Published in final edited form as:
Nat Prod Rep. 2013 April ; 30(4): 565–583. doi:10.1039/c3np20111b.

A global approach to analysis and interpretation of metabolic data for plant natural product discovery†

Manhoi Hur[a], **Alexis Ann Campbell**[b], **Marcia Almeida-de-Macedo**[c], **Ling Li**[d], **Nick Ransom**[e], **Adarsh Jose**[f], **Matt Crispin**[g], **Basil J. Nikolau**[h], and **Eve Syrkin Wurtele**[i]

[a]Human Computer Interactions and Department of Genetics Development and Cell Biology, 2624 Howe Hall, Iowa State University, Ames, IA 50011, USA. Fax: +1 515 294 0803; Tel: +1 515 708 3232; mhhur@iastate.edu

[b]Biochemistry, Biophysics and Molecular Biology and Center for Biorenewable Chemicals and Center for Metabolic Biology, 3254 Molecular Biology Building, Iowa State University, Ames, IA 50010, USA. Fax: +1 515 294 9423; Tel: +1 515 294 0453; alexis.a.campbell@gmail.com

[c]Department of Genetics Development and Cell Biology, 2624 Howe Hall, Iowa State University, Ames, IA 50011, USA. Fax: +1 515 294 5530; Tel: +1 515 294 3738; demacedo@iastate.edu

[d]Department of Genetics Development and Cell Biology, 443 Bessey Hall Iowa State University, Ames, IA 50011, USA. Fax: +1 515 294 1337; Tel: +1 515 294 6236; liling@iastate.edu

[e]Department of Genetics Development and Cell Biology, 2624 Howe Hall, Iowa State University, Ames, IA 50011, USA. Fax: +1 515 294 0803; Tel: +1 515 708 3232; rannic@gmail.com

[f]Bioinformatics and Computational Biology, Center for Biorenewable Chemicals, Iowa State University, Ames, IA 50010, USA. Fax: +1 515 294 1269; Tel: +1 515 230 3429; ajose@iastate.edu

[g]Department of Genetics Development and Cell Biology, 443 Bessey Hall Iowa State University, Ames, IA 50011, USA. Fax: +1 515 294 1337; Tel: +1 515 294 6236; crispy11@iastate.edu

[h]Biochemistry, Biophysics and Molecular Biology and Center for Biorenewable Chemicals and Center for Metabolic Biology, 3254 Molecular Biology Building, Iowa State University, Ames, IA 50010, USA. Fax: +1 515 294 9423; Tel: +1 515 294 0453; dimmas@iastate.edu

[i]Department of Genetics, Development and Cell Biology, Center for Metabolic Biology, and Center for Biorenewable Chemicals, 2624D Howe Hall, Iowa State University, Ames, IA 50011, USA. Fax: +1 515 294 0803; Tel: +1 515 708 3232; mash@iastate.edu

Abstract

Discovering molecular components and their functionality is key to the development of hypotheses concerning the organization and regulation of metabolic networks. The iterative experimental testing of such hypotheses is the trajectory that can ultimately enable accurate computational modelling and prediction of metabolic outcomes. This information can be particularly important for understanding the biology of natural products, whose metabolism itself is often only poorly defined. Here, we describe factors that must be in place to optimize the use of metabolomics in predictive biology. A key to achieving this vision is a collection of accurate time-resolved and spatially defined metabolite abundance data and associated metadata. One

†The opinions expressed in this review are those of the authors, not necessarily those of the US Government.

Correspondence to: Eve Syrkin Wurtele.

formidable challenge associated with metabolite profiling is the complexity and analytical limits associated with comprehensively determining the metabolome of an organism. Further, for metabolomics data to be efficiently used by the research community, it must be curated in publically available metabolomics databases. Such databases require clear, consistent formats, easy access to data and metadata, data download, and accessible computational tools to integrate genome system-scale datasets. Although transcriptomics and proteomics integrate the linear predictive power of the genome, the metabolome represents the nonlinear, final biochemical products of the genome, which results from the intricate system(s) that regulate genome expression. For example, the relationship of metabolomics data to the metabolic network is confounded by redundant connections between metabolites and gene-products. However, connections among metabolites are predictable through the rules of chemistry. Therefore, enhancing the ability to integrate the metabolome with anchor-points in the transcriptome and proteome will enhance the predictive power of genomics data. We detail a public database repository for metabolomics, tools and approaches for statistical analysis of metabolomics data, and methods for integrating these dataset with transcriptomic data to create hypotheses concerning specialized metabolism that generates the diversity in natural product chemistry. We discuss the importance of close collaborations among biologists, chemists, computer scientists and statisticians throughout the development of such integrated metabolism-centric databases and software.

1 Introduction

The metabolome of a biological sample defines the steady-state levels of the intermediates and end products of the metabolic networks that constitute that sample. Thus, metabolomic data reflect the ultimate expression (output) of a genome at the metabolic level.[1, 2] It follows, therefore, that by comparing the metabolomes of two samples that differ in their metabolic outputs, one gains insights as to the structure of the metabolic network that supports the metabolic outcome of these samples.

Moreover, because the structure of the metabolic network is the result of the programmatic expression of the genome, modified by environmental inputs, metabolomics data, integrated with additional ‘omics levels datasets, can provide insights into the systems level control and regulation of metabolic outcomes. For example, the quantitative determination of the metabolomes of tissues/organs that express different levels of a specific metabolic end-point, when integrated with additional –omics level expression profiles can facilitate the identification of genes/enzymes that are components of the biosynthetic pathway supporting that metabolic end-point. In the extreme, some specialized natural products of plant are synthesized and accumulate in dedicated structures (e.g., trichomes, glands, laticifers). Presupposing that there is no intercellular trafficking involved in the biosynthesis of the targeted metabolite, the metabolomes of the cells that hyperaccumulate the target metabolite will be populated by metabolic intermediates of its biosynthesis.

In a simple metabolic model, in which biosynthetic capacity is determined by transcriptional regulation, one would anticipate that the relative abundance of transcripts encoding enzymes involved in that biosynthetic pathway is proportional to the level of the product of that pathway. In this case, it would be a statistically straightforward task to correlate transcript levels to products of metabolism and to assign function to gene responsible to metabolism. However, the regulatory complexity of the interrelationships among genes, gene products, and metabolites often confounds the interpretation of multivariate datasets.

A second challenge to integrating transcriptomics and metabolomics data is the asymmetric nature of the analytical technologies that capture these datasets. Whereas RNA-seq technologies have the sensitivity to determine nearly the entire transcriptome of a sample, current metabolomics technologies are far from such capabilities. The plant metabolome has been estimated to be as large as 200,000 compounds,[3] but metabolomics datasets are likely 50 to a few thousand compounds, many of which are not chemically identified. Yet, metabolomics data is an absolute requirement for deciphering the functionality of transcripts and their translated protein products. Despite these limitations, the literature is replete with the successful findings from studies that are based upon correlations among transcript abundance data with the corresponding metabolic product.[4–21] For example, the relative abundance of transcripts encoding alkaloid biosynthetic enzymes correlate well with the induction of benzylisoquinoline accumulation in *Papaver somniferum*.[4, 17] Comparison of transcriptomes and metabolomes (particularly fatty acid and lipids) of developing seeds that accumulate “unusual” fatty acids has led to the discovery of a series of FAD2-related enzymes[6–11] that are responsible for the generation of hydroxy fatty acids,[6] epoxy-fatty acids,[7, 8] conjugated fatty acids,[9, 10] and acetylenic fatty acids.[7, 11] These fatty acids accumulate in taxonomically discrete clades, and they have properties that make them attractive commodity chemicals in industrial applications (e.g., as coatings, surfactants and varnishes).[22] Hence, comparing transcriptomics and metabolomics data has enabled the identification of a wide variety of genes.

A general framework for a metabolomics database, including the importance of data consistency and deposition of full metadata, has been described.[23] However, at present, only a few publically accessible metabolomics databases exist; most of these contain datasets from carefully defined samples with a common biological theme. For example, one such database, *Plant Metabolomics* (http://www.plantmetabolomics.org/),[24, 25] contains metabolomics data from Arabidopsis seedlings representing 200 mutants in genes of unknown function, and mutations in these genes do not show an obvious morphological phenotype.[24] Hence, the sole criteria for additional research on these mutants would be metabolic differences that are revealed to the research community via this data. Such data has enabled research on the role of novel plant lipids, such as Lipid A, a lipid that is considered unique to gram-positive bacteria,[26] the role of enzyme redundancies associated with FAE1-like[27] and ELO-like[28] fatty acid elongase components of Arabidopsis, and informed novel evolutionary and functional insights into the non-enzymatic FAP proteins.[29] *AtMetExpress*, (http://prime.psc.riken.jp/lcms/AtMetExpress/)[30] contains data and comprehensive metadata from carefully defined organs and developmental stages of Arabidopsis (for which microarray data is available from *AtGeneExpress*),[31] as well as from 20 ecotypes of this species. A third example, The *Medicinal Plant Metabolomics Resource* (*MPMR*, http://www.medicinalplantmetabolimics.org/),[32, 33] presents metabolomics data for 12 species. Its companion database, Medicinal Plant Genomics Resource (http://medicinalplantgenomics.msu.edu/) contains transcriptomics data from the same biological samples. To date, these resources have supported the identification of metabolic intermediates, reactions and genes from medicinal species including: identification of a gene encoding a cytochrome P450 which catalyzes a step in the synthesis of the alkaloid 19-O-acetylhorhammericine in *Catharanthus roseus*;[34] identification of unusual phloroacylglucinols in *Hypericum gentianoides*;[33] characterization of the evolutionary origin of different accessions of *Prunellla vulgaris*;[32] insight into the stereospecificity of quinoline alkaloid synthesis in *Camptotheca acuminata*;[32] cloning of three enzymes of cardenolide synthase, C4 sterol methyloxidase, and progesterone 5b-reductase from *Digitalis purpurea*;[35] and identification of genes encoding valerena-1,10-diene synthase in *Valeriana officinalis*,[32] and their role in the synthesis of sesquiterpenes in that species that have biological activities in mammals.[36]

Here, we describe approaches to construction of a metabolomics database that is scalable, flexible and can support researchers in deposition of metabolomics data and metadata. (In this context, metadata includes information detailing the growth conditions, biological material sampled, experimental protocols, and statistical and computational methods.) We also discuss methods for statistical analysis and visualization of metabolomics data; the usefulness of interactivity among data and metadata in data analysis; and statistical and technical approaches to integrate metabolomics data with other data types. We detail the importance of involving biologists, chemists, computer scientists and statisticians at every stage of development. These key features can enable intuitive use of the metabolomics data by researchers, and rapid and meaningful data analysis. Such capabilities empower researchers to create testable hypotheses concerning biological networks, energy flow in living organisms, or the ontogeny of the specialized metabolism that generates diversity in natural product chemistry.

2 Development of standardized, public, metabolomics databases

To best enable the use of metabolomics data by researchers worldwide, it is important to house them in public databases that provide a venue for researchers to deposit and analyze metabolomics data and metadata, and other data types. In planning a database and its associated user interface and software, the intended uses of the deposited data need to be kept in mind. Programming a database that will effectively house and efficiently retrieve data growing number of species, sample conditions, genotypes, analytical platforms, and metadata types, and also provide a user-friendly, adaptable interface by which biochemists, chemists and biologists can analyze the data, requires careful design and implementation. In general, a more flexible and generalizable database is more difficult to design and implement than a hard-coded database, however, flexibility and generalizability for metabolomics is important because it can better adapt to the rapid pace of technological improvements in data analysis and computation.

As a result of these considerations, the combined expertise of computer scientists, analytic chemists and biologists/biochemists is required in the planning stage and thereafter. Also important are iterative cycles of alpha-testing by a wide variety of intended users, followed by improvements by the designers/programmers. However, there are almost always large gaps in understanding of concepts and terminology used by different disciplines. Thus, developing a useful database requires considerable investment of time in careful planning and mutual understanding of aims and capabilities among computer scientists, biologists, statisticians and chemists, as well as constant communication among these groups during implementation.

Another aspect that is critical to a useful database is data and metadata quality. A database can only be as good as the quality of data and metadata it contains. To maximize the utility of the database, such that researchers most benefit by analysis of metabolomics data at a systems level, the members of each research group that contribute data must be responsible for ensuring submission of complete and clear data and metadata. One possible solution, which would require a bit more from reviewers, would be to require entry of metabolomics data and metadata prior to publication in a database; the entry would be cleared for completeness and released to the public at the time of acceptance of the manuscript for publication.

3 Approaches and innovation in metabolomics data storage and visualization

3.1 Metabolomics data storage

Combining metabolomics data from targeted and non-targeted analytical platforms is often a good approach for evaluating the composition of biological material.[24, 25] GC-MS platforms are typical for targeted metabolomic analyses because of the existing established protocols and the simplicity of chemical interpretation. LC-TOF MS, LC-QTOF MS and FT-ICR MS platforms are widely used for non-targeted metabolomics, as they provide data for large numbers of metabolites that can be concurrently evaluated in a scan. Storing data derived from multiple platforms requires additional considerations regarding the formats of the metadata and the metabolomics data itself.[23, 37]

Typically, metabolomics databases use generic file systems and/or Structured Query Language (SQL)-based database systems for storing and analyzing processed data and/or raw data. Several existing web resources enable researchers to store data, share data, and/or retrieve data in a flat file format. For example, researchers can download plant metabolomics data derived from a variety of analytical platforms from *Data Resources Of Plant Metabolomics* (*DROP Met,* http://prime.psc.riken.jp/?action=drop_index)(which includes the AtMetExpress datasets)) stores metabolomics data and any associated metadata on a file server, and provides links to a variety of files, including raw GC/MS data (ANDI-MS/NetCDF file (*.cdf)), excel files (*.xlsx), and simple text files (*.txt). Researchers can also submit their own metabolomics data to *DROP Met*.[30] The recently constructed *MedPITranscriptome*[38] stores metabolomics data, from two analytical platforms, of 23 medicinal plant species as simple text and excel files.

MetabolomeExpress (https://www.metabolome-express.org/) is a data resource for GC/MS, GC/Quadrupole MS, and GC/TOF MS platforms that uses a File Transfer Protocol (FTP) repository to store experimental data and enables virtual data exchange, combined with an SQL-based database which stores metadata and metabolite response statistics.[39] A significant feature is that researchers can upload the data as raw GC/MS files and/or processed data to the FTP repository. When a user uploads raw data, *MetabolomeExpress* is able to process it, using MSRI (Mass-spectral and retention-index) Library Matching algorithm.[39] The user can also download metabolomics data in several data formats.

