Metabolomics: Open Access
ISSN: 2153-0769

Madhavan et al., Metabolomics 2014, 4:1
DOI: 10.4172/2153-0769.1000130

Research Article Open Access

Integrative Analysis Workflow for Untargeted Metabolomics in Translational Research

Subha Madhavan[1,2]*, Robinder Gauba[1], Robert Clarke[2] and Yuriy Gusev[1]

[1]*Innovation Center for Biomedical Informatics, Georgetown University Medical Center, Washington, DC, USA*
[2]*Lombardi Comprehensive Cancer Center, Georgetown University Medical Center, Washington, DC, USA*

Abstract

Background: Metabolomics is an emerging 'omics' science that has demonstrated its fast gaining importance as a powerful profiling tool for determining an individual's response to a foreign stimulus such as a drug, toxin, or environmental change; or as an indicator of disease progression. Such small molecule profiles can used as biological markers of disease, and provide an indicator of drug efficacy or toxicity.Several studies have demonstrated that the results of any single 'omics' analysis may not be sufficient to decode extremely complex biological mechanisms.

Methods: We have developed a translational research workflow to enable researchers to perform cutting-edge integrative analysis of metabolomics data with transcriptomics (gene expression) data using knowledge-driven networks. This network based view of interconnected functional partners can provide valuable new insight about the underlying biochemical processes and pathways associated with the phenotype of interest. To enhance and simplify metabolomics annotation we built a fully cross-referenced database called MetPlus DB, which integrates data from the three most comprehensive metabolite databases tailored largely towards mammalian metabolomics: HMDB, HUAMNCYC, and LIPID MAPS.Cross-referencing information is provided for linking to several other mainstream cheminformatics/bioinformatics repositories including KEGG, METLIN, ChEBI, FooDB, Pubchem, and Chemspider to provide unambiguous knowledge on clinically and physiologically relevant metabolites. We have made the integrated database available freely to the research community *(https://github.com/ICBI/MetPlus-DB*).

Results: To demonstrate the usefulness and strength of our methodology, we have tested it on a multi-omics profiling dataset from NCI-60 breast cancer cell lines to explore the biological dynamics of breast cancer.

Conclusions: Our results demonstrate a streamlined approach for the comprehensive annotation of metabolites using MetPlus DB, and the subsequent integration of metabolomics and transcriptomics data to explore potentially relevant biological interactions and candidate biomarkers associated with disease phenotype.

Keywords: Untargeted metabolomics; Integrative analysis; Systems biology; Knowledge-driven network; NCI-60

Introduction

With the recent advances in technology related to ultra-high performance mass spectrometry, especially those concerned with liquid chromatography/mass spectrometry (LC/MS), it is now possible to perform a comprehensive assessment of thousands of physiochemically distinct small molecules such as lipids, central carbon metabolites, sugars and amino acids, both exogenous or endogenous in nature, to gain insights into cellular and physiological responses. The field of metabolomics is showing tremendous promise for providing fast, accurate, and nonbiased profiling methods to elucidate underlying disease mechanisms and diagnosis, develop new effective strategies for the treatment, and for other potential applications in diverse clinical areas such as personalized medicine, nutritional metabolomics, population profiling, and molecular epidemiology.

Metabolomics analyses are commonly practiced in two fundamental ways. A targeted metabolomics study with the help of internal standards and specific mass spectrometer (MS) conditions aims at measuring previously identified metabolites for achieving better quantitation and validation. Untargeted metabolomics research largely focuses on discovery to detect and quantify any small molecule (<1500 Da) that can be ionized by mass spectrometry to understand biochemical function and change within a biological system.

The ability to easily quantify changes in the biological system makes metabolomics an attractive translational research tool that can help identify biomarkers of disease through non-invasive measurements of biofluids or tissue [1]. In combination with other molecular measurements it has shown great potential for accurate diagnosis, prognosis, and evaluation of therapeutic response and drug efficacy [2].

The cancer metabolome may be the most studied to date, particularly in the area of breast cancer diagnostics [1]. Several breast cancer studies using multivariate analysis of metabolite profiles from tissue have shown very high sensitivity and specificity in discriminating a malignant phenotype from benign tissue, and even cancer grade and hormone status [3-5]. Numerous studies have also indicated that changes in lipid metabolism can be an indicator of cancer (often characterized by an elevation of total choline-containing compounds (tCho) and phosphocholine). Choline levels have been extensively studied as a biomarker to stratify various cancers, including brain tumor types and grade [6,7], and distinguish between benign and malignant breast tumors [8,9], and prostate tumors [10,11].

***Corresponding author:** Subha Madhavan, Innovation Center for Biomedical Informatics, Georgetown University Medical Center, Washington, DC, USA, Tel: 2026873294; Fax: 2026875011; E-mail: sm696@georgetown.edu

Received January 17, 2014; **Accepted** February 25, 2014; **Published** March 02, 2014

Citation: Madhavan S, Gauba R, Clarke R, Gusev Y (2014) Integrative Analysis Workflow for Untargeted Metabolomics in Translational Research. Metabolomics 4: 130. doi:10.4172/2153-0769.1000130

The use of a multi-omics, systems biology approach to study cellular metabolism is key to understanding disease etiology and outcome. Integration of omics data types can help identify biological pathways affected and provides a more complete picture disease biology and prognosis, and can even provide new drug targets. One recent study integrated the metabolomics profile of tissue from breast cancer patients with protein expression data for glycerol-3-phosphate acyltransferase (GPAM), an enzyme involved in lipid biosynthesis. This data was overlaid with gene expression data to show that GPAM expression in breast cancer is associated with changes in the cellular metabolism, especially an increased synthesis of phospholipids [12]. Another study integrated transcriptomics and metabolomics data from NCI60 cell lines to show biological pathways associated with drug sensitivity, which demonstrates potential as a methodology to predict response to drug therapy [13].

Several approaches have been recently introduced to integrate metabolomics data with gene expression data [14,15]. Some of the current methods utilize a data-driven approach based on a statistical framework of correlation or multivariate analysis [16] producing computationally derived associations between the abundance of metabolites and level of gene expression. Such methods have proven to be effective in elucidating possible connections between changes observed in metabolomics data and gene expression data. The existence of correlated changes, however, does not necessarily represent real biological interactions; for instance, correlated metabolomics and gene expression data may indicate independent co-regulation by upstream regulators or environmental factors (e.g. hypoxia etc.). These computationally derived associations do not provide a clear path to biologically meaningful conclusions about molecular interactions and require further downstream analysis for biological interpretation. Recent studies have been also reported on applications of pathway enrichment analysis for gene expression and metabolomics data [17,18] utilizing biological pathway information for integration of metabolomics and transcriptomics for biomarker discovery.

Metabolite identification and systematic curation of public resources to obtain high confidence annotations is one of the biggest challenges faced by researchers working in the field of metabolomics. One of the current methods to address this challenge is to use consensus based reporting Metabosearch [19], and the Manchester Metabolomics Database (MMD) [20] by merging annotations from multiple resources to obtain annotations of higher confidence and also achieve widespread coverage since none of the existing bioinformatics/cheminformatics resources claims to cover the entire metabolome [21]. Another useful approach to identify metabolites is to carry out a restrictive search using organism-specific metabolomics databases since each species is characterized by a biochemically distinguishable and unique metabolome. Disregarding the source of origin of metabolites can lead to false and inconsequential identifications leading to wasted downstream validation efforts [22,23].

