require visualization of data in the context of pre-defined functional interactions, benefit from the former approach. On the other hand, data from experiments that are seeking to define hypothesis can be tested using either a similar approach or would benefit from the pre-defined pathway independent strategy. In both the above cases however, one needs to examine a minimum subset of metabolites (at least > 20) to get meaningful results.

**Strategies-based on pre-defined pathway maps** Enrichment analysis methods under this category can be broadly classified into three groups: (i) over-representation analysis, (ii) set enrichment analysis and (iii) network based enrichment methods. These methods are found as modules in commercially available pathway mapping tools like Gene Spring, Ingenuity Pathway Analysis, and GeneGO, and freely-available software such as Gene Set Enrichment Analysis (GSEA), GSA, NetGSA, etc. Although these methods were originally adopted for analysis of gene expression dataset, their functionality is by far independent of the type of profiling data used, extending their application to proteomics and more recently metabolomics datasets. In spite of their application in evaluating metabolomics datasets, there exist a number of issues that need to be considered, as discussed in the following section.

Over-representation analysis techniques (ORA), the standard approach for commercial products, use variations of the following basic strategy: an input list containing altered levels of metabolites between two conditions being compared is selected at a certain threshold (e.g. FDR at 5%). Following this, for each pathway, the number of metabolites from the input list belonging to the pathway are counted and the process repeated for an appropriate background or null set comprising of the entire universe of measured metabolites in the experiment. Each pathway that has metabolites from the differential list mapped to it is tested for over- or under-representation using tests based on hypergeometric, chi-square, or binomial distribution. The major advantage of such an approach lies in its simplicity, while the drawback lies in its strong dependence on the pre-defined threshold used to define the differential metabolic signature (input list). Furthermore, this approach is not sensitive to correlations or interactions among the metabolites within the signature, or pathways represented by them, which in turn could accentuate the significance associated with the enrichments.
To overcome these shortcomings, functional class scoring (FCS) methods can be used with the most prominent example being GSEA. These methods rely on the use of the entire data without any threshold-based filtering. The general strategy of FCS methods relies on computing a statistic for each of the measure analyte in the specimen. Popular statistics that are employed include t-test, fold-change, Z-scores or correlation coefficient, with the condition/perturbation under investigation. Measured analytes along with their associated statistics are assembled in a pathway context using pre-defined maps (example from KEGG) and the aggregate statistic is computed per pathway. Methods like sum, mean, median, max-mean, rank sum, Kolmogorov-Smirnov, etc, are popularly used to compute aggregate statistics in standalone software in GSEA, GSA, sigPath and SAM-GS. The final step in FCS is to assess the statistical significance of the pathway level statistic, which is done computationally by permuting the labels of the samples between the two conditions being compared (example treatment vs. control). In general, FCS methods are more powerful than ORA, but are computationally intensive, represent a univariate measure that is independent of underlying interactions between the analytes, and the statistical power obtained is sensitive to the size of the pathway examined.

To address these shortcomings, pathway (network) topology methods have been recently developed: ScorePage that takes into consideration similarities between members of the pathway and uses those as weights when calculating a pathway level statistic, Impact Factor Analysis that models each pathway as a linear system and hence takes into consideration the interactions between its members, and NetGSA developed by Co-I Michailidis, that models all the pathways together as a graph and hence takes into account interactions both between pathways and within pathways. NetGSA can also account for changing pathway topologies observed between the experimental conditions. In general, network based methods exhibit superior statistical power, at the expense of increased computational time.

Overall the four major challenges in examine metabolomic data using a pathway-centric approaches are 1) unlike gene expression data are sparse 2) few metabolites easily map to pathways and 3) some pathways contain relatively few nodes and 4) metabolites behave either as a substrate or a product within a reactome and hence exhibit dual
directionality. The first three drawbacks make ORA and FCS potentially ineffective. Given this, in cases where pathway maps are known, our group will rely mainly on NetGSA-type analysis that provide enough statistical power to the results irrespective of pathway size, number of metabolites detected or their membership profile in pathways.

**Strategies that do not require pre-defined information on pathways:** Here we propose to use Oncomine Concept Maps (OCM, private.molecularconcepts.org) to examine the metabolic data. In OCM, genes derived (biosynthetic or catalytic genes associated with the metabolite in various reactomes) from the metabolic data are used to seed the analysis. The null set contains the universe of genes obtained from the entire measured metabolome. OCM carries out all versus all enrichment analysis using ORA against sets of molecular components, i.e., molecular concepts associated with different publically available databases, all of which are plugged into the backend of the OCM server. There are about 12 different databases that include 1) gene and protein annotations from external databases and 2) computationally-derived regulatory networks. Annotations include chromosomal locations, protein domains and families, molecular functions, cellular localizations, biological processes, signaling and metabolic pathways, protein-protein interaction networks, protein complexes, and gene expression signatures. The regulatory networks are derived by scanning human promoters for known transcription factor motifs and by comparative analyses.