SetupX (http://fiehnlab.ucdavis.edu:8080/m1/) stores experimental data and its metadata in SQL-based database systems (MySQL, Oracle DB and native XML DB).[40] According to the report by SetupX, this resource contains data from 75 species and provides annotated compounds that have been identified by BinBase,[41] which it utilizes for automated metabolite annotation. Furthermore, *SetupX* allows the sharing of files, such as documents, metadata, and image files uploaded by researchers. Researchers are able to download metabolomics data as raw GC/MS and GC/TOF MS files from the site. A significant feature is that after installing the *SetupX* program (https://code.google.com/p/setupx), researchers can use it to analyze their private data and its metadata.

Plant Metabolomics data is linked to phenotypic information and data concerning gene function[24] To store data and its metadata, *Plant Metabolomics* employs an SQL-based database, MySQL. This resource contains data from 9 experiments and 8 analytical platforms (Ceramide; Fatty Acids; Amino Acids; Phytosterols; Isoprenoids; Lipidomics; Non-targeted GC-TOF; CE-MS). Researchers can download the data and metadata in a flat file format.

Initial metabolomics datasets were somewhat limited in size, containing information on tens to a few thousand of metabolites (only some of which were chemically identified), thus relatively little storage was required and conventional flat file or SQL-based storage was sufficient to store and retrieve the data. With the acceleration of metabolomics technologies and the increased accessibility of these technologies, metabolomics datasets are capturing information on thousands to tens of thousands of known and unknown compounds. In addition, if metabolomics and transcriptomics data (obtained from the same biological materials) are to be integrated, further increases in storage and access capacity are needed.

Because metabolomics and transcriptomics have different types of metadata (for example, molecular mass and structural metadata versus sequence metadata), it is conceptually a challenge to design and implement a schema for an integrated database.

To address the rapidly expanding amount of data, *Plant Metabolomics Resource* (*PMR*; http://www.metnetdb.org/pmr/), a database that is newly established for metabolomics and transcriptomics data and associated metadata for plants and microorganisms, has been designed and implemented. *PMR* incorporates a data storage system that considers scalability of data storage and facile data integration with transcriptomics data. It also emphasizes the analysis of data and its associated metadata, which is rich in information concerning each sample, metabolite and transcript. *PMR* uses a unique hybrid storage system that combines the SQL-based MySQL database with a NoSQL-based database (http://nosql-database.org/).[42–44] NoSQL (also called "Not only SQL") is a new database platform designed to address the storage and retrieval of non-relational, distributed, and horizontally scalable data. NoSQL is schema-free, as opposed to more traditional databases that are built primarily using tables and rely on queries in SQL. As such, NoSQL databases are highly optimized for retrieve-and-append operations, making them ideal for large volumes of data. NoSQL databases fall into several categories according to the way they store the data, three of the more common being: key–value stores (schema-free and compatible with many programming languages);[45] document-oriented databases (information is stored as a flexible "document", and is usually encoded into JSON (JavaScript Object Notation) or BSON (Binary JSON) format.);[46, 47] and graph and network databases (ideal for data that has strong relationships among entities).[48]

With the increasing quantity and variety of biological data generated by metabolomics and transcriptomics analyses, there is a need to provide curated information from large volumes of data with high-performance and interoperability. Therefore, in order to construct the *PMR* database, we used a document-oriented NoSQL database, MongoDB (10gen Co., USA).[49–51] Relational database models such as SQL require a given relationship model be established among the entities. They also require, for performance reasons, organizing and optimizing the fields and tables in the database. In contrast, document-oriented storage is already optimized to: 1) readily integrate diverse metadata (between metabolite and transcript entities) without considering the relationships among these entities in the database design; 2) reduce development and design time for the programmers of the database, analysis, and visualization; and 3) enable an expandable storage system that contains the data (in our case, e.g., levels of metabolites, and transcripts (RNA-seq or microarray) and metadata (sequence data, motif information, Gene Ontology(GO) annotations, pathway/network information, regulon membership, and other data types). Consequently, a large dataset can be quickly retrieved from *PMR*'s database server, and analyzed and visualized by user-selected protocols. Fig. 1 shows the systems architecture of *PMR* including the hybrid storage system for data integration. The system is primarily composed of three components: 1) SQL-based storage system for metabolomics data and metadata; 2) a NoSQL-based storage system for metabolomics and transcriptomics data and metadata; 3) a portal for analysis and visualization that includes a service (*iMetaTrans, detailed in Section 4.2*) for co-analysis of transcriptomics data with metabolomics data.

Metabolomics data for the 18 species (34 experiments most using multiple analytical platforms) currently in *PMR* can be downloaded, and the associated unprocessed raw data is available upon request to an administrator. The integrated metabolomics and transcriptomics (RNA-seq) data, with additional metadata for transcript sequences, (predicted) protein amino acid sequences, GO annotations and more, are available for two of these species (*Echinacea purpurea* and *Hypericum perforatum*). Data that is deposited as private will be converted to public data after publication. In its present iteration, metabolomics data, related metadata, and (if available) associated transcriptomics data, can be submitted to *PMR* by researchers in the community; soon, researchers will be able to use *PMR* to submit, compare, and download transcriptomics and metabolomics data gathered from the same samples for all species.

3.2 Processing and statistical analysis of metabolomics data

Metabolomics data analysis, predominantly statistical, can be considered as encompassing three aspects, and in some cases, a fourth: 1) The raw metabolomics data is processed and analyzed. 2) Comparison/quality assessments are made across replicates. 3) Metabolite levels/types are compared among samples. 4) Metabolomics data can be integrated with a wide variety of other data types, such as transcriptomics, pathway and network annotations, or text mining data. Because different sources of metabolomics data often use differing scales, data normalization is an essential step for analysis of the data.[52, 53] Normalization also helps to reduce the noise of variability induced by sources other than biological variability.[54]

Often, raw metabolomics data is semi-manually processed; however manual analysis is time-consuming, impractical for large datasets, and unwieldy for non-specialists. Thus, the need for user-friendly software for raw data processing is immense. Recently, a web-based platform to automate processing and to some extent analysis of non-targeted LC/MS data, *XCMS online* (https://xcmsonline.scripps.edu/), has been developed.[37] The software includes MetAlign,[55, 56] Mzmine,[57] and XCMS,[58] all of which are widely accepted by the metabolomics community. *XCMS Online* allows researchers to upload data, compare several samples using Principal Component Analysis (PCA) or Multidimensional Scaling (MDS), and thus obtain an overview of the variances among samples or to identify potential outliers in replicates, and to visualize the results. A meaningful tool for analysis is the mirror plot for comparing metabolites between two samples and it represents up- or down-regulated compounds between two samples.[37]

MetabolomeExpress, which also processes raw data, contains a variety of data analysis methods for the GC/MS, GC/Quadrupole MS, and GC/TOF MS platform, including combined statistics with more than two samples.[39] One feature allows analysis of metabolite-metabolite correlations within GC/MS datasets and creation of its correlation network graphs using *Cytoscape* (http://cytoscape.org/).[59]

Chromatoplots[60] is a very different approach for analyzing raw metabolomics data. Designed for GC/MS data, it combines visualizations from several different Graphic User Interface (GUI) toolkits and the *XCMS package* with an LC/MS based data analysis approach in *R* (http://www.r-project.org/) to assist the user's understanding of the data.

It would be extremely informative to compare the results derived from various processing methods for raw data, using a variety of molecular (and synthetic) materials, including: experimental data from biological material; simulated data; data derived from biological material that has been “spiked” with known levels of standard metabolites; and data derived from a set of known levels of define metabolite standards. Although such types of comparisons are rare for ‘omics data, they are important both for processing data and for analysis of processed data, because they enable a better understanding of how the data and its distribution might affect the analysis and results.

Our focus here is on statistical approaches to analysis. Statistical considerations for comparisons of metabolite patterns across biological materials, and of metabolomics data to other data types are addressed in Section 4.1.

3.3 Metabolomics data visualization: comparison of two samples

Visualization of metabolomics data is distinct from its analysis, in that it deals with the way each type of analysis is displayed and how the user can interact with the data, but it is also used to extract information from the data. Data visualization is a well-developed area for multivariate statistical analysis for small data display. However, visualization of large ‘omics datasets requires more consideration.[61–63]

Visualizing complex, multidimensional datasets is extremely challenging, even when only two samples are being compared. Researchers often use ratio plots (Fig. 2) in order to visualize metabolite differences between two samples.[24, 25] In ratio plots, the x-axis shows the mean fold-change of the relative abundance of each metabolite between the two samples that the user selected. Metabolites that have a relatively low fold-change between these two samples are close to the central vertical y-axis; metabolites that have a higher fold-change are farther from the central vertical y-axis.

An alternate visualization is a volcano plot (Fig. 3).[63–66] Volcano plots are sometimes used for visualization of statistical results of omics data such as differential expression of genes measured through microarrays. The interactive volcano plot has the power to show at a click of a mouse button which metabolites show a stronger combination of fold change and statistical significance. They represent significance from a statistical test (such as a p-value) on the y-axis and fold-change on the x-axis. They can also compare metabolite levels with different experimental conditions. As a consequence, metabolites in the volcano plot that have a relatively low fold-change between the two samples appear near the center and metabolites that have significant p-values are found in the upper-right or upper-left.

The volcano plot provided by *PMR* (Fig. 3) illustrates a powerful implementation of interactive visualization and improved interpretation capabilities. To implement the interactive volcano plot, we employed HighCharts software (Highsoft Solutions AS, USA) written in JavaScript. It has high performance and easier to implement the volcano plot than software based on browser plugin technologies, such as Silverlight (Microsoft, USA)[67] and Adobe Flash (Adobe, USA)[68]. The plots can be interactively manipulated by researchers, and a variety of information such as p-values, fold-change, chemical formulae, and chemical classifications can be integrated.

3.4 Metabolomics data visualization: interactive display and comparison of multiple samples

When analyzing data from more than two biological samples, it is important to integrate data interactively with existing metadata and chemical information. Researchers require visualization tools for extracting specific information from their data analysis and improving the data interpretation. Some web-based platforms, such as *XCMS online*[37] and *AtMetExpress*,[30] mostly provide results as an image file. Others, like *MetabolomeExpress*,[39] provide a variety of interactive visualizations such as heatmaps, Chromatogram viewers for GC/MS data, and 3D PCA plots. It also gives image files of the results: 2D PCA score plots for the first two dimensionless Principal Components (PC), screen plots for the Eigen values of the covariance matrix, and loading plots for each vector. However, in order to easily and efficiently access massive, multiplatform metabolomics data with many types of metadata, simple and intuitive design with interactive visualization is important.

Recent advances in informatics using open source platforms have resulted in the development of many tools to implement interactive visualization. One of them, *D3js* (http://d3js.org), is a framework to create interactive plots with great functionality. This increases the visibility of data, and users ability to interoperate complex data. Hur *el at.* introduced a wide variety of approaches for visualizing complex and large datasets, in this case FT-ICR MS analysis of 20 petroleum samples and its correlation with chemical and physical properties[69, 70] with Circos[71] diagrams (http://circos.ca/). Because Circos diagrams support an image that allows linkable and clickable images,[72] they could be applied within a web-based application for the visualization of metabolomics and transcriptomics data. Hive plots (http://www.hiveplot.net/)[71, 73] have been used to provide a framework for the interactive visualization of a network relationship or the correlation among datasets, for example, Hive plots designed to show quantitative comparisons of metabolomics and transcriptomics data. Large-scale metabolomics studies could also benefit by visualization with Circos diagrams and Hive plots. Both tools could help a user recognize and identify significant biological features when comparing metabolomics data or transcriptomics data.

Implementing software that enables interactive visualization of integrated data provide a technical challenge. Several analysis platforms, including *PMR*, *XCMS online*, *SetupX*, and *MetabolomeExpress*, have developed tools that combine interactive visualization of metabolomics and transcriptomics data. However, these platforms are limited by the fact that they only allow the display of relatively small datasets (under several thousands data points). In order to display larger amounts of data, a plot program that allows interactive visualization could be combined with a browser plugin based technologies such as Silverlight (Microsoft, USA)[67] and Adobe Flash (Adobe, USA)[68].

3.5 Emergent technologies for data storage, retrieval, and analysis of metabolomics data

The technology used for storing and analyzing metabolomics data can be critical to the utility of the data. Until recently, conventional storage systems such as SQL-based databases and FTP file systems have been used for storage and web-based analysis platforms. Newer types of storage and informatics technologies like NoSQL, as is used in *PMR*, can enable more rapid and efficient storage, retrieval, and analysis of very large metabolomics and other ‘omics datasets. For example, Hadoop,[74] a document-oriented NoSQL system that has been integrated with the statistical analysis software R (http://www.r-project.com), can be used directly to statistically analyze metabolomics and transcriptomics data.[75, 76] Another example is Neoj4,[77] a graph and network based NoSQL system with a high-performance database, a flexible network structure, and a graph database. Discovering natural plant products and understanding of how these products are integrated into the overall network of metabolism and its regulation, analyze and visualize many relationships among different types of data and its metadata. Consequently, the variety of NoSQL-based approaches can facilitate data integration between metabolomics and other types of data, including high performance real-time data analysis. Thus, utilizing this type of technology will inform systems biology investigations.

4 Correlation as a measure of association between biomolecules

Systems-level data, and a systems-level approach to its analysis can enrich our understanding of all aspects of world around us.[78] There are a wide number of systems-level approaches for analysis of biological data. Here, we focus on the development and implementation of an interactive statistical approach to metabolomics and transcriptomic data.

4.1 Statistical co-analysis of different ‘omics data types: metabolomics and transcriptomics

Among the many considerations in the integrated statistical analysis of different data types (e.g., fluxomics, metabolomics, transcriptomic networks, proteomics) and analytical platforms (GC-MS, QTOF, etc.), we address three that pertain specifically to co-analysis of transcriptomics (measured either by microarrays or Next Generation Sequencing) and metabolomics. These are: a) different scales of measurement, for instance microarrays, proteomics, and metabolomics have continuous scales of measurements whereas transcripts measured through Next Generation Sequencing come in total count of short reads, i.e. a discrete scale; b) arbitrary scales of measurement that require normalization procedures in order to make measurements comparable; and c) data heteroskedasticity (a term used in statistics to describe variances that are not constant across measurements) that is present across omics data and includes induced technical and biological variability components that differ across different platforms.