We describe a streamlined biological context-driven workflow for turning crude MS information into reliable and actionable knowledge applicable to profiling studies focusing on untargeted metabolomics.As part of this workflow we introduce a metabolite annotation database, MetPlus DB to address gaps in the current tools and provide the scientific community with a freely available resource for reliable annotation of metabolites. MetPlus DB isa fully cross-referenced database that integrates data from the three most comprehensive metabolite databases: HMDB [24], LIPID MAPS [25] and HUMANCYC [26]. Another aspect of this workflow is the ability to combine the information from complimentary 'omics' technologies, particularly metabolomics and transcriptomics, to build integrative, knowledge-driven networks. This method provides a novel systems biology based way to explore metabolite-to-gene interactions that can facilitate biological interpretation of the findings and generation of new hypotheses and be incorporated into an overall systems biology approach to metabolite profiling.

Materials and Methods

Integrative analysis workflow

The entire workflow for carrying out integrative analysis of metabolomics datasets can be broadly categorized into 4 steps – 1) data acquisition and processing, 2) metabolite identification, 3) data integration and 4) data interpretation. The workflow diagram with all the major steps is shown in Figure 1.

The first step in the workflow involves extraction of high-resolution LC/MS signals. Several frameworks are being developed for feature detection (and alignment) including MarkerLynx (Waters, commercial), MathDAMP [27], MetAlign [28], XCMS [29] and MZmine [30] (open source, freely-available). Using various univariate and multivariate statistical techniques the screening involves identification of hundreds to thousands of features that are differentially regulated between two or more sample groups or those strongly associated with the phenotype of interest. While the first step has now become straightforward and routine practice there is a pressing need for improvements in the latter steps of the workflow (i.e. metabolite identification and data interpretation especially in the case of multi-omics studies). Our method allows for omics data integration to provide a more comprehensive biological context for interpretation of metabolomics findings.

Database design and implementation

MetPlus DB is implemented as a SQLite database to deliver high confidence annotations with significantly reduced redundancy and eliminates the complexity of extracting metabolite annotations from individual databases. The data is primarily integrated from three comprehensive metabolite databases including HMDB (Version: 3.0), HUMANCYC (Version: 16.0) and LIPID MAPS (Updated: April 12, 2012) by merging records based on IUPAC International Chemical Identifier (InChIKey). The InChIKey is a fixed-length hashed version (27-character, including two hyphens) of the full InChI with five distinct components: 1) the first block (14-characters) encodes molecular skeleton (connectivity); 2) the second block (8-characters) contributes to proton positions (tautomers), isotopomers and stereochemistry; 3) a single flag character; 4) a single version character; and 5) a deprotonation indicator [http://www.iupac.org/home/publications/e-resources/inchi.html].

The InChIKey works very effectively in linking the chemical information together across multiple databases. It also provides a viable strategy for eliminating redundancy, inconsistency and inadequate representation of metabolite entry pulled from different sources. Since untargeted metabolomics experiments aims at identifying putative annotations for m/z signals through mass based search within a specified mass tolerance (ppm), we designed MetPlus DB exclusively to include records searchable through molecular weights preferably monoisotopic weights. However, mass-based queries can result in multiple numbers of putative identifications for any single

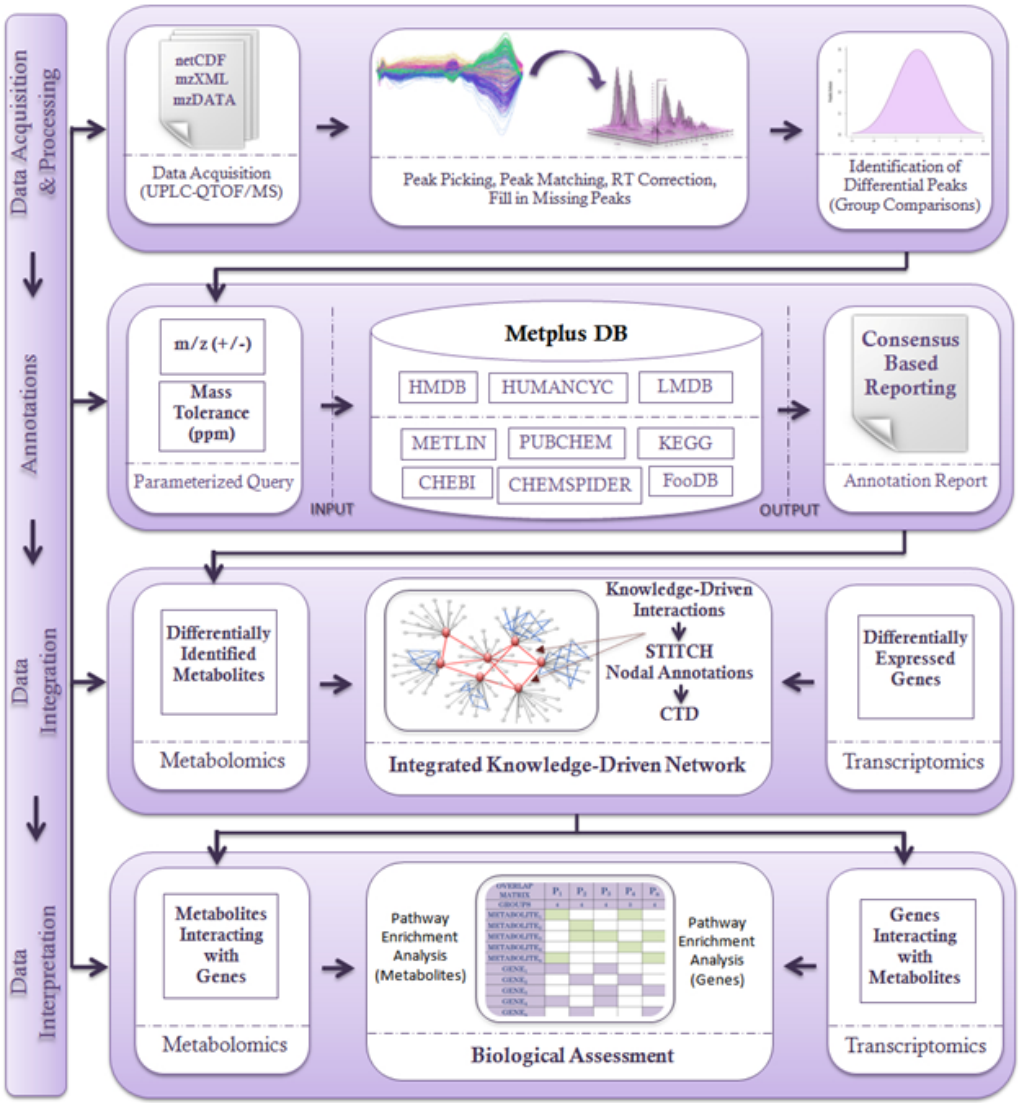

Figure 1: The bioinformatics . Outlining major steps for constructing the integrative knowledge-driven network and its biological interpretation using complimentary 'omics' datasets.

m/z signal. It is always advised to further investigate and validate all putative annotations by UPLC-ESI MS/MS protocol to determine the metabolite identity. For the HUMANCYC database where only InChI string is available, a conversion to InChIKey was carried out using InChI resolver service [http://www.chemspider.com/inchi-resolver] provided by Chemspider. In the case of LIPID MAPS, records are included for integration having either valid corresponding Pubchem identifiers and/or HMDB identifiers. Metabolites with more than one primary identifier having the same InChIKey are merged as a single record to minimize redundancy. Cross-referencing information for linking to other mainstream cheminformatics/bioinformatics databases (e.g. ChEBI, Pubchem and Chemspider) is based on matching information on InChIKey. Additionally, HMDB and LIPID MAPS are used to provide pre-existing cross-referencing information in the case of METLIN, Foodb, and KEGG.