Normalization procedures can be used to convert data that are collected on arbitrary scales, enabling data types to be compared to each other. Typical biological examples include quantile and scale normalization for microarray data,[79] and RPKM and FPKM for RNA-seq data,[80] both of which make measurements that are independent of total counts in the samples analyzed. Because FPKM and RPKM normalization overcome the problem of mixing discrete and continuous scales, they enable an integrated analysis across metabolomics and transcriptomics without the requirement of developing specific methods for mixing continuous/discrete scales. Unlike transcriptomics data, metabolomics data lacks a “gold” standard for data normalization. Van den Berg *et al*.[81] analyzed 10 types of normalization procedures for metabolomics data, and reported that the conclusions of which metabolites were most abundant/important depended on the normalization method used.

Understanding data variability is a fundamental step that can guide the choice of the right statistical model for data analysis and inference. The appropriate statistical model will reduce variance estimates and thus provide more accurate inferences. Heteroskedasticity of microarray data has been extensively studied.[82–85] For instance, variances in microarray data are proportional to gene expression levels for transcripts that are near average in their level of accumulation. Because of this, a log transformation stabilizes variances of microarray data for such mid-range values. Technical variance of raw counts of short sequences in RNA-seq data have a linear relationship with transcript expression, whereas biological variance displays a larger, non-linear, and increasing relationship with the measured level of transcript accumulation.[86–88]

Heteroskedasticity of many metabolomics measurements shows a similar pattern to the one observed for microarray data, in which there seems to be a proportional relationship between variances and measured abundances of metabolites (the larger the abundance, the larger the variance). But this is not the case for metabolite abundances close to the detection limits. Our experience with the data in the *MPMR* database[32] was that log transformations stabilized variances of abundances that were not close to the detection limit. However, metabolomic data contrasts to transcriptomic data in that for metabolomic, the variability among biological replicates is more erratic. This can be seen in the scatterplots of Fig. 4, where metabolites detected in only one of a pair of biological replicates are represented as circles parallel to the x- or the y-axis. The Spearman correlation coefficients between pairwise combinations of biological replicates in scatterplot matrix of Fig. 4 also show that biological (and/or extraction) variation of metabolites present at levels above the detection limit is considerably larger than the variability that would be expected from microarray or RNA-seq biological variation (e.g., Marioni *et al.*)[89]. The Venn diagram (Fig. 4) illustrates this point numerically: a total of 449 metabolites were above detection limit across 3 biological replicates of *Atropa belladonna* flower-buds, but only 222 of these metabolites were above detection limit in all three biological replicates.

The integrated analysis of metabolomics and transcriptomics involves the following three types of interrelated associations: metabolite vs. metabolite, transcript vs. transcript, and metabolite vs. transcript. Since its introduction as a tool to measure co-expression in microarray studies by Spelman *et al*, Pearson Correlation Coefficients have been a popular method for estimating associations between transcript expression vectors.[90] Pearson correlation coefficients owes popularity to its ability to detect similarity between genomic elements that have different absolute values of expression levels. Because it is a parametric measure (it assumes that the data is normally distributed), it provides more accurate results if the parameters are met. Alternative non-parametric approaches such as Jackknife[91] and Spearman correlation[92] have also been used in order to overcome the issue of outliers and to avoid relying on strong assumptions about the statistical distribution of the data.

Pearson and Spearman correlation coefficients are also common tools used for exploring metabolite-metabolite associations.[93] Spearman rank correlation is preferred by some in analyzing metabolomics data because the very high dynamic range of metabolomics data has a greater tendency to produce outliers.[94] Steuer *et al*. hypothesized that the experimentally observable patterns of correlation between metabolites result from biological fluctuations of metabolic levels caused by the changing environment and the regulatory steps of the metabolic network.[95] The fluctuations propagate and ultimately create the observable patterns of correlation. This is not always the case, for example, in samples of Arabidopsis shoots harvested diurnally, the topology of the metabolite correlation network was different from the KEGG[96] biochemical pathways, indicating correlations among metabolites do not always reflect a shared functionality.[21, 97] Metabolomics data from yeast exposed to varied conditions have shown high correlation of metabolites to genes that are part of related biological processes, although the nature and extent of the correlations depended on the metabolites involved and the physiological conditions.[98]

Interpretation of correlations, or the lack thereof, between metabolites and their function is confounded because of incompleteness (or incorrectness) of data. For example, typically both the metabolic and regulatory pathway data and the metabolomics data are incomplete. Furthermore, studies comparing ‘omics data types involve biological material that is compartmentalized (i.e., the data is derived from a mixture of different cell types, different organellar compartments, and/or cells at different conditions of development).

Even though the specific mechanisms that impact the correlations among accumulation of metabolites and transcripts are not completely understood under any environmental, developmental, or genetic perturbation, a correlation analyses can be very useful in identifying novel metabolite-metabolite and metabolite-transcript interactions. As detailed in the introduction, parallel analysis of transcriptome and metabolome has successfully identified small molecules and genetic elements involved in their metabolism.[4–21]

We give here a method for integrated ‘omics analysis using correlation coefficients as a measure of similarity between transcript-metabolite profiles across samples under various biological conditions (its implementation to create interactive software is described in Section 4.2). A correlation coefficient equal to 1 would describe the ideal situation of perfect similarity between metabolite/transcript profiles. Therefore, it can be argued that a metabolite-transcript pair showing similar profiles would be affected in similar ways by the biological processes taking place across various conditions. The task of finding which transcripts have similar profiles to the one observed for a metabolite of interest is illustrated in Fig. 5.

Briefly, given the profile of a chosen metabolite (it could also be a transcript) the algorithm estimates Pearson correlation coefficients between profiles of the chosen metabolite and all transcripts present in the database. We used the Pearson correlation coefficient between log-transformed metabolite abundance and transcript expressions (normalized as RPKM) because an exploratory analysis of the data revealed that a log transformation stabilized variances and caused the data to be fairly normally distributed. We obtained statistical significance of correlation coefficients by testing the hypothesis Ho: ρ=0 through the statistics $t\,(r) = r\,/\,\delta_r$, where $\delta_r = \sqrt{1 - r^2/n - 1}$ corresponds to the standard deviation of the Pearson correlation coefficient *r*. Under the specific conditions of the null hypothesis Ho: ρ=0, only one of the variables in the correlated pair is required to be normally distributed so that *t(r)* follows a Student's t distribution with *(n−2)* d.f..[99] The task of finding transcripts and metabolomics with similar profiles involves tens of thousands of statistical tests; the multiple tests problem has a high impact on the statistical significance of the analysis. Therefore, we correct for multiple tests and report pFDR-values for each of the pairwise comparisons. For example, a p-value cut-off of 0.05 allows 5% false positives, and therefore a set of 1000 simultaneous tests will contain on average 50 false positives. This is where calculation of False Discovery Rate (FDR) becomes relevant. The FDR is the expected fraction of significant tests that are false positives. If we further consider that 200 of the 1000 tests in the example above are identified as statistically significant results, a tolerance of 5% FDR will result in only 10 false discoveries. Several multiple testing correction methods have been proposed in the literature, and the FDR correction proposed by Benjamin and Hochberg[100] is very popular. However, the FDR method has been found to be too conservative for genome wide studies.[101] Hishikawa *et al.* introduced a quantity called pFDR which is defined as $pFDR = E[F/S|S>0]$ where *F* is the number of False Positives and *S* is the expected number of true Positives. Storey[101, 102] defined a new quantity called q-value, which is the minimum *pFDR* at which the given feature can be called significant.

In order to check if our parametric approach suited the integrated analysis of metabolomics and transcriptomics data, we performed a computer experiment where non-parametric null distributions were generated for 35,992 randomly selected metabolite/transcripts correlation coefficients, by randomly re-labelling the samples 5000 times and estimating the Pearson correlation coefficient between randomly scrambled metabolite and transcript measurements. Statistical significances of correlation coefficients obtained through the parametric and non-parametric methods were compatible in 95% of cases, which corroborates the validity of a parametric approach for the combined analysis of metabolomics and transcriptomics data. Furthermore, the parametric approach overcomes extensive computational time and storage requirements of the re-labelling method. Similar comparisons between parametric and non-parametric approaches have been done for transcript-transcript[103] and metabolite-metabolite comparisons,[94] both resulting in the same conclusions as our experiment.

A bootstrap-based method has been used to create a null distribution for the correlation data of metabolite to microarray data.[104] However, to the best of our knowledge, a comparison of parametric and non-parametric approaches had not been previously used for metabolite-RNA-seq data.

4.2 Implementation of a co-analysis database and software for ‘omics data

Several challenges lie in implementing software to visualize and co-analyze metabolomics data with other ‘omics data. For example, consider co-analysis of metabolomics data with transcriptomics data. First, a well-structured schema for the database is essential. Metabolomics and transcriptomics data and their metadata have complex interdependencies, which causes difficulty in data integration. One design solution is to utilize a schema-free database system such as NoSQL, allowing for flexibility and more rapid redesign.

The second important issue is the ever-increasing need for computing speed to enable real-time analysis of large and larger datasets. NoSQL methods provide solutions that address this issue as well. As previously mentioned, in order to accelerate the co-analysis time between metabolomics and transcriptomics data, *PMR* uses a document-oriented NoSQL system, which has advantages in flexibility, scalability and performance, to reduce the time to retrieve data from storage.[50] The *iMetaTrans service* uses an internally-developed application server to quickly calculate correlation coefficient values, p-values (using the alglib library), and pFDR values.[102]

Fig. 5 shows a flowchart of co-analysis by *PMR*. First, the appropriate transcriptomic samples are matched to the metabolomics data. Co-analysis is then performed and p-values are calculated; if the data includes large numbers of metabolites/transcripts, pFDR values are computed. The co-analysis is then loaded into *PMR*. From the resulting co-analysis table (an example in shown in Fig. 6), researchers can view the sequences, sort the table, and interact with detailed information by clicking the sequence ID.

These features integrate sequence and gene ontology information, permit real-time statistical analysis and visualization, and thus allow interactive comparisons between large amounts of metabolomics data and associated transcriptomics data. Using these approaches, analysis of high dimensional data (e.g., 500 metabolites correlated pairwise with 20,000 genes, see Section 3.1) can be addressed computationally. The asymmetry of the *completeness* (or lack thereof) of metabolomics datasets, due to technical limitations of the analytical technology remains a constraint on the ability of co-analysis of metabolomics data with transcriptomics data.

5 Case study: Utilization of a metabolomics database to understand biochemical function and gene redundancy - Insights into the FAE Elongation (ELO) genes of Arabidopsis

Advances in whole genome sequencing have revealed the extent to which gene paralogs exist within a genome.[105–108] Deducing the biochemical function of each gene paralog and obtaining insight into why this redundancy is retained within the genome becomes an interesting yet complex problem. This case study illustrates how the *PMR* database can be used to integrate metabolic profiling data with reverse genetics to provide insight and testable hypothesises about gene redundancy within the Arabidopsis fatty acid elongase (FAE) system. FAE is a complex system of integral membrane proteins that elongates pre-existing saturated and unsaturated fatty acids of 16 and 18 carbon chain lengths to fatty acids of greater than 18 carbons in length (VLCFA). This system is composed of four enzyme components that act in an iterative cycle of Claisen condensation, reduction, dehydration and a second reduction. Ketoacyl-CoA synthases (KCS) have been biochemically shown to catalyse the Claisen condensation reaction. There are 21 KCS-coding genes in the Arabidopsis genome.[109, 110] In mammalian and yeast systems, the enzymes that catalyses these Claisen condensation reactions are termed ELO.[28, 111–115] The KCS enzymes share no sequence similarity to the ELO enzymes, thus two apparently unrelated proteins occur able to catalyse this Claisen condensation reaction. An unanticipated finding that genome sequences have revealed is that plants contain ELO-like and KCS genes. Four such ELOs are in Arabidopsis: AT3G06460 (*AtELO1*), AT4G36830 (*AtELO2*, *HOS3*), AT3G06470 (*AtELO3*), and AT1G75000 (*AtELO4*)[116, 117] (and Cahoon unpublished work). These two types of condensation enzymes are both located in the endoplasmic reticulum, which indicates that Arabidopsis ELO enzymes may provide a plant-specific FAE Claisen condensation reaction functionality.

The role of the ELOs in plant VLCFA biosynthesis has yet to be confirmed, although it has been hypothesized that the fatty acid products synthesized by ELO-containing FAE systems are utilized in sphingolipid and phospholipid biosynthesis.[116, 117]

An Arabidopsis knock-down series was generated from three of the Arabidopsis ELO (*Atelo*) knock-out lines, as part of the mutant lines evaluated by the *Plant Metabolomics* 2010 Consortium.[25] These lines, generated by Dietrich and Cahoon (unpublished), exploit a reverse genetic approach and use publically available Arabidopsis single knockout ELO T-DNA lines (SALK_075185, *Atelo2*; SALK_109405, *Atelo3*; and SALK_083029, *Atelo4*). The detailed metabolite profile produced for this knock-down series, includes data[24] from the three single knock-out lines; three double knock-out lines *Atelo2*: :*Atelo3*; *Atelo2*: :*Atelo4*; *Atelo3*: :*Atelo4*; and the triple knock-out line *Atelo2*: :*Atelo3*: :*Atelo4*. Integrating these metabolomics datasets with the statistical power and tools on *PMR* is enabling a deeper understanding of the roles of enzyme and gene redundancy in the FAE metabolic network.