Knowledge-driven network construction

In this workflow, we have created an integrated knowledge-driven network by combining two profiles (metabolomics and transcriptomics) using STITCH ('Search Tool for Interactions of Chemicals') (Version: 3.1) [31] a repository of data that captures the publicly available knowledge on protein–chemical associations. STITCH is an aggregated database of interactions connecting over 300,000 chemicals and 2.6 million proteins from 1,133 organisms. As an exploratory tool STITCH includes 254,000 high-confidence human protein–chemical edges/interactions (confidence score ≥ 0.7). A confidence score is assigned taking into account both level of significance and certainty of an interaction [31]. Following metabolite annotation using MetPlus DB, a combined list of identified metabolites and genes/proteins that are strongly associated with the phenotype of interest was uploaded to the STITCH database to produce an interaction network. Edges in the network depict knowledge-driven interactions (e.g. based on text mining or experimental evidence) between genes and/or proteins. Hypotheses generated from this approach enables data-driven knowledge discovery. Next, the Comparative Toxicogenomics Database (CTD) [32] is used as an additional source of information

to annotate both metabolites and proteins as molecular markers or therapeutic targets in the context of given phenotype or disease. A complete workflow for the data integration and interpretation steps of the pipeline resulting in a gene-metabolite knowledge-driven interaction network is shown in Figure 2.

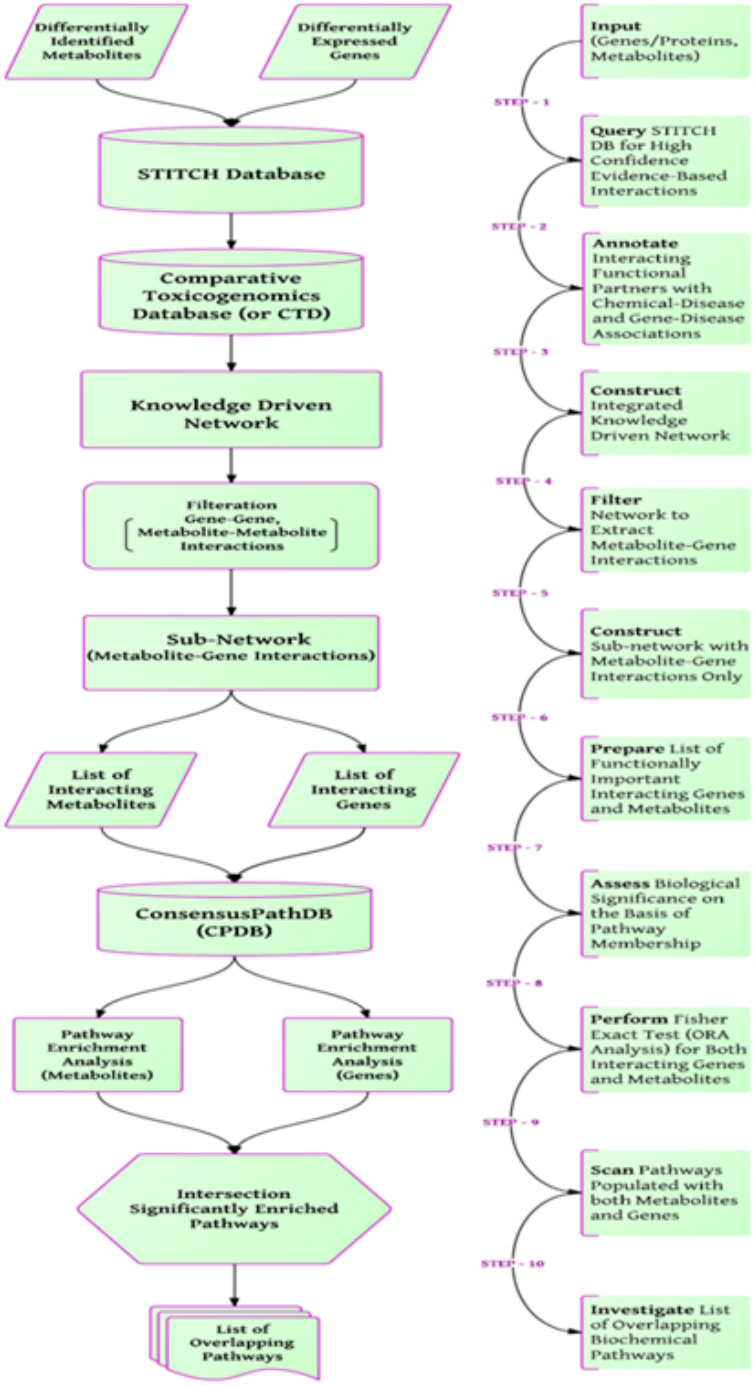

Figure 2: A detailed stepwise for assessing the biological signi of knowledge-driven integrative network of gene and metabolite interactions. Step-1 Upload a combined list of differentially identified metabolites and genes/proteins into STITCH; STEP-2 Retrieval of high-confidence human protein–chemical interactions (confidence score ≥ 0.9); STEP-3 Use the Comparative Toxicogenomics Database (CTD) as an additional source of information to annotate both metabolites and proteins that are molecular markers or therapeutic targets in the context of given phenotype or disease; STEP-4 Sub-set the interaction network by filtering out the gene-to-gene and metabolite-to-metabolite interactions; STEP-5 Re-construct the knowledge driven network based on direct interacting metabolites and genes/proteins; Step-6 Prepare the list of interacting genes/proteins and metabolites from sub-network constructed in the previous step; Step-7 Perform pathway enrichment analysis using CPDB (ConsensusPathDB) database; Step-8 Perform Fisher exact test (ORA Analysis) for both list of interacting metabolites and genes/proteins; Step-9 Scan for pathways populated with both metabolites and genes; Step-10 Investigate the list of overlapping biochemical pathways to assess the biological significance.

This network view of interacting functional partners can provide new insights about their association with the phenotype of interest and a more granular understanding of interdependence and interconnectivity between underlying biochemical processes and pathways at a systems level.

Case study

We tested our computational workflow on two breast cancer cell lines: BT-549 (basal breast cancer) and MCF-7 (luminal breast cancer) using a combination of two complimentary 'omics' datasets: 1) transcriptomics dataset –expression profiles were measured on the Affymetrix U133 Plus 2.0 microarray platform downloaded from the Gene Expression Omnibus (GEO; accession number GSE32474 [33]); and 2) metabolomics dataset –the NCI60 metabolomics data was downloaded from DTP web portal (http://dtp.nci.nih.gov/mtargets/download.html; August 2010 Data Release).

Case study data processing

The metabolomics data was generated by Metabolon Inc. [http://www.metabolon.com] and consists of measurements on 352 metabolites with 154 putatively identified. After removing metabolites having no variation across samples and excluding retired annotations 137 metabolites were selected. The raw gene expression data from Affymetrix U133 Plus 2.0 chip was pre-processed using RMA normalization [34] equating to 54,675 probes. The processed data was log transformed and non-specific filtering was performed using nsfilter function from the Bioconductor package genefilter [35] to remove probe sets that have no Entrez Gene identifier and have low intergroup variability, as these are likely to be uninformative. To ensure that each probe set maps to exactly one Entrez Gene ID, if multiple probes were found that mapped to the same Entrez Gene ID a probe with the largest IQR was selected for downstream analysis.

Case study statistical analysis

Statistical group comparison was performed using a t-test (Bioconductor). A total of 54 differential metabolites with adjusted p-value cut-off 0.05 were identified between basal (BT-549) vs. luminal (MCF-7) breast cancer cell lines (based on three biological replicates for each cell line).Similarly, 428 DEGs were identified with an adjusted p-value cut-off at 0.01 and absolute log-ratio of 2. The multiple testing corrections were computed using the Bioconductor package qvalue [36] (Supplemental File 1).

Case study network inference

The differentially identified metabolites (using KEGG identifiers) and differentially expressed genes (using HGNC Symbols) were used as input to query chemical-gene interactions using the STITCH database (illustrated in Supplemental File 2). Using a confidence score of ≥ 0.9, a set of highest-confidence interactions was outputted and the resultant interaction network was downloaded locally in text format (Supplemental File 3). Furthermore, the Comparative Toxicogenomics Database (CTD) was used as an additional source to provide manually curated information on interacting metabolites and genes that are annotated as molecular markers or therapeutic targets in the context of breast cancer. The network-based representation of metabolite – protein interactions derived from STITCH and CTD was created using cytoscape [37].

Case study pathway enrichment analysis

To assess the biological significance of an integrative network in the context of significantly affected biological pathways, a pathway enrichment analysis was performed on the list of 26 differentially identified metabolites annotated by MetPlus DB to acquire external database identifiers including KEGG, HMDB etc. and 54 differentially expressed genes that are interconnected through gene-to-metabolite interactions using ConsensusPathDB [38]. A generalized workflow outlining steps for assessing the significance of integrative network in the context of biological pathways is illustrated in Figure 2.

Pathway analysis was performed on the ConsensusPathDB website using overrepresentation analysis based on Fisher's exact test. The lists of significantly enriched pathways with an adjusted p-value cut-off of 0.05 were retained. The enrichment analysis was performed using KEGG, Reactome, SMPDB, EHMN, and Humancyc.By intersecting the two separate lists of enriched pathways we obtained overlapping biochemical pathways containing both metabolite and gene functional partners found to be altered in the study sets.

Results

We have developed a novel pipeline for integrative analysis of untargeted metabolomics data in conjunction with gene expression data derived from the same samples. A complete workflow for this pipeline is presented in Figure 1.

This methodology includes several consecutive steps for streamlining a process of metabolite annotation (MetPlus DB) as well as downstream systems biology analysis of networks and pathways by incorporating known information about gene-metabolite interactions. This pipeline has been already utilized in our current translational research projects on multi-omics profiling of breast and colorectal cancer (data unpublished).

To illustrate the utility of this methodology we have applied it to the analysis of open access breast cancer data from the NCI-60 collection.

Implementation of MetPlus DB

We have built a fully cross-referenced database (MetPlus DB) by integrating the data from the three most comprehensive metabolite databases tailored largely towards mammalian metabolomics: HMDB, HUAMNCYC and LIPID MAPS. MetPlus DB has cross-referencing information for linking to several other mainstream cheminformatics/bioinformatics repositories including KEGG [39], METLIN [40], ChEBI [41], FooDB [http://www.foodb.ca], Pubchem [42] and Chemspider [http://www.chemspider.com] to provide unambiguous knowledge on clinically and physiologically relevant metabolites. Technical details of database construction are presented in Methods section. MetPlus DB is available for download at (*https://github.com/ICBI/MetPlus-DB*).

There are several features that have made MetPlus DB a customizable and unique alternative to current annotation resources [19, 20]: 1) Non-dependency on remote database availability and schema. Remote databases often change their database schema when releasing newer versions thereby making older versions obsolete. Under such situations, web-based applications with a dependency on remote databases often require changes in their software architecture and/or computational module to tune-in with these resources from time to time. Consequently, reproducibility of the same results from one time or another cannot be guaranteed; 2) Fully customizable, extensible, accessible, and tailored towards supporting a variety of systems biology workflows for annotating massive amounts of metabolomics data; 3) Assembly of a large amount of information useful for searching

annotations specific to mammalians thereby reducing the chances of hitting spurious and inconsequential identification of metabolites of different origins; and 4) Flexible architecture that not only allows for mass-based or external identifier specific searches but also enables searching for metabolites based on a common chemical name, IUPAC name, systematic or trade name. In short, MetPlus DB saves a large amount of time for other research activities that would otherwise be devoted to manually annotate metabolites across multiple sources to filter down to meaningful results for downstream validation purposes.

To demonstrate the utility of our pipeline we used MetPlus DB to re-annotate a list of 26 differentially present metabolites in the case study presented below. MetPlus DB acquires multiple external database identifiers including (e.g. KEGG, HMDB) as shown in Table 1 and enables downstream analysis of metabolite pathway enrichment.

Case Study: Integrative analysis of metabolomics and transcriptomics data for NCI-60 breast cancer cell lines

We tested our computational workflow on two breast cancer cell lines - BT-549 (basal breast cancer) and MCF-7 (luminal breast cancer) using a combination of two complimentary 'omics' datasets – transcriptomics (microarray - gene expression) and metabolomics. The technical details of data processing and network and pathway analysis for this study are presented in the Methods section.

A knowledge-driven network of interactions was generated between differentially expressed genes (DEGs) and differentially identified metabolites (Supplemental File 1) using literature-derived information on chemical-gene interactions from the STITCH knowledge base as illustrated in (Supplemental File 2) and redrawn using cytoscape in the circular layout in Figure 3A. To focus specifically on knowledge-driven networks connected through metabolite and gene interactions, we filtered out the gene-to-gene and metabolite-to-metabolite interactions and re-constructed the network as shown in Figure 3B retaining only metabolite-to-gene interactions. This interaction network indicated significant differences in metabolism and gene expression between two breast cancer cell lines of basal and luminal subtypes.

The network allowed us to clearly identify several interacting functional modules (Figure 3B). The maximum number of connections in this network involved phosphate (24 edges connected to genes) followed by glycerol (12 edges). Identification of metabolites with multiple gene connections could provide additional criteria for selection of potential candidates for targeted metabolomics validation experiments.

Genes in the network were found to have smaller number of connections to metabolites in a range from 1 to 5 metabolites per gene with the majority of genes (34 out of 54 genes) connected to only one metabolite. The maximum number of connections was for NT5E (5 edges) followed by UPP1 and FNP1 (4 edges each).Genes connected to several metabolites could also help identify potential candidate metabolites for validation experiments.

Joint pathway enrichment analysis

To assess the biological significance of an integrative network in the context of significantly affected biological pathways, a pathway enrichment analysis was performed on the list of 26 differentially identified metabolites and 54 differentially expressed genes that are interconnected through gene-to-metabolite interactions using ConsensusPathDB [38]. The metabolites were annotated by MetPlus DB to acquire external database identifiers as shown in Table 1.

The lists of significantly enriched pathways were generated using KEGG, Reactome, SMPDB, EHMN, and Humancyc databases. By intersecting the two lists, overlapping biochemical pathways were obtained and populated with both functional partners (i.e. metabolites and genes as shown in Figure 4). The results of joint pathway analysis presented as a heat-map showing presence of metabolites (highlighted in green color) and genes (highlighted in blue color) in the enriched pathways. Each column represents individual pathways and each row – metabolites or genes.

Noticeably, only few genes were mapped to a large number of enriched metabolic pathways: ABAT (16 pathways), ALDHSA1 (12 Pathways), NT5E (9 pathways). Similarly, only few metabolites were mapped to a large number of pathways: Glutamate (20 pathways), Succinate (15 pathways), beta-Alanine (15 pathways).