In this study, each individual *Atelo* knock-out line, the double knock-out lines, the triple knock-out line, and the wild-type Columbia control were compared against each other interactively, and the data visualized by volcano plot. *PMR* enables *t*-test options (estimate of variance and one/two tailed analysis); for this analysis the default settings were used (Estimate of Variance using Auto-select and Two tails), a p-value cut-off of 0.05, and detection limit values were ignored for non-detected metabolites. Using the knowledge that ELO proteins in non-plant systems function as the condensing enzymes of FAE to inform our computational research, we postulated that metabolites in fatty acid and its derivatives would be the most affected in the mutant lines. Examining the entire set of nearly 1500 identified and unidentified metabolites that are profiled by the six platforms in the plant metabolomics analysis of the ELO mutant lines, we determined that FAE-related compounds were the main class of metabolites whose level was affected significantly across the mutants lines. The investigation was thus focused on the analytical platforms that specifically detect the FAE-related compounds. These platforms are: fatty acids, cuticular waxes (data generated within Dr. Basil Nikolau's research group), and lipidomics (data generated within Dr. Ruth Welti's research group, Kansas State University).[25]

The volcano plot for all pairwise comparisons of the *Atelo* knock-out lines indicates that fatty acid accumulation for each of the three single gene knock-out lines are generally unaffected; however, knocking down ELO activity further via double and triple knock-outs decreases fatty acid accumulation significantly (Fig. 7). Presupposing that ELO activity is decreased as more *elo* mutant alleles are combined, these data indicate that as ELO activity is knocked-down, the impact on the plant's metabolome, specifically fatty acid metabolism, becomes more acute.

Volcano plots of all pairwise combinations of all reverse genetic lines and all analytical platform data (shown in Fig. 8 for the pairwise comparison of WT versus double knock-out *Atelo2*: :*Atelo4*) revealed that of the approximately 1,500 metabolites detected across the platforms, *virtually all of the metabolites that are differentially accumulated are fatty-acid-derived metabolites*. Volcano plots of all pairwise combinations of all reverse genetic lines and all analytical platform data also revealed the knock-out line with the most severe chemical phenotype (chemotype) is the double knock-out *Atelo2*: :*Atelo3*. This indicates that when the third mutation (*Atelo4*) is introduced some type of recovery mechanism is induced which lessens the numbers and types of metabolites affected.

A comparison of each Arabidopsis line to the wild type control reveals a unique chemical profile with defining metabolite characteristics. For example, analysis of the cuticular VLCFA derivatives reveals an interesting trend. These metabolites were generally increased in the ELO knock-out lines, indicating that FAE KCS may be upregulated and 'over-producing' cuticular wax components as a consequence of the decreased ELO-type FAE activity (Fig. 10).

In addition, VLCFA containing phospholipids and in some cases glycoglycerolipids are significantly affected across the knock-out lines (figure not shown). Each *Atelo* knock-out causes a distinct fingerprint of change in the accumulation of particular VLCFA-containing phospholipids. For example, *Atelo2* knock-out causes an increase in 36:1 species of phosphatidylethanolamine (PE), phosphatidylglycerol (PG), and phosphatidylserine (PS) while decreasing PS 40:1 and PS 42:1. In contrast, the major effect within the *Atelo3* knock-out line is decreased accumulation of phosphatidylinositol (PI) and PE metabolites and an increase in C40 species of PC. Although the fatty acid profiles in the different ELO knock-out lines have major differences, many VLCFA derivatives are unaffected by decreased ELO gene expression.

One possible explanation is that as ELO-type FAE activity is knocked-down, the remaining ELO, (i.e., AtELO1) and/or KCS-type FAE activity compensates for the decrease in ELO functionality, but collectively these changes cannot fully recover the diverse fatty acid profile expressed in the wild-type. An alternate explanation is that ELO proteins have a unique role in VLCFA synthesis. A cursory examination of transcript expression profiles for the ELO-like and KCS genes across multiple biological conditions visualized on the MetaOmGraph database (http://www.metnetdb.org/MetNet_MetaOmGraph.htm)[118–122], shows that KCS transcripts appear generally more abundant than ELO-like transcripts (Fig. 8); RNA-seq data for a more limited number of biological conditions bears out this observation quantitatively (Li and Wurtele, unpublished). Further, the accumulation of the three ELO-like transcripts reveals very diverse patterns of accumulation for each gene across multiple conditions. For example, *ELO1* is more abundant in roots, whereas *ELO2* is relatively more abundant in hypocotyl and cotyledons. *ELO3* accumulates to a higher level and is abundant under most other conditions. It will be interesting to experimentally assay transcript abundance of the ELO encoding gene(s) and the KCS encoding genes across the knock-down series to better understand how the remaining VLCFAs are synthesized. This research will be facilitated because the collections of biological material from the mutant seedling of the *Plant Metabolomics* Consortium also includes biological material for analysis of transcriptomics; thus, RNA-seq can be analyzed for the same samples used for metabolomics analysis.

Collectively, these metabolomics analyses were greatly facilitated by having the metabolomics data in a metabolomics database with software that enables highly interactive analyses. The analysis indicate that ELO paralogs affect VLCFA accumulation, each providing different substrate specificities, and that they likely function similarly to their yeast and mammalian orthologs as condensing enzymes for FAE. Further, the ELO-dependent FAE products appear to contribute to the pool of phospholipids and neutral glycoglycerolipids, but do not appear to contribute to cuticular wax accumulation. Other VLCFA-derived molecules, such as ceramides, wax esters (not targeted in the analytical platforms of *Plant Metabolomics*) would be strong candidates for further investigation into the additional roles of the ELO proteins.

6 Conclusion

Considerable resources in effort and monetary expenditures are used in obtaining natural product and other metabolomics data. Deposition of this data into a flexible database with a user-friendly interface enables members of the research community to leverage this data on a global scale for a variety of purposes. Such databases enable researchers to intuitively and quickly assess data quality, compare metabolite levels across samples, annotations, evaluate redundancies among different platforms in metabolite identification, or compare metabolomics to other data types.

Biological systems, even at the cellular level, are complex and dynamic; existing models capture only a small proportion of behavior. However, despite these limitations, current approaches to systems-level studies are enabling extensive biological discovery. Bringing metabolomics more intimately into the mix will further these advances.

By aggressively incorporating interactive visualization capabilities, the statistical analysis of metabolomics data and its associated complex chemical information can be more easily understood and placed in context. Co-analysis of metabolomics data with transcriptomics data enables users to develop hypotheses that associate metabolites with genes that maybe involved in their synthesis or regulation. The problems associated with the asymmetry between the size of metabolomics and transcriptomics datasets can be prevailed with appropriate statistical analysis. *PMR* provides an example of a metabolomics database with data-dependent statistical analysis, interactive links among metadata, the genome and metabolome and enables researchers to extract novel information concerning genes and markers of natural compound biosynthesis.

Several key computational, biological and technical challenges have been discussed. Analysis of high dimensional data can be addressed computationally for example using NoSQL-type distributed approaches. However, interactive real-time computations of relationships such as pFDR, pairwise value comparisons for large datasets, and various network analyses with integrated data require significant computing resources, thus computing speed needs to be further optimized. Models for co-analysis of combined data types and for exploring data in the context of metabolic and regulatory networks are conceptually challenging. To develop models that can be used to strengthen biological understanding requires the combined effort for domain experts (biologists/biochemists/chemists) and computer scientists. Interactive visualization using graphs, tables, Circos diagrams, and Hive plots provide powerful ways to understand large-scale biological data. These capabilities must be adapted to the data types, and much additional research is needed to better visually represent the multidimensional nature of the data such that a researcher can grasp salient features.

Other factors are due to current technological limitations. One of the greatest challenges, which is due to limitations in analytical technologies, is that metabolomics analyses can identify and measure only a small proportion of the metabolites in a biological material. For example, in the case of natural products, many precursors are present in only very low concentrations. Another example is that of unstable metabolites, which are often degraded before their detection. Even more pervasive is current inabilities to identify detected compounds. Thus, we are left with an incomplete picture of the metabolome. Another major issue is that current methodologies of ‘omics analyses cannot decipher data on a compartmental basis.

A user-friendly database with associated analysis capabilities is critical to the accessibility and use of metabolomics data. As an expanding public resource, this capability can lead to discovery and development of testable hypotheses. In the future, new innovations will further extend these capabilities and facilitate the road to predictive biology.

Acknowledgments

We deeply appreciate the helpful comments of all members of the Medicinal Plants Consortium, especially Joe Chappell, Dean Dela Pena, Robin Buel, Dan Jones, Sarah O'Conner, and Natalia Dudareva, on the design of the user interface for *Medicinal Plant Metabolomics Resource*, on which we have modeled much of the *PMR* interface. We are grateful to Preeti Bais and Stephanie Quanbeck for their kind help with provision of the Arabidopsis data from *Plant Metabolomics*. We thank the members of the MetNet team and the Wurtele lab, most especially Micheline Ngaki, for much valuable feedback during database construction. This material is based in part upon work supported by the National Science Foundation under award numbers: EEC-0813570 (E.S.W. and B.J.N); MCB-0951170 (to E.S.W. and L.L); Medicinal Plant Consortium GM092521, National Institutes of Health; and the Center for Metabolic Biology, Iowa State University (to E.S.W). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation of the National Institutes of Health.

Biographies

Manhoi Hur completed his B.S in Computer Science and Engineering from Paichai University, Korea, in 2002. He is pursuing his M.S in Bioinformatics at Korea University, Korea. Currently, he is working with Dr. Wurtele as a research scientist in developing PMR site using NoSQL based approaches.

He is interested in developing data analysis software for -omics study such as metabolomics, transcriptomics and Petroleomics using FT-ICR MS and developing big-data acquisition system of mass spectrometry.

Alexis A Campbell earned her Ph.D. in Plant Biology in 2011 for her work in genetic redundancy within fatty acid elongase systems in biochemistry, at Iowa State University. To continue the growth and expansion of her current research interests, she was awarded both the HHMI Educational Development and a NSF-AGEP Post-Doctoral Fellowships, which has allowed continued work within the field of fatty acid metabolism. This includes the investigation of how plant and yeast systems regulate and define the set of fatty acid products they produce for the purposes of biorenewable chemicals.

Dr. Li, Adjunct Assistant Professor at Iowa State University, completed her undergraduate studies in biology at Peking University, Beijing, P.R.China and her Ph.D. (major genetics, minor statistics) in Iowa State University in 2006. Dr. Li stayed at Iowa State University after her graduation. Dr. Li has been developing an integrated approach to identify the factors that regulate plant. She combines bioinformatics, structural studies and molecular genetics to explore and expand the starch metabolic and regulatory network, to identify previously uncharacterized genes (protein with obscure features).

Adarsh did his bachelors in Electrical and Biomedical Engineering from Cochin University of Science and Technology, India and his master's degree in Biomedical Engineering from University of Akron, Ohio. He is pursuing his PhD in Bioinformatics and Computational Biology at Iowa State University working with Dr.Basil Nikolau in developing methods to integrate different types of biological data to generate testable biological hypotheses.

Nick Ransom received his BS in Computer Science from Iowa State University. He is an expert in Java and JavaScript. He has created bioinformatics software for massive datasets, most notably MetaOmGraph, and has designed the database, Medicinal Plant Metabolomics Resource (MPMR). He is developing Meta!Blast, a computer game for cell and metabolic biology.

Eve Wurtele, Professor at Iowa State University, received a B.S. from UC-Santa Cruz, and a Ph.D. in Biology from UCLA in 1980. After a postdoctoral fellowship at UC-Davis and Senior Scientist at NPI, a biotechnology company, she joined Iowa State. Wurtele's research, juxtaposed at the interface between biological and computational sciences, centers on the interplay between metabolic and regulatory signals. The research is revealing a complex network that mediates accumulation of proteins, starches, oils, and specialized natural products. Wurtele directs award-winning computer game, Meta!Blast, which takes a player into an interactive metabolic adventure within a 3D photosynthetic cell.

References

1. Bino RJ, et al. Trends Plant Sci. 2004; 9:418–425. [PubMed: 15337491]
2. Oliver DJ, Nikolau B, Wurtele ES. Metab. Eng. 2002; 4:98–106. [PubMed: 11800579]
3. Weckwerth W, Fiehn O. Curr. Opin. Biotechnol. 2002; 13:156–160. [PubMed: 11950569]
4. Ounaroon A, et al. Plant J. 2003; 36:808–819. [PubMed: 14675446]

5. Gennity JM, Stumpf PK. Arch. Biochem. Biophys. 1985; 239:444–454. [PubMed: 4004273]

6. Bafor M, et al. Biochem. J. 1991; 280:507–514. [PubMed: 1747126]

7. Lee M, et al. Science. 1998; 280:915–918. [PubMed: 9572738]

8. Bafor M, et al. Biochem. Biophys. 1993; 303:145–151.

9. Cahoon EB, et al. J. Biol. Chem. 2001; 276:2637–2643. [PubMed: 11067856]

10. Dyer JM, et al. Plant Physiol. 2002; 130:2027–2038. [PubMed: 12481086]

11. Sperling P, et al. Eur. J. Biochem. 2000; 267:3801–3811. [PubMed: 10848999]

12. Cakir T, et al. Mol. Syst. Biol. 2006; 2:50. [PubMed: 17016516]

13. Murray DB, Beckmann M, Kitano H. Proc. Natl. Acad. Sci. USA. 2007; 104:2241–2246. [PubMed: 17284613]

14. Kresnowati MT, et al. Mol. Syst. Biol. 2006; 2:49. [PubMed: 16969341]

15. Smith MA, et al. Biochem. J. 1992; 287:141–144. [PubMed: 1417766]

16. Amiour N, et al. J. Exp. Bot. 2012; 63:5017–5033. [PubMed: 22936829]

17. Dang TT, Facchini PJ. Plant Physiol. 2012; 159:618–631. [PubMed: 22535422]

18. Jauhiainen A, et al. Biostatistics. 2012; 13:748–761. [PubMed: 22699861]

19. Osorio S, et al. Plant Physiol. 2012; 159:1713–1729. [PubMed: 22685169]

20. Singh SP, et al. Curr. Opin. Plant Biol. 2005; 8:197–203. [PubMed: 15753001]

21. Gibon Y, et al. Genome Biol. 2006; 7:R76. [PubMed: 16916443]

22. Broun P, et al. Science. 1998; 282:1315–1317. [PubMed: 9812895]

23. Jenkins H, et al. Nat. Biotechnol. 2004; 22:1601–1606. [PubMed: 15583675]

24. Bais P, et al. Plant Physiol. 2010; 152:1807–1816. [PubMed: 20147492]

25. Quanbeck SM, et al. Front. Plant Sci. 2012; 3:15. [PubMed: 22645570]

26. Raetz CR, et al. J. Lipid Res. 2009; 50:S103–S108. [PubMed: 18974037]

27. James DWJ, et al. Plant Cell. 1995; 7:309–319. [PubMed: 7734965]

28. Oh CS, et al. J. Biol. Chem. 1997; 272:17376–17384. [PubMed: 9211877]

29. Ngaki MN, et al. Nature. 2012; 485:530–533. [PubMed: 22622584]

30. Matsuda F, et al. Plant Physiol. 2010; 152:566–578. [PubMed: 20023150]

31. Goda H, et al. Plant J. 2008; 55:526–542. [PubMed: 18419781]

32. Wurtele ES, et al. Metabolites. 2012 In Press.

33. Crispin, MC.; Wurtele, ES. Biotechnology for Medicinal Plants. Chandra, S.; Lata, H.; Varma, A., editors. Vol. ch. 17. Springer Berlin Heidelberg; 2013. p. 395-411.

34. Giddings LA, et al. J. Biol. Chem. 2011; 286:16751–16757. [PubMed: 21454651]

35. Yeo YS, et al. J. Biol. Chem. 2012 In Review.

36. Takahashi S, et al. Biotechnol. Bioeng. 2007; 97:170–181. [PubMed: 17013941]

37. Tautenhahn R. Anal. Chem. 2012; 84:5035–5039. [PubMed: 22533540]

38. Lewis N, et al. MedPlTranscriptome. http://uic.edu/pharmacy/MedPlTranscriptome/database.html.

39. Carroll AJ, Badger MR, Harvey Millar A. BMC Bioinformatics. 2010; 11:376. [PubMed: 20626915]

40. Scholz M, Fiehn O. Pacific Symp. Biocomp. 2007; 12:169–180.

41. Skogerson K, et al. BMC Bioinformatics. 2011; 12:321. [PubMed: 21816034]

42. Clark B. NoSQL: If only it was that easy. http://bjclark.me/2009/08/04/nosql-if-only-it-was-that-easy/.

43. Lith, AJM. Master Thesis. 2010.

44. Sadalage P, Fowler M. NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence. 2012

45. Kumar S. What is a Key/Value store database? http://dba.stackexchange.com/questions/607/what-is-a-key-valuestore-database.

46. bsonspec.org, BSON, http://bsonspec.org/.

47. json.org, Introducing JSON, http://www.json.org/.

48. Renzo Angles CG. Survey of graph database models. http://dl.acm.org/citation.cfm?id=1322433.

49. Chodorow, K.; Horowitz, E.; Merriman, D. mongoDB website. http://www.mongodb.org/.

50. Kennedy, M. MongoDB vs. SQL Server 2008 Performance Showdown. http://blog.michaelckennedy.net/2010/04/29/mongodbvs-sql-server-2008-performance-showdown/.

51. Kennedy, M. The NoSQL Movement, LINQ, and MongoDB – Oh My!. http://blog.michaelckennedy.net/2010/04/22/the-nosqlmovement-linq-and-mongodb-oh-my/.

52. Parmigiani, G., et al. The Analysis of Gene Expression Data. New York: Springer; 2003.

53. Bolstad BM, et al. Bioinformatics. 2003; 19:185–193. [PubMed: 12538238]

54. Veselkov KA, et al. Anal. Chem. 2011; 83:5864–5872. [PubMed: 21526840]

55. Lommen A. Anal. Chem. 2009; 81:3079–3086. [PubMed: 19301908]

56. Lommen A, Kools HJ. Metabolomics. 2012; 8:719–726. [PubMed: 22833710]

57. Katajamaa M, Miettinen J, Oresic M. Bioinformatics. 2006; 22:634–636. [PubMed: 16403790]

58. Smith CA, et al. Anal. Chem. 2006; 78:779–787. [PubMed: 16448051]

59. Shannon P, et al. Genome Res. 2003; 13:2498–2504. [PubMed: 14597658]

60. Lawrence M, et al. Chromatoplots. https://github.com/tengfei/chromatoplots/.

61. Healey CG, Enns JT. IEEE Trans. Vis. Comput. Graph. 1999; 5:145–167.

62. Vickers P, Faith J, Rossiter N. IEEE Trans. Vis. Comput. Graph. 2012 In Press.

63. Hertogh BD, et al. BMC Bioinformatics. 2010; 11:17. [PubMed: 20064233]

64. Perkins EJ, et al. BMC Bioinformatics. 2006; 7:S22. [PubMed: 17217515]

65. Brusniak M-Y, et al. BMC Bioinformatics. 2008; 9:542. [PubMed: 19087345]

66. Cui X, Churchill GA. Genome Biol. 2003; 4:210. [PubMed: 12702200]

67. Microsoft, Silverlight. http://silverlight.codeplex.com/releases/view/78435/.

68. Adobe, Adobe Flash. http://get.adobe.com/flashplayer/.

69. Marshall AG, Rodgers RP. Proc. Natl. Acad. Sci. USA. 2008; 105:18090–18095. [PubMed: 18836082]

70. Hur M, et al. Energy Fuels. 2010; 24:5524–5532.

71. Krzywinski M, et al. Genome Res. 2009; 19:1639–1645. [PubMed: 19541911]

72. Krzywinski M, et al. IMAGE MAPS - INTRODUCTION AND CLICKABLE IDEOGRAMS. http://circos.ca/documentation/tutorials/image_maps/.

73. Krzywinski M, et al. Brief. Bioinform. 2012; 13:627–644. [PubMed: 22155641]

74. Group AH. Hadoop project. http://hadoop.apache.org/.

75. Smith D. R and Hadoop: Step-by-step tutorials. http://blog.revolutionanalytics.com/2012/03/r-and-hadoop-step-by-step-tutorials.html.

76. Smith D. Data distillation with Hadoop and R. 2012 http://blog.revolutionanalytics.com/2012/06/data-distillation-with-hadoop-and-r.html.

77. Neo4J, Neo4J project. http://neo4j.org/.

78. Churchman, CW. The systems approach. New York: Delacorte Press; 1968.

79. Irizarry RA, et al. Biostatistics. 2003; 4:249–264. [PubMed: 12925520]

80. Mortazavi A, et al. Nat. Methods. 2008; 5:621–628. [PubMed: 18516045]

81. van den Berg RA, et al. BMC genomics. 2006; 8:142. [PubMed: 16762068]

82. Ji T, Liu P, Nettleton D. Stat. Appl. Genet. Mol. Biol. 2012; vol. 11:12.

83. Wang K, et al. IEEE Trans. Inf. Technol. Biomed. 2009; 13:848–853. [PubMed: 19527960]

84. Durbin B, Rocke DM. Bioinformatics. 2004; vol. 20:660–667. [PubMed: 15033873]

85. Rocke DM, Durbin B. J. Comput. Biol. 2001; vol. 8:557–569. [PubMed: 11747612]

86. Robinson MD, Smyth GK. Bioinformatics. 2007; 23:2881–2887. [PubMed: 17881408]

87. Robinson MD, Smyth GK. Biostatistics. 2008; 9:321–332. [PubMed: 17728317]

88. McIntyre LM, et al. BMC genomics. 2011; 12:293. [PubMed: 21645359]

89. Marioni JC, et al. Genome Res. 2008; 18:1509–1517. [PubMed: 18550803]

90. Dehmer M, et al. Applied statistics for network biology:methods in systems biology. 2011

91. Heyer LJ, Kruglyak S, Yooseph S. Genome Res. 1999; 9:1106–1115. [PubMed: 10568750]

92. Jiang D, Tang C, Zhang A. IEEE Trans. Knowl. Data Eng. 2004; 16:1370–1386.

93. Dastani Z, et al. PLoS Genet. 2012; 8:e1002607. [PubMed: 22479202]

94. Fukushima A, et al. BMC Syst. Biol. 2011; 5:1. [PubMed: 21194489]

95. Steuer R, et al. Biochem. Soc. Trans. 2003; 31:1476–1478. [PubMed: 14641093]

96. Kanehisa M, et al. Nucleic Acids Res. 2010; 38:D355–D360. [PubMed: 19880382]

97. Carrari F, et al. Plant Physiol. 2006; 142:1380–1396. [PubMed: 17071647]

98. Bradley PH, et al. PLoS Comput. Biol. 2009; 5:e1000270. [PubMed: 19180179]

99. Snedecor, GW.; Cochran, WG. Statistical methods. Ames, IA: Iowa State University Press; 1989.

100. Benjamini Y, Hochberg Y. JR. Stat. Soc. 1995; 57:289–300.

101. Storey JD, Tibshirani R. Proc. Natl. Acad. Sci. USA. 2003; 100:9440–9445. [PubMed: 12883005]

102. Storey JD. JR. Stat. Soc. 2002; 64:479–498.

103. Lee HK, et al. Genome Res. 2004; 14:1085–1094. [PubMed: 15173114]

104. Allen E, et al. BMC Syst. Biol. 2010; 4:62. [PubMed: 20465807]

105. Kanehisa M, et al. Nucleic Acids Res. 2006; 34:D354–D357. [PubMed: 16381885]

106. Koonin EV. Annu. Rev. Genet. 2005; 39:309–338. [PubMed: 16285863]

107. Remm M, Storm CE, Sonnhammer EL. J. Mol. Biol. 2001; 314:1041–1052. [PubMed: 11743721]

108. Tatusov RL, Koonin EV, Lipman DJ. Science. 1997; 278:631–637. [PubMed: 9381173]

109. Blacklock BJ, Jaworski JG. Biochem. Biophys. Res. Commun. 2006; 346:583–590. [PubMed: 16765910]

110. Joubes J, et al. Plant Mol. Biol. 2008; 67:547–566. [PubMed: 18465198]

111. Leonard AE, et al. Prog. Lipid Res. 2004; 43:36–54. [PubMed: 14636670]

112. Denic V, Weissman JS. Cell. 2007; 130:663–677. [PubMed: 17719544]

113. Beaudoin F, et al. J. Biol. Chem. 2002; 277:11481–11488. [PubMed: 11792704]

114. Jump DB. Methods Mol. Biol. 2009; 579:375–389. [PubMed: 19763486]

115. Toke DA, Martin CE. J. Biol. Chem. 1996; 271:18413–18422. [PubMed: 8702485]

116. Li X, et al. Mol. Plant. 2009; 2:138–151. [PubMed: 19529829]

117. Dunn TM, et al. Ann. Bot. 2004; 93:483–497. [PubMed: 15037448]

118. Feng Y, et al. Chem. Biodivers. 2012; 9:868–887. [PubMed: 22589089]

119. Li L, et al. Plant J. 2009; 58:485–498. [PubMed: 19154206]

120. Li L, et al. J. Exp. Bot. 2007; 58:3323–3342. [PubMed: 17890231]

121. Mentzen WI, Wurtele ES. BMC Plant Biol. 2008; 8:99. [PubMed: 18826618]

122. Mentzen WI, et al. BMC Plant Biol. 2008; 8:76. [PubMed: 18616834]

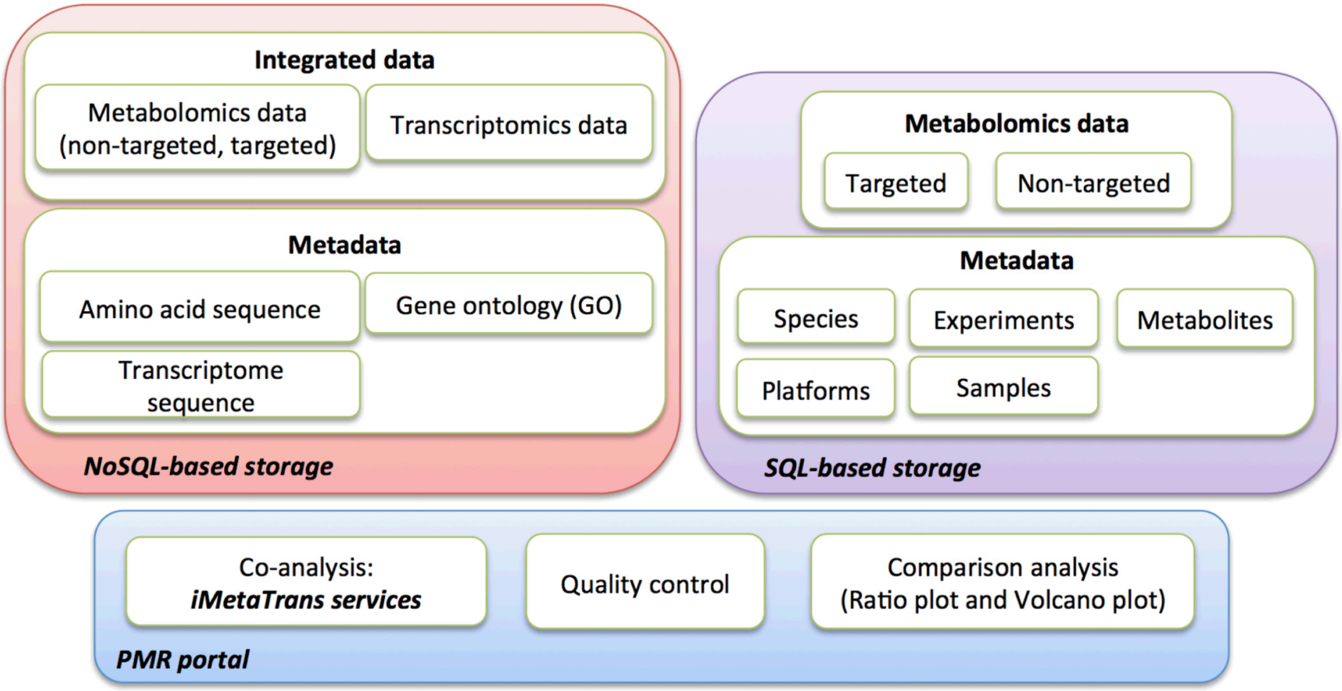


Fig. 1.

System architecture of the hybrid data-storage for PMR. *NoSQL-based storage for integrated data* contains transcriptomics and metabolomics data and metadata. This *NoSQL-based storage* system works with the *iMetaTrans services* for the integrative *co-analysis* of metabolomics, transcriptomics and metadata. *SQL-based storage* contains targeted and non-targeted metabolomics data and its metadata. The *PMR portal* provides quality control methods, comparison analyses and co-analysis.