Enriched pathways were primarily represented by metabolites belonging to known metabolic pathways. However, several more specific pathways were also represented by multiple molecules, including transmembrane transport of small molecules from Reactome (23 molecules), metabolism of nucleotides (17 molecules), metabolism of amino acids and derivatives (17 molecules), and Urea cycle and metabolism of arginine, proline, glutamate, aspartate and asparagine from the Edinburgh Human Metabolic Network (16 molecules).

Discussion

We discuss a pipeline using open source tools for metabolomics analysis that allows for metabolite integration with transcriptomics data. The integration of both molecular data types provides a more accurate understanding of the biological processes implicated in disease. The novelty of our methodology is the sequential execution of analytical steps allowing for both annotation of metabolomics peaks, and knowledge driven integration of annotated metabolites with gene expression data in a flexible, reusable, and configurable workflow.

The computational workflow is based on loosely coupled analytical modules to support custom data analysis in a rapidly changing environment. The pipeline is easy to implement and reusable for users with some programming skills including database handling and writing basic SQL (Structured Query Language) queries. It utilizes our new resource for annotation of metabolites (MetPlus DB) and knowledge-driven workflow for integrating metabolite abundance data with gene expression data into interaction networks based on prior biological knowledge. A final step of this pipeline allows to us to identify biological pathways that are significantly enriched within these subsets of interacting genes and metabolites. This methodology streamlines the identification of biologically meaningful subsets of data and assists in focusing on potential biological mechanisms that are relevant to observed phenotypic changes.

To demonstrate utility of this approach a case study was conducted on publicly available NCI-60 metabolomics profiling and gene expression data from breast cancer cell lines. The data were compared between two breast cancer cell lines with known distinct phenotypes of basal and luminal types of breast tumors.

A network integrating genes and metabolites was constructed based on the STITCH knowledge base (Figure 3). After subtraction of gene-gene and metabolite-metabolite interactions a gene-metabolite network emerged that allowed us to clearly identify a subset of metabolites connected to multiple DEGs (ranging from 1 to 24 edges). In contrast, the genes were connected to markedly less metabolites

Common name	Inchikey	Monoisotopic weight	Formula	HMDB id	LMDB id	Humancyc Compound name	Pubchem CID	Chemspider id	Chebi id	Metlin id	KEGG id	foodb id
NADP	XJLXINKUBYWONI-NNYOXOHSSA-O	744.0833	C21H29N7O17P3	HMDB00217	--	NADP	5886	5675	18009	5227	C00006	FDB021908
Galactitol	FBPFZTCFMRRESA-GUCUJZIJSA-N	182.079	C6H14O6	HMDB00107	--	--	11850	11357	16813	5148	C01697	FDB006453
L-Aspartic acid	CKLJMWTZIZZHCS-REOHCLBHSA-N	133.0375	C4H7NO4	HMDB00191	--	--	5960; 44367445	5745	17053	5206	C00049	FDB012567
Uracil	ISAKRJDGNUQOIC-UHFFFAOYSA-N	112.0273	C4H4N2O2	HMDB00300	--	Uracil	1174	1141	17568	258	C00106	FDB006426
Glycerol	PEDCQBHIVMGVHV-UHFFFAOYSA-N	92.04734	C3H8O3	HMDB00131	--	Glycerol	753	733	17522; 17754	105	C00116	FDB000756
Succinic acid	KDYFGRWQOYBRFD-UHFFFAOYSA-N	118.0266	C4H6O4	HMDB00254	LMFA01170043	Succinic acid	1110; 21952380	1078	15741	114	C00042	FDB001931
Phosphate	NBIIXXVUZAFLBC-UHFFFAOYSA-K	94.95342	O4P	HMDB01429	--	Phosphate	1061	1032	18367	3231	C00009	FDB022617
L-Glutamic acid	WHUUTDBJXJRKMK-VKHMYHEASA-N	147.0532	C5H9NO4	HMDB00148	--	L-Glutamic acid	33032; 44272391	30572	16015	5174	C00025	FDB012535
S-Adenosylhomocysteine	ZJUKTBDSGOFHSH-WFMPWKQPSA-N	384.1216	C14H20N6O5S	HMDB00939	--	--	439155; 25246222	388301	16680; 57856	296	C00021	FDB022327
Glutathione	RWSXRVCMGQZWBV-WDSKDSIN-SA-N	307.0838	C10H17N3O6S	HMDB00125	--	--	124886; 25246407	111188	16856; 60836	44	C00051	FDB001498
Niacinamide	DFPAKSUCGFBDDF-UHFFFAOYSA-N	122.048	C6H6N2O	HMDB01406	--	Niacinamide	936	911	17154	1497	C00153	FDB012485
Cytidine	UHDGCWIWMRVCDJ-XVFCMESISA-N	243.0855	C9H13N3O5	HMDB00089	--	--	6175	5940	17562	3376	C00475	FDB021809
L-Kynurenine	YGPSJZOEDVAXAB-QMMMGPOBSA-N	208.0848	C10H12N2O3	HMDB00684	--	--	161166; 6971029	141580	16946; 57959	72	C00328	FDB022181
Uridine	DRTQHJPVMGBUCF-XVFCMESISA-N	244.0695	C9H12N2O6	HMDB00296	--	Uridine	6029	5807	16704	90	C00299	FDB007411
N-Acetylneuraminic acid	SQVRNKJHWKZAKO-PFQGKNLYSA-N	309.106	C11H19NO9	HMDB00230	--	--	445063	392810	45744	3321	C19910	FDB001209
Hydroxyproline	PMMYEEVYMWASQN-DMTCNVIQ-SA-N	131.0582	C5H9NO3	HMDB00725	--	--	5810; 6971053	5605	18095; 58375	257	C01157	FDB013511
L-Valine	KZSNJWFQEVHDMF-BYPYZUCNSA-N	117.079	C5H11NO2	HMDB00883	--	L-Valine	6287; 6971018	6050	16414; 57762	5842	C00183	FDB000465
Cholesterol	HVYWMOMLDIMFJA-DPAQBDIFSA-N	386.3549	C27H46O	--	LMST01010001	--	5997	5775	16113	--	C00187	--
Cholesterol	HVYWMOMLDIMFJA-FNOPAARDSA-N	386.3549	C27H46O	HMDB00067	--	--	11025495	9200676	--	163	C00187	FDB013269
Cholesterol(d7)	HVYWMOMLDIMFJA-IFAPJKRJSA-N	393.3988	C27H39D7O	--	LMST01010093	--	5314029	4473448	--	--	C00187	--
5'-Methylthioadenosine	WUUGFSXJNOTRMR-IOSLPCCCSA-N	297.0896	C11H15N5O3S	HMDB01173	--	5'-Methylthioadenosine	439176	388321	17509	3425	C00170	FDB022465
L-Alanine	QNAYBMKLOCPYGJ-REOHCLBHSA-N	89.04768	C3H7NO2	HMDB00161	--	L-Alanine	5950; 7311724	5735	16977; 57972	--	C00041	FDB000556
L-Leucine	ROHFNLRQFUQHCH-YFKPBYRVSA-N	131.0946	C6H13NO2	HMDB00687, HMDB13773	--	--	6106; 7045798	5880	15603; 57427	24	C00123	FDB001946
L-Glutamine	ZDXPYRJPNDTMRX-VKHMYHEASA-N	146.0691	C5H10N2O3	HMDB00641	--	--	5961; 6992086	5746	18050; 58359	5614	C00064	FDB012164
Guanosine	NYHBQMYGNKIUIF-UUOKFMHZSA-N	283.0917	C10H13N5O5	HMDB00133	--	--	6802	6544	16750	87	C00387	FDB003632
Inosine	UGQMRVRMYYASKQ-KQYNXXCU-SA-N	268.0808	C10H12N4O5	HMDB00195	--	Inosine	6021	5799	17596	84	C00294	FDB011802
Glucose 6-phosphate	NBSCHQHZLSJFNQ-GASJEMHNSA-N	260.0297	C6H13O9P	HMDB01401	--	--	5958	5743; 17216117	4170	145	C00092	FDB021818
Beta-Alanine	UCMIRNVEIXFBKS-UHFFFAOYSA-N	89.04768	C3H7NO2	HMDB00056	--	Beta-Alanine	239; 4755801	234	16958; 57966	5119	C00099	FDB002253

List of 26 metabolites annotated using MetPlus DB directly involved in gene-to-metabolite interactions within the knowledge-driven integrated network

Table 1: Breast cancer metabolites annotated using MetPlus DB.