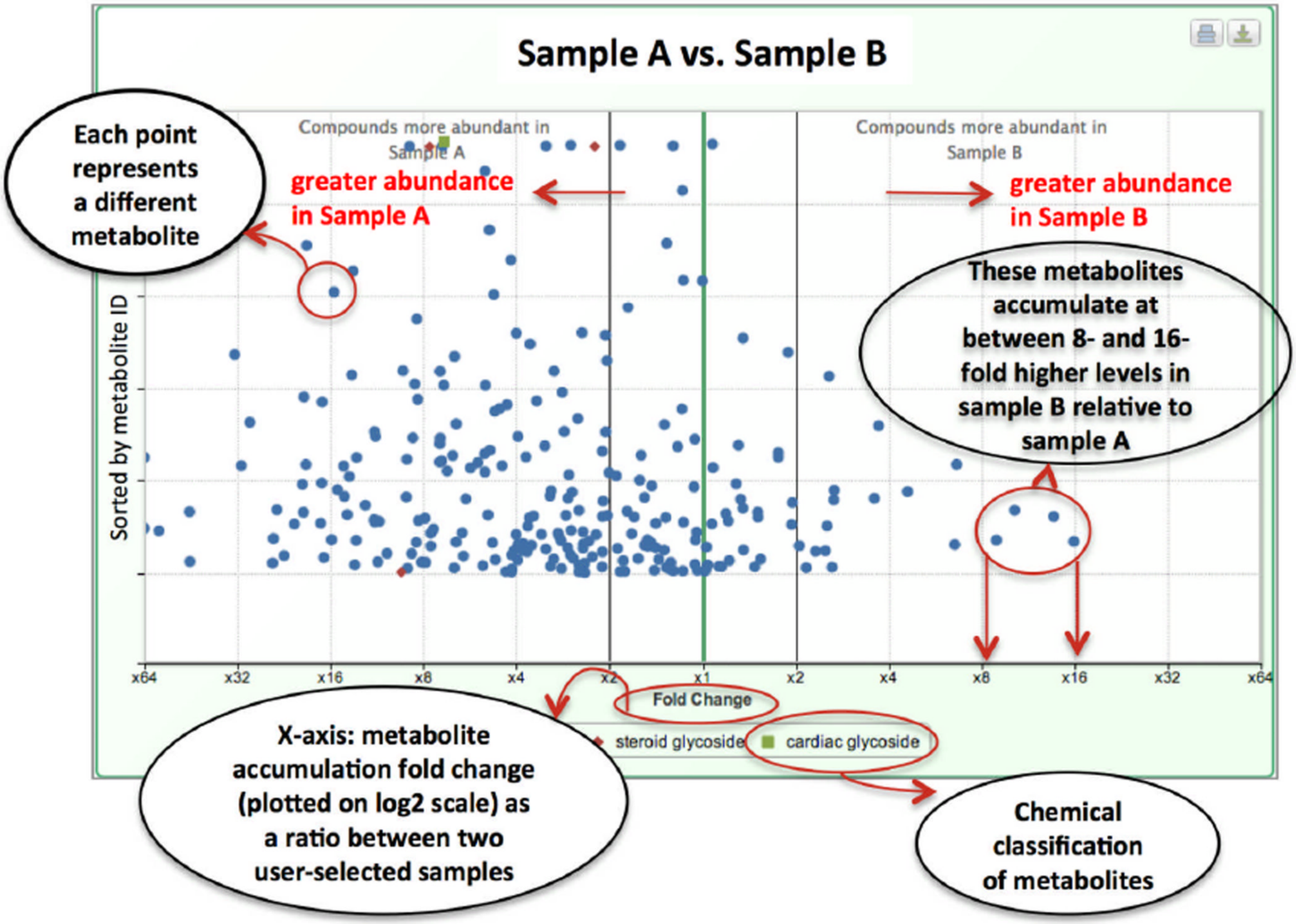


Fig. 2.

Log-ratio plot of metabolomics data. The x-axis is the mean ratio fold-change (plotted on a log 2 scale) of the relative abundance of each metabolite between the two samples that the user has selected. The y-axis represents each metabolite that has been analyzed. Metabolites whose abundance is unchanged between the two samples will plot at the x-axis origin (the green vertical line; ratio fold-change=1). Metabolites that hyper-accumulate in one of the two samples under analysis will plot either to the left or right of the x-axis origin.

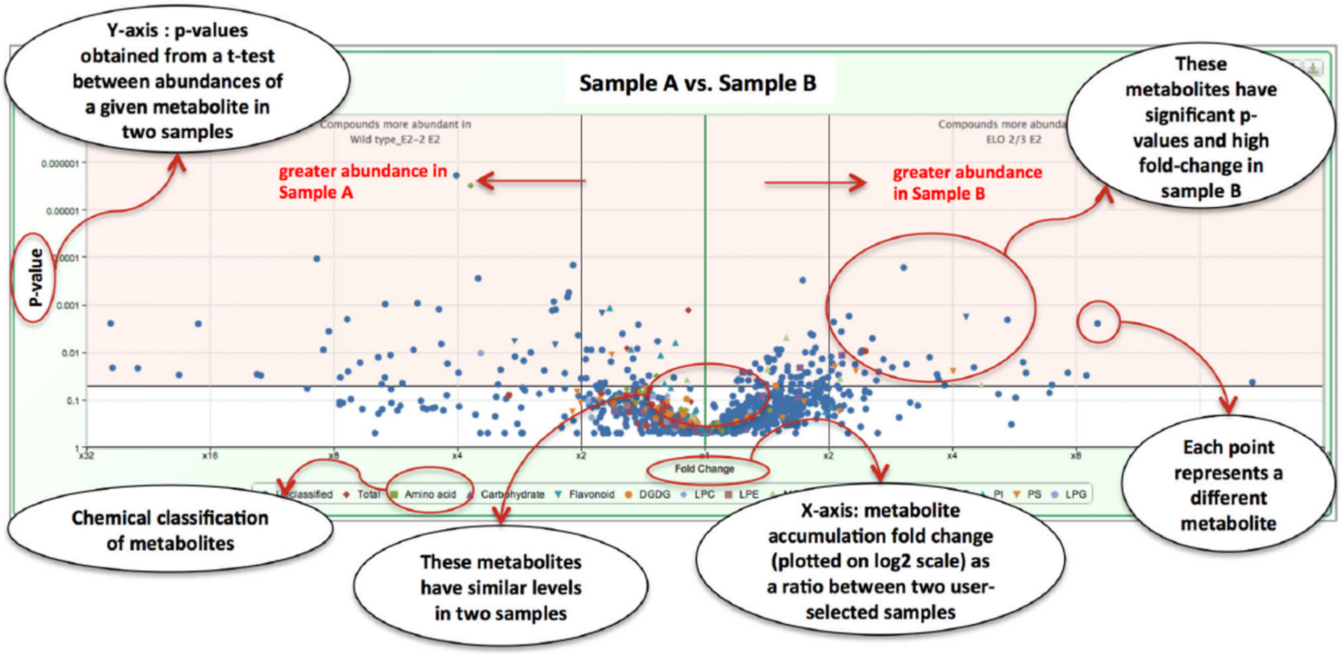


Fig. 3.

Volcano plot of metabolomics data. The x-axis is the mean ratio fold-change (plotted on a log 2 scale) of the relative abundance of each metabolite between the two samples that the user has selected (identical to the ratio plot, Figure 2). The y-axis represents the statistical significance p-value of the ratio fold-change for each metabolite. Metabolites whose abundance is unchanged between the two samples will plot at the x-axis origin (the green vertical line; ratio fold-change=1). Metabolites that hyper-accumulate in one of the two samples under analysis will plot either to the left or right of the x-axis origin. The order of the metabolites on the y-axis is determined by the statistical significance p-value of the ratio fold-change. The chemical nature of each metabolite is indicated by the color and shape of each data-point.

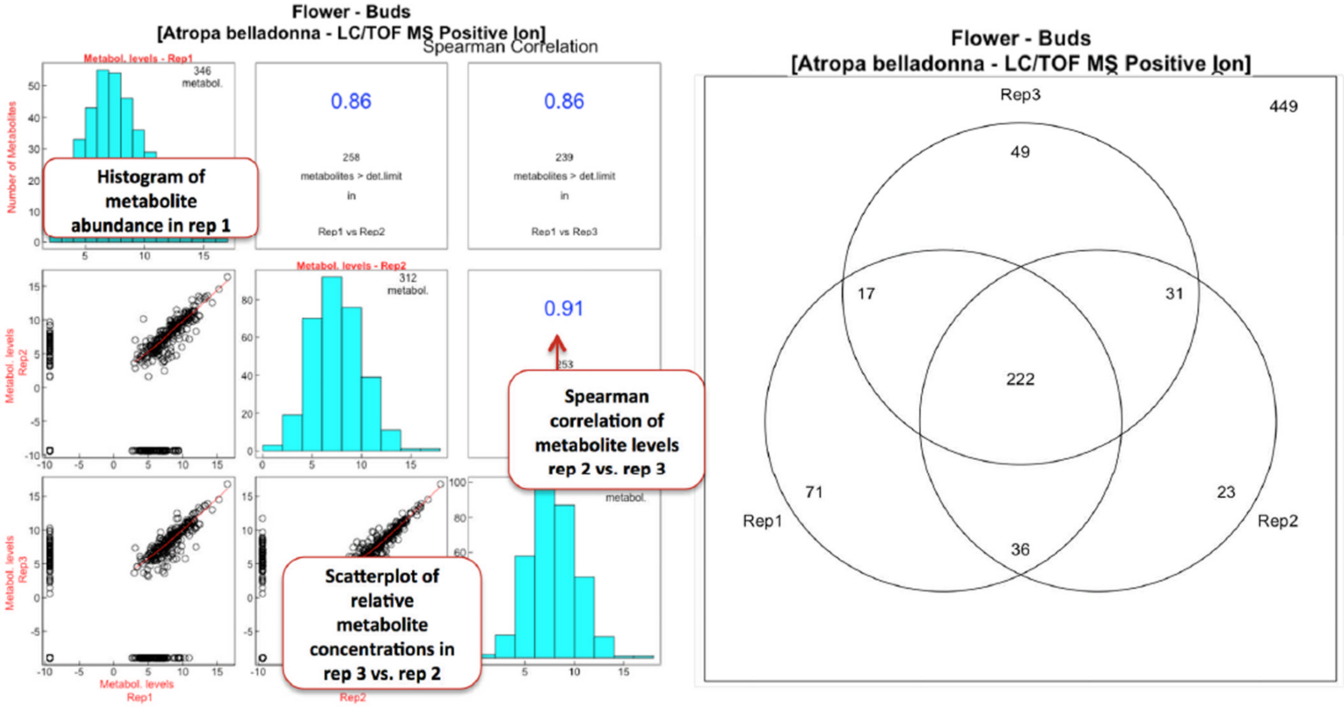


Fig. 4.

Reproducibility of abundance data for individual metabolites compared across biological replicates. The figure illustrates the relative abundance of metabolites measured in 3 biological replicates of *Atropa belladonna* flower buds. Left panel (bar charts): the log transformation of metabolite abundances shows that the data fits a normal distribution. The number of metabolites that are above the analytical detection limit in each biological replicate is shown at the upper right corner of the histograms. Scatterplots indicate the reproducibility of the abundance data for individual metabolites across pairs of biological replicates; each circle data-point represents an individual metabolite. Metabolites with similar abundances in both replicates are located near the diagonal line. Metabolites detected in only one biological replicate are indicated as data-points parallel to the x- and y-axis. Spearman correlations estimate overall data reproducibility across each pair of replicates. The number of metabolites above detection limit in each pair of biological replicates is given below the Spearman correlation coefficient. Right panel. Venn diagram shows the number of metabolites (above analytical detection limit) measured in the 3 biological replicates. Although the total number of distinct metabolites detected across the three biological replicates is 449 (upper right corner of the plot), only 222 metabolites were detected in all 3 biological replicates.