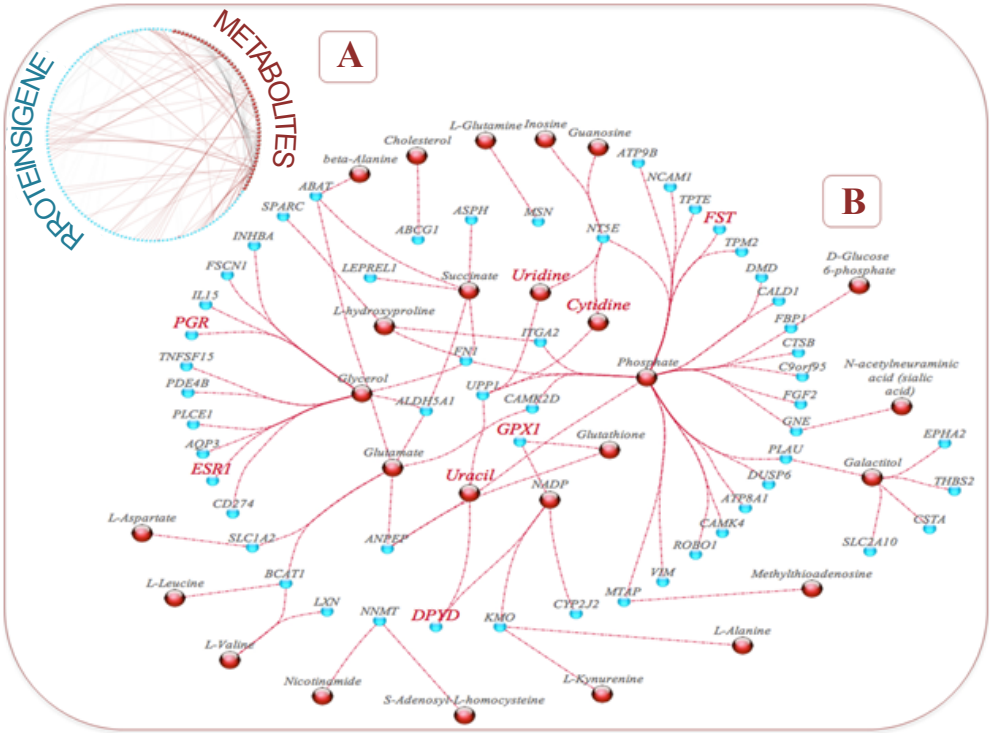

Figure 3: The integrative knowledge-driven network combining two 'omics' technologies including microarray gene expression and metabolomics.A) the circular view of entire network showing gene-to-metabolite, gene-to-gene and metabolite-to-metabolite interactions. The network is color-coded with nodes in red color corresponds to metabolites and cyan color genes. Node label highlighted in red color corresponds to CTD curated genes and metabolites associated with breast cancer. Grey color solid edges corresponds to undirected gene-to-gene and metabolite-to-metabolite interactions, and dotted red color undirected gene-to-metabolites interactions. B) Sub-network view showing exclusive gene-to-metabolite interactions to gain novel insights into the molecular mechanisms of breast cancer dynamics.

(ranging from 1 to 5 per gene) with majority of genes (34 out of 54 genes) connected to only one metabolite. Identification of metabolites with high connectivity to genes provides selection criteria for potential small molecule biomarkers associated with phenotype differences between basal and luminal breast cancer cell lines. In this case study we found that phosphate has the largest number of connections in the network with 26 genes. Differences in phosphate metabolite levels has been reported in the early 31P NMR study [43] showing that human breast cancer cells with the phenotype of pleiotropic drug resistance (derivative from wt MCF7 cell line) exhibited significantly elevated levels of phosphate metabolites as compared to the wild-type, drug-sensitive parent cell line of wt MCF7 cells. Interestingly, in our case study we have compared the MCF7 cell line with the BT-549, a basal subtype B triple negative cell line [44] that has been reported to have a multi-drug resistant phenotype [45].

The differential expression of genes connected to multiple metabolites (such as NT5E with 5 connections) could indicate a key role in the alteration of several metabolic pathways associated with specific breast cancer cell line phenotypes. Connectivity of genes might serve as additional criteria for designing follow up experiments.

The last step of our pipeline, a joint analysis of pathway enrichment, enables the identification of specific metabolic and regulatory pathways that are significantly altered between the two cell lines. Such an approach provides an opportunity to expand analysis beyond interactions between single entities and identify those genes and metabolites that are involved in the same pathways where there might be no direct association or interaction between them (Figure 4). Using this approach for analysis of breast cancer cell lines we have identified several genes that were mapped to many enriched pathways: ABAT (16 pathways), ALDHSA1 (12 Pathways), NT5E (9 pathways). While some of these genes have multiple connections in metabolite-gene networks as well (such as NT5E), other genes mapped to multiple pathways did not have large number of connections to metabolites in the network (ABAT – only 3 connections).

During a recent validation study on 42 breast cancer cell lines the role of ABAT expression in breast cancer has been reported as part of multi-gene predictors for use in capturing response to chemotherapy [46]. ABAT was also shown to contribute to the mechanisms of mammary tumor progression [47].

These observations have clearly demonstrated the complementarity of the two integrative analysis steps incorporated into our pipeline. In addition, the joint pathway enrichment analysis step helps to reveal relatively important (and not obvious) roles for genes and metabolites participating in multiple pathways relevant to a specific phenotype. Such differences can only be obtained by integrative analysis of two data types together.

While specific findings from this case study require follow up validation, the results clearly demonstrate that such a systems biology approach for the analysis of metabolomics data offers a unique way to connect data within a biological context. The methodology also provides novel criteria for selection of candidate biomarkers based on joint analysis of differentially expressed genes and metabolites.