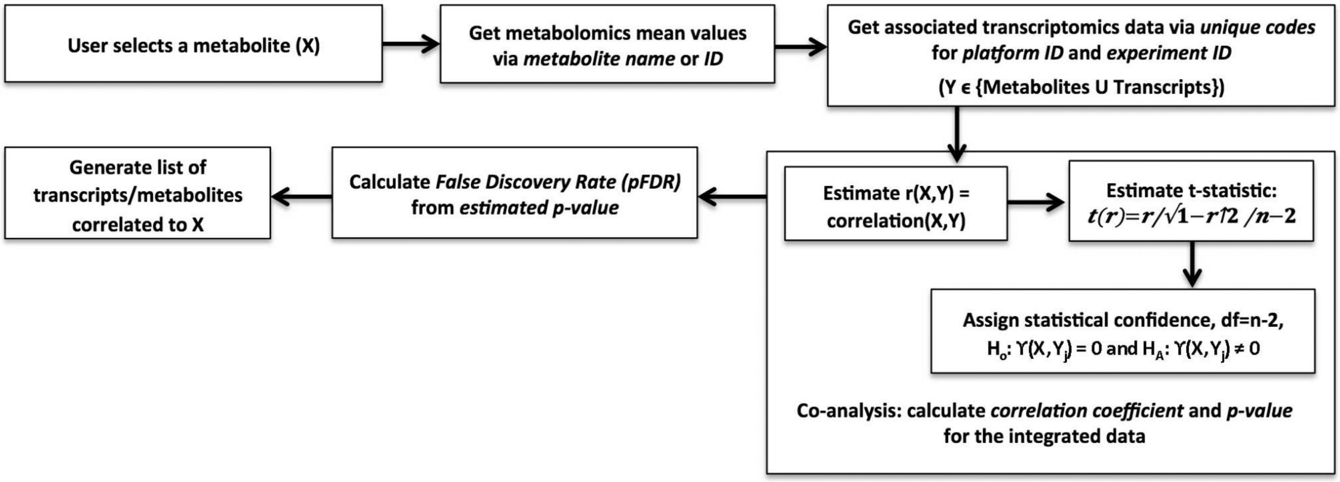
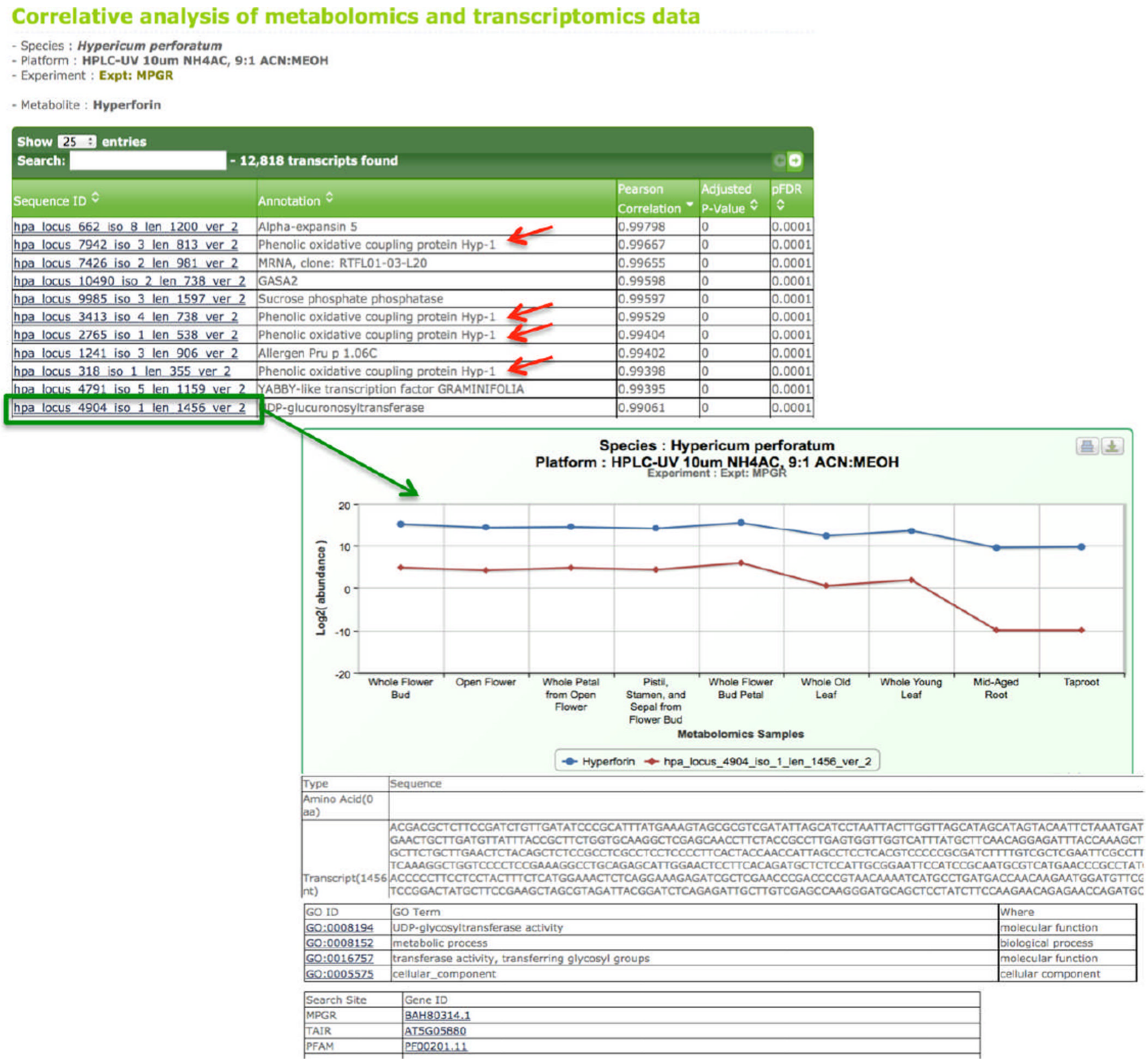

User selects a metabolite (X)
Get metabolomics mean values via metabolite name or ID
Get associated transcriptomics data via unique codes for platform ID and experiment ID
(Y ϵ {Metabolites U Transcripts})
Estimate r(X,Y) = correlation(X,Y)
Estimate t-statistic: t(r)=r/√1−r↑2 /n−2
Assign statistical confidence, df=n-2, Ho: ϒ(X,Yj) = 0 and HA: ϒ(X,Yj) ≠ 0
Co-analysis: calculate correlation coefficient and p-value for the integrated data
Calculate False Discovery Rate (pFDR) from estimated p-value
Generate list of transcripts/metabolites correlated to X

Fig. 5.

Flow-chart for parametric-based statistical approach to identify which the transcripts that have similar expression profiles as compared to the abundance of a user-selected metabolite (pFDR value). This approach has been implemented into *iMetaTrans services* in *PMR*. The user selects a metabolite within an experiment, and the unique metadata is used to retrieve the associated transcriptomic data. Metabolomic and transcriptomic data are integrated, and correlation coefficients and p-values are calculated. Finally, pFDR values are computed from the estimated p-values. The statistical approach to the co-analysis is details in section 4.1.



Correlative analysis of metabolomics and transcriptomics data

- Species : *Hypericum perforatum*
- Platform : **HPLC-UV 10um NH4AC, 9:1 ACN:MEOH**
- Experiment : **Expt: MPGR**

- Metabolite : **Hyperforin**

Show 25 entries
Search: - 12,818 transcripts found

Sequence ID	Annotation	Pearson Correlation	Adjusted P-Value	pFDR
hpa_locus_662_iso_8_len_1200_ver_2	Alpha-expansin 5	0.99798	0	0.0001
hpa_locus_7942_iso_3_len_813_ver_2	Phenolic oxidative coupling protein Hyp-1	0.99667	0	0.0001
hpa_locus_7426_iso_2_len_981_ver_2	MRNA, clone: RTFL01-03-L20	0.99655	0	0.0001
hpa_locus_10490_iso_2_len_738_ver_2	GASA2	0.99598	0	0.0001
hpa_locus_9985_iso_3_len_1597_ver_2	Sucrose phosphate phosphatase	0.99597	0	0.0001
hpa_locus_3413_iso_4_len_738_ver_2	Phenolic oxidative coupling protein Hyp-1	0.99529	0	0.0001
hpa_locus_2765_iso_1_len_538_ver_2	Phenolic oxidative coupling protein Hyp-1	0.99404	0	0.0001
hpa_locus_1241_iso_3_len_906_ver_2	Allergen Pru p 1.06C	0.99402	0	0.0001
hpa_locus_318_iso_1_len_355_ver_2	Phenolic oxidative coupling protein Hyp-1	0.99398	0	0.0001
hpa_locus_4791_iso_5_len_1159_ver_2	YABBY-like transcription factor GRAMINIFOLIA	0.99395	0	0.0001
hpa_locus_4904_iso_1_len_1456_ver_2	UDP-glucuronosyltransferase	0.99061	0	0.0001

Species : Hypericum perforatum
Platform : HPLC-UV 10um NH4AC, 9:1 ACN:MEOH
Experiment : Expt: MPGR

Type	Sequence
Amino Acid(0 aa)	
Transcript(1456 nt)	ACGACGCTCTTCCGATCTGTTGATATCCCGCATTTATGAAAGTAGCGCGTCGATATTAGCATCCTAATTACTTGGTTAGCATAGCATAGTACAATTCTAAATGAT GAACTGCTTGATGTTATTTACCGCTTCTGGTGCAAGGCTCGAGCAACCTTCTACCGCCTTGAGTGGTTGGTCATTTATGCTTCAACAGGAGATTTACCAAAGCT GCTTCTGCTTGAACTCTACAGCTCTCCGCCTCGCCTCCTCCCCTTCACTACCAACCATTAGCCTCCTCACGTCCCCCGCGATCTTTTGTCGCTCGAATTCGCCTT TCAAAGGCTGGTCCCCTCCGAAAGGCCTGCAGAGCATTGGAACTCCTTCACAGATGCTCTCCATTGCGGAATTCCATCCGCAATGCGTCATGAACCCGCCTAT ACCCCCTTCCTCCTACTTTCTCATGGAAACTCTCAGGAAAGAGATCGCTCGAACCCGACCCCGTAACAAAATCATGCCTGATGACCAACAAGAATGGATGTTCG TCCGGACTATGCTTCCGAAGCTAGCGTAGATTACGGATCTCAGAGATTGCTTGTCGAGCCAAGGGATGCAGCTCCTATCTTCCAAGAACAGAGAACCAGATGC

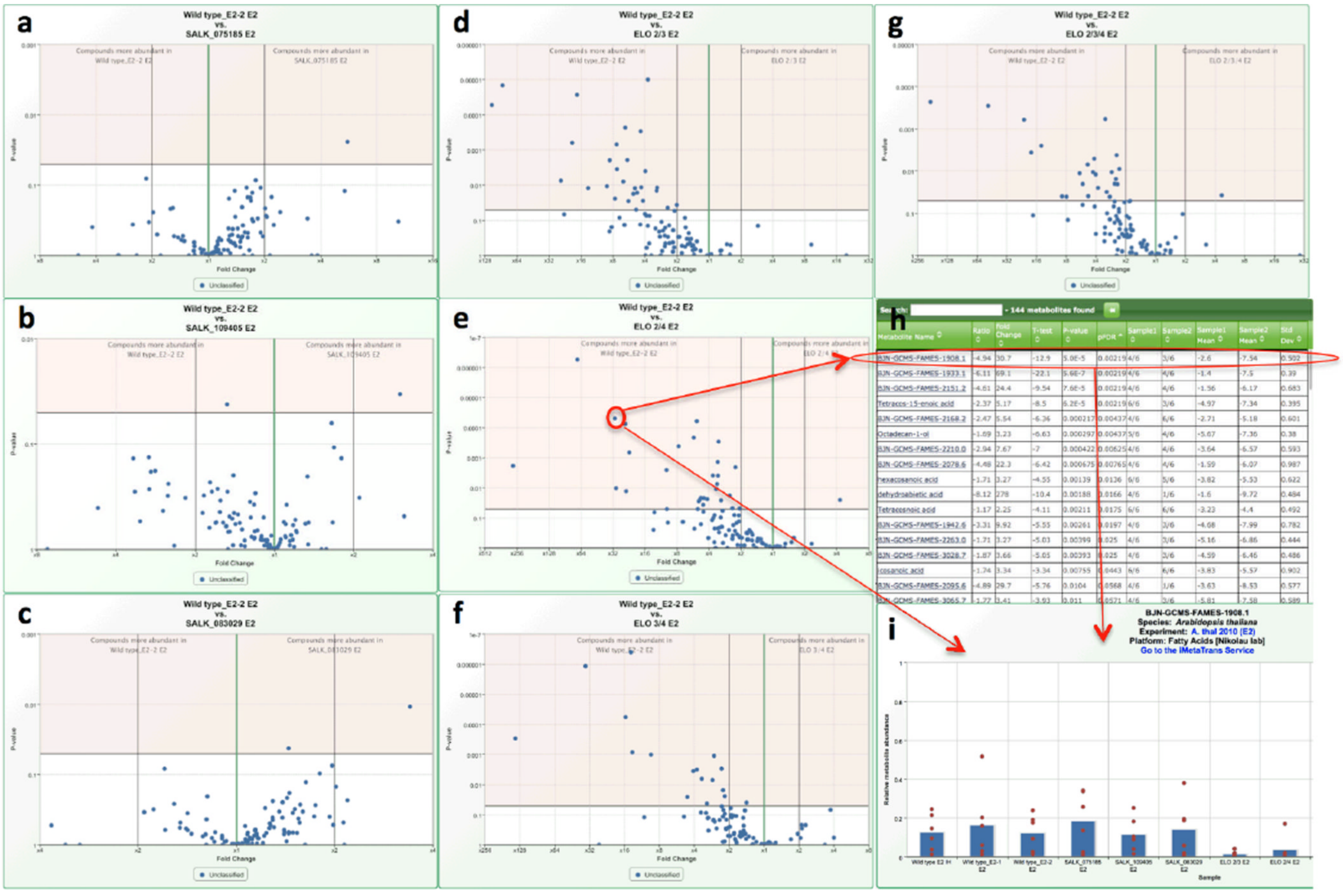
GO ID	GO Term	Where
GO:0008194	UDP-glycosyltransferase activity	molecular function
GO:0008152	metabolic process	biological process
GO:0016757	transferase activity, transferring glycosyl groups	molecular function
GO:0005575	cellular_component	cellular component

Search Site	Gene ID
MPGR	BAH80314.1
TAIR	AT5G05880
PFAM	PF00201.11

Fig. 6.

Interactive co-analysis table showing the correlation between the abundance of a user-selected metabolite with abundances of all transcripts in a dataset, as implemented in *PMR*. The table in the upper panel exemplifies the correlation between the abundance of the bioactive molecule hyperforin to the abundances of all transcripts obtained from RNAseq analysis of reproductive and vegetative organs of *Hypericum perforatum*. The results include Pearson correlation coefficients, p-values and pFDR values. This table contains all transcripts (12,818) whose accumulation is correlated to hyperforin with a pFDR value of < 0.05 (a user-determined parameters). The list is ordered by most significant pFDR values; thus, of the 30,000 transcripts detected by RNAseq analysis in *H. perforatum*, hyperforin accumulation four Hyp-1 like transcripts (red arrows) are among the 10 most tightly correlated. The experimental data used in these analysis are from the Medicinal Plants Genomics Consortium; biological materials and metabolomics analysis[32, 33] were from Matt

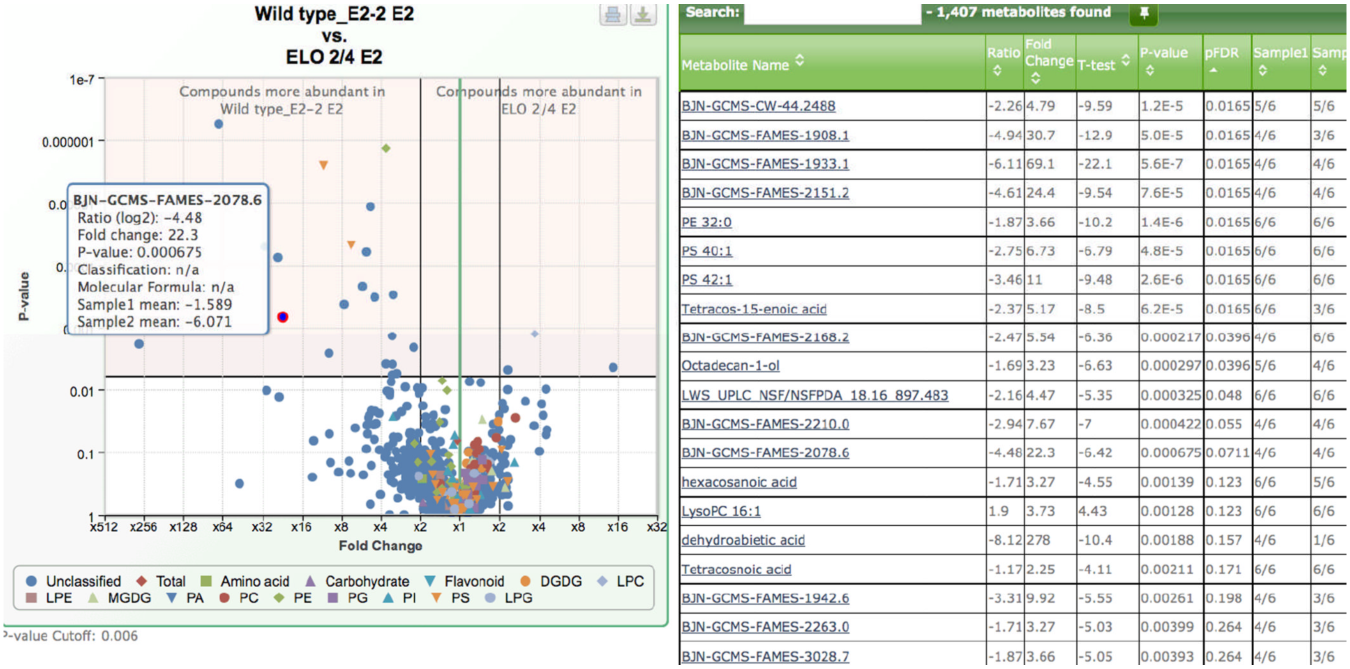
Crispin, Iowa State University. Clicking on Sequence ID row for UDP-glucuronosyl transferase (green arrow) brings the user to detailed information for that sequence (lower panel). This latter page provides GO annotations, external links associated with the selected Sequence ID, and interactive comparisons between transcriptomics data and metabolomics data. Red represents transcriptomic data and blue represents user-selected data.