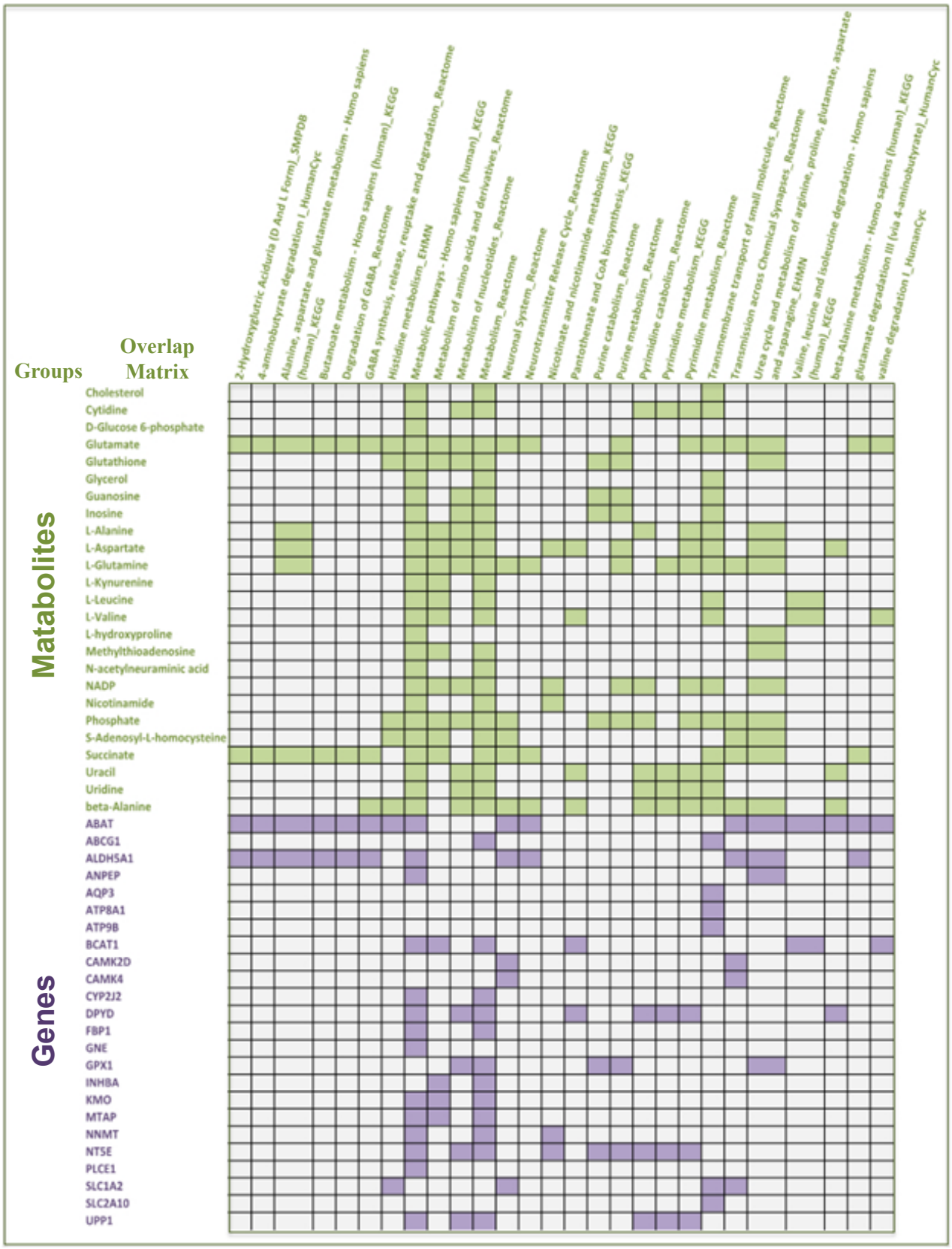

Figure 4: Diagram of Enriched Overlapping Pathways. Adj. P-value < 0.05 identified by Pathway Enrichment Analysis for metabolites and genes. Colored cells indicate pathways containing corresponding metabolite (green) or genes (blue).

Future Directions

Our novel integrative analysis pipeline provides a streamlined approach for integrating untargeted metabolomics data with transcriptomics data. The key strength of the pipeline is in the potential for broader applicability for integrating metabolomics data with other complementary 'omics' data types to allow for rapid generation of new testable hypotheses based on the interactions of functional partners (genes, proteins, miRNA and metabolites etc.) in the knowledge-driven network. The workflow with some modifications can be extended to integration of metabolomics data with other omics data such as proteomics and genomics (e.g. DNA-seq variant data and microRNA profiling data).

Several steps of pipeline presented here have been recently incorporated into the metabolomics analysis module of the Georgetown Database of Cancer (G-DOC).Downstream analysis steps will eventually be incorporated into G-DOC and made available for public use on cancer datasets currently in G-DOC. We are also exploring the analysis of human metabolomics data from bio-fluids (blood, urine) and plan to integrate this with other types of omics data types.

MetPlus DB as a standalone SQLite database is developed to facilitate external identifier or mass-based searches from the collection of comprehensive metabolite databases. The data is primarily integrated by merging records based on IUPAC International Chemical Identifier (InChIKey) to deliver high confidence annotations with a significantly reduced redundancy and eliminates the complexity of extracting metabolite annotations from individual databases. To further extend the coverage on mammalian specific metabolites of clinical and physiological importance, more databases can be integrated to provide annotation-rich and curated knowledgebase of mammalian metabolomes for putative identification of metabolites.

Although several current methods are available for accurate metabolite identification and annotation, the process is cumbersome and involves curation using multiple databases. We describe novel computational solutions for streamlining the process of metabolite annotation using MetPlus DB, and the subsequent integration of metabolomics and transcriptomics data to explore potentially relevant biological interactions and candidate biomarkers associated with disease phenotype.

Availability of Supporting Data

The data sets supporting the results of this article are available in the DTP Web portal: http://dtp.nci.nih.gov/mtargets/download.html; (August 2010 Data Release) and expression profiles downloaded from Gene Expression Omnibus (GEO; accession number GSE32474.

MetPlus DB is located in GitHub: https://github.com/ICBI/MetPlus-DB. Instructions for using MetPlusdb are available on Github to load and run the SQLite database, and examples are provided for querying annotations.

Acknowledgements

This research was supported by the National Cancer Institute (NCI) of the National Institutes of Health under Cooperative Agreement: U54-CA149147; and the NCI *in silico* Research Center of Excellence (ISRCE) contract: HHSN26120080000IE.

We would also like to thank Dr. Laura Sheahan for editing the manuscript and providing invaluable advice and support.

References

1. Spratlin JL, Serkova NJ, Eckhardt SG (2009) Clinical applications of metabolomics in oncology: a review. Clin Cancer Res 15: 431-440.
2. Corona G, Rizzolio F, Giordano A, Toffoli G (2012) Pharmaco-metabolomics: an emerging "omics" tool for the personalization of anticancer treatments and identification of new valuable therapeutic targets. J Cell Physiol 227: 2827-2831.
3. Li M, Song Y, Cho N, Chang JM, Koo HR, et al. (2011) An HR-MAS MR metabolomics study on breast tissues obtained with core needle biopsy. PLoS One 6: e25563.
4. Bathen TF, Jensen LR, Sitter B, Fjösne HE, Halgunset J, et al. (2007) MR-determined metabolic phenotype of breast cancer in prediction of lymphatic spread, grade, and hormone status. Breast Cancer Res Treat 104: 181-189.
5. Giskeødegård GF, Grinde MT, Sitter B, Axelson DE, Lundgren S, et al. (2010) Multivariate modeling and prediction of breast cancer prognostic factors using MR metabolomics. J Proteome Res 9: 972-979.
6. Porto L (2012) Cutoff value of choline concentration reliably reveals high-grade brain tumors among other contrast-enhancing brain lesions. J NeurolSurg A Cent EurNeurosurg. 73(3): p. 147-52.
7. Howe FA, Barton SJ, Cudlip SA, Stubbs M, Saunders DE, et al. (2003) Metabolic profiles of human brain tumors using quantitative in vivo 1H magnetic resonance spectroscopy. MagnReson Med 49: 223-232.
8. Bartella L, Morris EA, Dershaw DD, Liberman L, Thakur SB, et al. (2006) Proton MR spectroscopy with choline peak as malignancy marker improves positive predictive value for breast cancer diagnosis: preliminary study. Radiology 239: 686-692.
9. Thakur SB, Brennan SB, Ishill NM, Morris EA, Liberman L, et al. (2011) Diagnostic usefulness of water-to-fat ratio and choline concentration in malignant and benign breast lesions and normal breast parenchyma: an in vivo (1) H MRS study. J Magn Reson Imaging 33: 855-863.
10. Swanson MG, Keshari KR, Tabatabai ZL, Simko JP, Shinohara K, et al. (2008) Quantification of choline- and ethanolamine-containing metabolites in human prostate tissues using 1H HR-MAS total correlation spectroscopy. MagnReson Med 60: 33-40.
11. Roberts MJ, Schirra HJ, Lavin MF, Gardiner RA (2011) Metabolomics: a novel approach to early and noninvasive prostate cancer detection. Korean J Urol 52: 79-89.
12. Brockmöller SF, Bucher E, Müller BM, Budczies J, Hilvo M, et al. (2012) Integration of metabolomics and expression of glycerol-3-phosphate acyltransferase (GPAM) in breast cancer-link to patient survival, hormone receptor status, and metabolic profiling. J Proteome Res 11: 850-860.
13. Cavill R, Kamburov A, Ellis JK, Athersuch TJ, Blagrove MS, et al. (2011) Consensus-phenotype integration of transcriptomic and metabolomic data implies a role for metabolism in the chemosensitivity of tumour cells. PLoS Comput Biol 7: e1001113.
14. García-Alcalde F, García-López F, Dopazo J, Conesa A (2011) Paintomics: a web based tool for the joint visualization of transcriptomics and metabolomics data. Bioinformatics 27: 137-139.
15. Wägele B, Witting M, Schmitt-Kopplin P, Suhre K (2012) MassTRIX reloaded: combined analysis and visualization of transcriptome and metabolome data. PLoS One 7: e39860.
16. Adourian A, Jennings E, Balasubramanian R, Hines WM, Damian D, et al. (2008) Correlation network analysis for data integration and biomarker selection. Mol Biosyst 4: 249-259.
17. Connor SC, Hansen MK, Corner A, Smith RF, Ryan TE (2010) Integration of metabolomics and transcriptomics data to aid biomarker discovery in type 2 diabetes. Mol Biosyst 6: 909-921.
18. Kamburov A, Cavill R, Ebbels TM, Herwig R, Keun HC (2011) Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA. Bioinformatics 27: 2917-2918.
19. Zhou B, Wang J, Ressom HW (2012) MetaboSearch: tool for mass-based metabolite identification using multiple databases. PLoS One 7: e40096.
20. Brown M, Dunn WB, Dobson P, Patel Y, Winder CL, et al. (2009) Mass spectrometry tools and metabolite-specific databases for molecular identification in metabolomics. Analyst 134: 1322-1332.

21. Moco S, Vervoort J (2007) Metabolomics technologies and metabolite identification. TrAC Trends in Analytical Chemistry. 26: 855-866.

22. Jewison T, Knox C, Neveu V, Djoumbou Y, Guo AC, et al. (2012) YMDB: the Yeast Metabolome Database. Nucleic Acids Res 40: D815-D820.

23. Scalbert A, Brennan L, Fiehn O, Hankemeier T, Kristal BS, et al. (2009) Mass-spectrometry-based metabolomics: limitations and recommendations for future progress with particular focus on nutrition research. Metabolomics 5: 435-458.

24. Wishart DS, Jewison T, Guo AC, Wilson M, Knox C, et al. (2013) HMDB 3.0--The Human Metabolome Database in 2013. Nucleic Acids Res 41: D801-D807.

25. Fahy E, Sud M, Cotter D, Subramaniam S (2007) LIPID MAPS online tools for lipid research. Nucleic Acids Res 35: W606-W612.

26. Romero P, Wagg J, Green ML, Kaiser D, Krummenacker M, et al. (2005) Computational prediction of human metabolic pathways from the complete human genome. Genome Biol 6: R2.

27. Baran R, Kochi H, Saito N, Suematsu M, Soga T, et al. (2006) MathDAMP: a package for differential analysis of metabolite profiles. BMC Bioinformatics 7: 530.

28. Tikunov Y, Lommen A, de Vos CH, Verhoeven HA, Bino RJ, et al. (2005) A novel approach for nontargeted data analysis for metabolomics. Large-scale profiling of tomato fruit volatiles. Plant Physiol 139: 1125-1137.

29. Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G (2006) XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. Anal Chem 78: 779-787.

30. Katajamaa M, Oresic M (2005) Processing methods for differential analysis of LC/MS profile data. BMC Bioinformatics 6: 179.

31. Kuhn M, Szklarczyk D, Franceschini A, von Mering C, Jensen LJ, et al. (2012) STITCH 3: zooming in on protein-chemical interactions. Nucleic Acids Res 40: D876-880.

32. Davis AP, Murphy CG, Johnson R, Lay JM, Lennon-Hopkins K, et al. (2013) The Comparative Toxicogenomics Database: update 2013. Nucleic Acids Res 41: D1104-1114.

33. Pfister TD, Reinhold WC, Agama K, Gupta S, Khin SA, et al. (2009) Topoisomerase I levels in the NCI-60 cancer cell line panel determined by validated ELISA and microarray analysis and correlation with indenoisoquinoline sensitivity. Mol Cancer Ther 8: 1878-1884.

34. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, et al. (2003) Summaries of AffymetrixGeneChip probe level data. Nucleic Acids Res 31: e15.

35. Gentleman R, Carey V, Huber W, Hahne F (2003) genefilter: methods for filtering genes from microarray experiments.

36. Dabney A, Storey JD, Gregory R (2004)Warnesqvalue: Q-value estimation for false discovery rate control.

37. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T (2011) Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics 27: 431-432.

38. Kamburov A, Wierling C, Lehrach H, Herwig R (2009) ConsensusPathDB--a database for integrating human functional interaction networks. Nucleic Acids Res 37: D623-D628.

39. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, et al. (2006) From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res 34: D354-357.

40. Smith CA, O'Maille G, Want EJ, Qin C, Trauger SA, et al. (2005) METLIN: a metabolite mass spectral database. Ther Drug Monit 27: 747-751.

41. Degtyarenko K, de Matos P, Ennis M, Hastings J, Zbinden M, et al. (2008) ChEBI: a database and ontology for chemical entities of biological interest. Nucleic Acids Res 36: D344-D350.

42. Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, et al. (2009) PubChem: a public information system for analyzing bioactivities of small molecules. Nucleic Acids Res 37: W623-W633.

43. Cohen JS (1986) Differences in phosphate metabolite levels in drug-sensitive and -resistant human breast cancer cell lines determined by 31P magnetic resonance spectroscopy. Cancer Res 46: 4087-4090.

44. Kao J, Salari K, Bocanegra M, Choi YL, Girard L, et al. (2009) Molecular profiling of breast cancer cell lines defines relevant tumor models and provides a resource for cancer gene discovery. PLoS One 4: e6146.

45. Bartholomeusz C, Yamasaki F, Saso H, Kurisu K, Hortobagyi GN, et al. (2011) Gemcitabine Overcomes Erlotinib Resistance in EGFR-Overexpressing Cancer Cells through Downregulation of Akt. J Cancer 2: 435-442.

46. Shen K, Song N, Kim Y, Tian C, Rice SD, et al. (2012) A systematic evaluation of multi-gene predictors for the pathological response of breast cancer patients to chemotherapy. PLoS One 7: e49529.

47. Kretschmer C, Sterner-Kock A, Siedentopf F, Schoenegg W, Schlag PM, et al. (2011) Identification of early molecular markers for breast cancer. Mol Cancer 10: 15.