Metabolite Name	Ratio	Fold Change	T-test	P-value	pFDR	Sample1	Sample2	Sample1 Mean	Sample2 Mean	Std Dev
BJN-GCMS-FAMES-1908.1	-4.94	30.7	-12.9	5.0E-5	0.00219	4/6	3/6	-2.6	-7.54	0.502
BJN-GCMS-FAMES-1933.1	-6.11	69.1	-22.1	5.6E-7	0.00219	4/6	4/6	-1.4	-7.5	0.39
BJN-GCMS-FAMES-2151.2	-4.61	24.4	-9.54	7.6E-5	0.00219	4/6	4/6	-1.56	-6.17	0.683
Tetracos-15-enoic acid	-2.37	5.17	-8.5	6.2E-5	0.00219	6/6	3/6	-4.97	-7.34	0.395
BJN-GCMS-FAMES-2168.2	-2.47	5.54	-6.36	0.000217	0.00437	4/6	6/6	-2.71	-5.18	0.601
Octadecan-1-ol	-1.69	3.23	-6.63	0.000297	0.00437	5/6	4/6	-5.67	-7.36	0.38
BJN-GCMS-FAMES-2210.0	-2.94	7.67	-7	0.000422	0.00625	4/6	4/6	-3.64	-6.57	0.593
BJN-GCMS-FAMES-2078.6	-4.48	22.3	-6.42	0.000675	0.00765	4/6	4/6	-1.59	-6.07	0.987
hexacosanoic acid	-1.71	3.27	-4.55	0.00139	0.0136	6/6	5/6	-3.82	-5.53	0.622
dehydroabietic acid	-8.12	278	-10.4	0.00188	0.0166	4/6	1/6	-1.6	-9.72	0.484
Tetracosanoic acid	-1.17	2.25	-4.11	0.00211	0.0175	6/6	6/6	-3.23	-4.4	0.492
BJN-GCMS-FAMES-1942.6	-3.31	9.92	-5.55	0.00261	0.0197	4/6	3/6	-4.68	-7.99	0.782
BJN-GCMS-FAMES-2263.0	-1.71	3.27	-5.03	0.00399	0.025	4/6	3/6	-5.16	-6.86	0.444
BJN-GCMS-FAMES-3028.7	-1.87	3.66	-5.05	0.00393	0.025	4/6	3/6	-4.59	-6.46	0.486
icosanoic acid	-1.74	3.34	-3.34	0.00755	0.0443	6/6	6/6	-3.83	-5.57	0.902
BJN-GCMS-FAMES-2095.6	-4.89	29.7	-5.76	0.0104	0.0568	4/6	1/6	-3.63	-8.53	0.577
BJN-GCMS-FAMES-3065.7	-1.77	3.41	-3.93	0.011	0.0571	4/6	3/6	-5.81	-7.58	0.589

Fig. 7.

Exemplary use of interactive volcano plots comparing fatty acid accumulation in *ELO* mutants; visualized using *PMR*. These data are chosen by selecting a species (*Arabidopsis thaliana*) and "Experiment 2" labeled as "A. thal 2010 (E2)". These plots compare the fatty acid data for each *Atelo* knock-out to wild-type (E2-2 E2) generated by the Fatty Acid analytical platform. **a)** *Atelo2*-KO against wild-type; **b)** *Atelo3*-KO against wild-type; **c)** *Atelo4*-KO against wild-type; **d)** *Atelo2*: :*Atelo3*-KO against wild-type; **e)** *Atelo2*: :*Atelo4*-KO against wild-type; **f)** *Atelo3*: :*Atelo4*-KO against wild-type and **g)** *Atelo2*: :*Atelo3*: :*Atelo4*-KO against wild-type. Blue data-points represent relative abundances of individual metabolites. Statistical significance is indicated within the pale-pink shaded regions on the plots, as defined by p-value of <0.05. Metabolites to the left of the green x-axis origin (×1 fold change line) are less abundant in the knock-out mutant lines, whereas metabolites to the right of this origin are more abundant in the knock-out lines. **h)** clicking on any metabolite data-point on these panels (e.g., clicking on the metabolite in Panel **e**) brings the user to the associated statistical information for that metabolite (compound assignment, ratio, fold-change, t-test results, p-value, pFDR [in this case 5 × 10^{-5}]. In this case, its level is 30-fold more in the wild type control. **i)** Clicking on metabolite name in panel **h)** leads the user to a bar plot, which reveal that each of the double and triple ELO mutants express lower levels of this putative fatty acid.



BJN-GCMS-FAMES-2078.6
Ratio (log2): -4.48
Fold change: 22.3
P-value: 0.000675
Classification: n/a
Molecular Formula: n/a
Sample1 mean: -1.589
Sample2 mean: -6.071

P-value Cutoff: 0.006

Search: - 1,407 metabolites found

Metabolite Name	Ratio	Fold Change	T-test	P-value	pFDR	Sample1	Samp
BJN-GCMS-CW-44.2488	-2.26	4.79	-9.59	1.2E-5	0.0165	5/6	5/6
BJN-GCMS-FAMES-1908.1	-4.94	30.7	-12.9	5.0E-5	0.0165	4/6	3/6
BJN-GCMS-FAMES-1933.1	-6.11	69.1	-22.1	5.6E-7	0.0165	4/6	4/6
BJN-GCMS-FAMES-2151.2	-4.61	24.4	-9.54	7.6E-5	0.0165	4/6	4/6
PE 32:0	-1.87	3.66	-10.2	1.4E-6	0.0165	6/6	6/6
PS 40:1	-2.75	6.73	-6.79	4.8E-5	0.0165	6/6	6/6
PS 42:1	-3.46	11	-9.48	2.6E-6	0.0165	6/6	6/6
Tetracos-15-enoic acid	-2.37	5.17	-8.5	6.2E-5	0.0165	6/6	3/6
BJN-GCMS-FAMES-2168.2	-2.47	5.54	-6.36	0.000217	0.0396	4/6	6/6
Octadecan-1-ol	-1.69	3.23	-6.63	0.000297	0.0396	5/6	4/6
LWS_UPLC_NSF/NSFPDA_18.16_897.483	-2.16	4.47	-5.35	0.000325	0.048	6/6	6/6
BJN-GCMS-FAMES-2210.0	-2.94	7.67	-7	0.000422	0.055	4/6	4/6
BJN-GCMS-FAMES-2078.6	-4.48	22.3	-6.42	0.000675	0.0711	4/6	4/6
hexacosanoic acid	-1.71	3.27	-4.55	0.00139	0.123	6/6	5/6
LysoPC 16:1	1.9	3.73	4.43	0.00128	0.123	6/6	6/6
dehydroabietic acid	-8.12	278	-10.4	0.00188	0.157	4/6	1/6
Tetracosnoic acid	-1.17	2.25	-4.11	0.00211	0.171	6/6	6/6
BJN-GCMS-FAMES-1942.6	-3.31	9.92	-5.55	0.00261	0.198	4/6	3/6
BJN-GCMS-FAMES-2263.0	-1.71	3.27	-5.03	0.00399	0.264	4/6	3/6
BJN-GCMS-FAMES-3028.7	-1.87	3.66	-5.05	0.00393	0.264	4/6	3/6

Fig. 8.

Exemplary use of interactive volcano plots comparing the metabolite accumulation patterns from six analytical platforms in *Atelo2*: :*Atelo4*-KO mutant versus wild-type control (E2-2 E2); visualized using *PMR*. The plot represents all available metabolite data (1407 metabolites). Individual metabolites are plotted according to classification, and unclassified metabolites appear in blue. Significance is indicated within the pale-pink shaded regions on the plot, and p-value of < 0.05 was used during the analysis. Mousing over a metabolite brings up its identity and statistics. The table at the right (sorted by pFDR value) shows metabolites whose level is significantly altered. Metabolites to the left of the green ×1 fold change line are less abundant in the knock-out, whereas metabolites to the right are more abundant in the knock-out. Almost every metabolite whose accumulation is altered in an ELO mutant is a known or suspected fatty-acid derived compound.

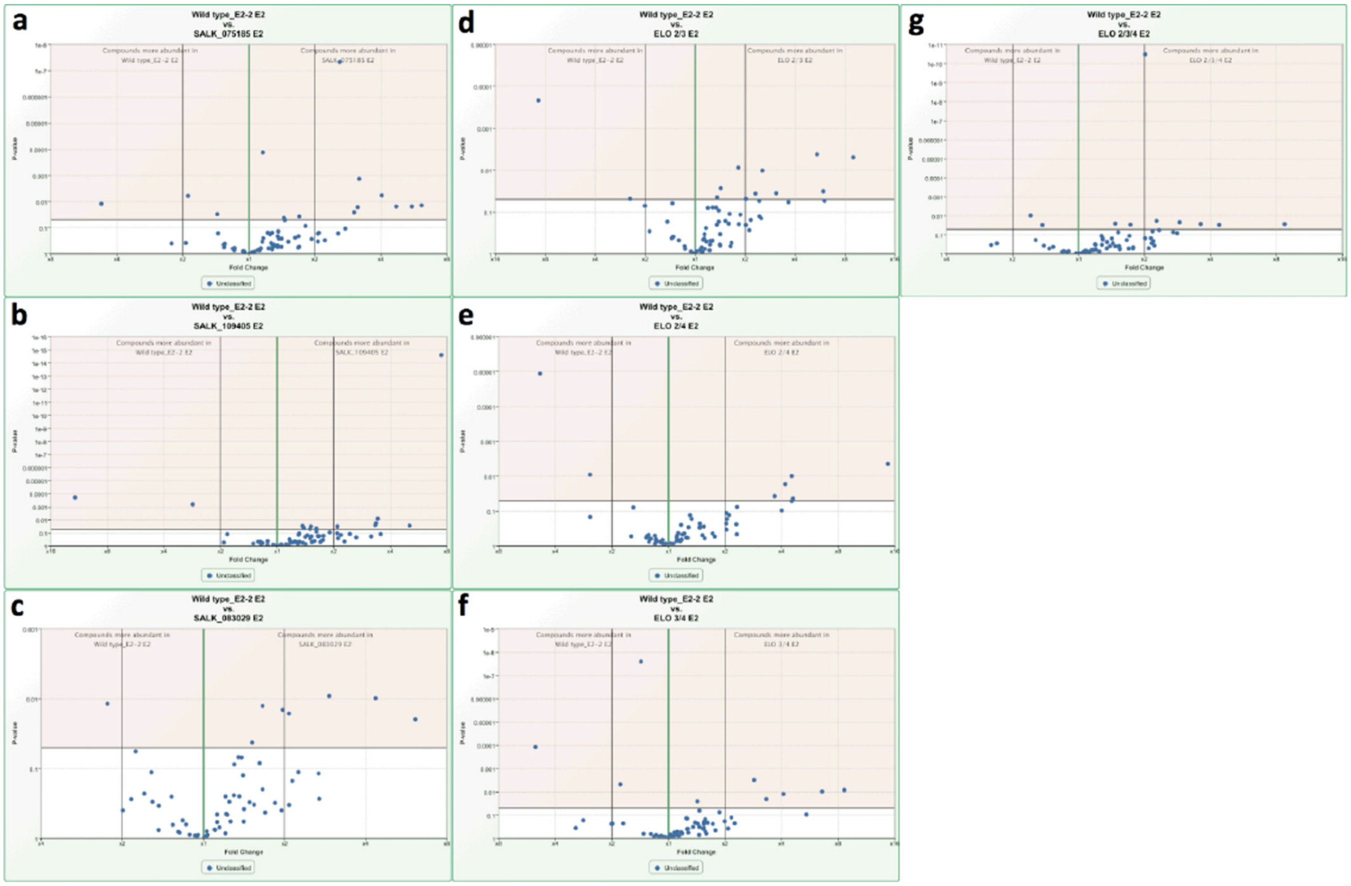


Fig.9.

Exemplary use of volcano plots comparing cuticular wax accumulation; visualized using *PMR*. These plots compare the data generated by the Cuticular Wax analytical platform for each *Atelo* knock-out line to the wild-type lines (E2-2 E2). **a)** *Atelo2*-KO against wild-type; **b)** *Atelo3*-KO against wild-type; **c)** *Atelo4*-KO against wild-type; **d)** *Atelo2*: :*Atelo3*-KO against wild-type; **e)** *Atelo2*: :*Atelo4*-KO against wild-type; **f)** *Atelo3*: :*Atelo4*-KO against wild-type and **g)** *Atelo2*: :*Atelo3*: :*Atelo4*-KO against wild-type. Blue data-points represent relative abundances of individual metabolites. Statistical significance is indicated within the shaded regions on the plot. Metabolites to the left of the green x-axis origin (×1 fold change line) are less abundant in the knock-out mutant lines, whereas metabolites to the right of this origin are more abundant in the knock-out mutant lines.

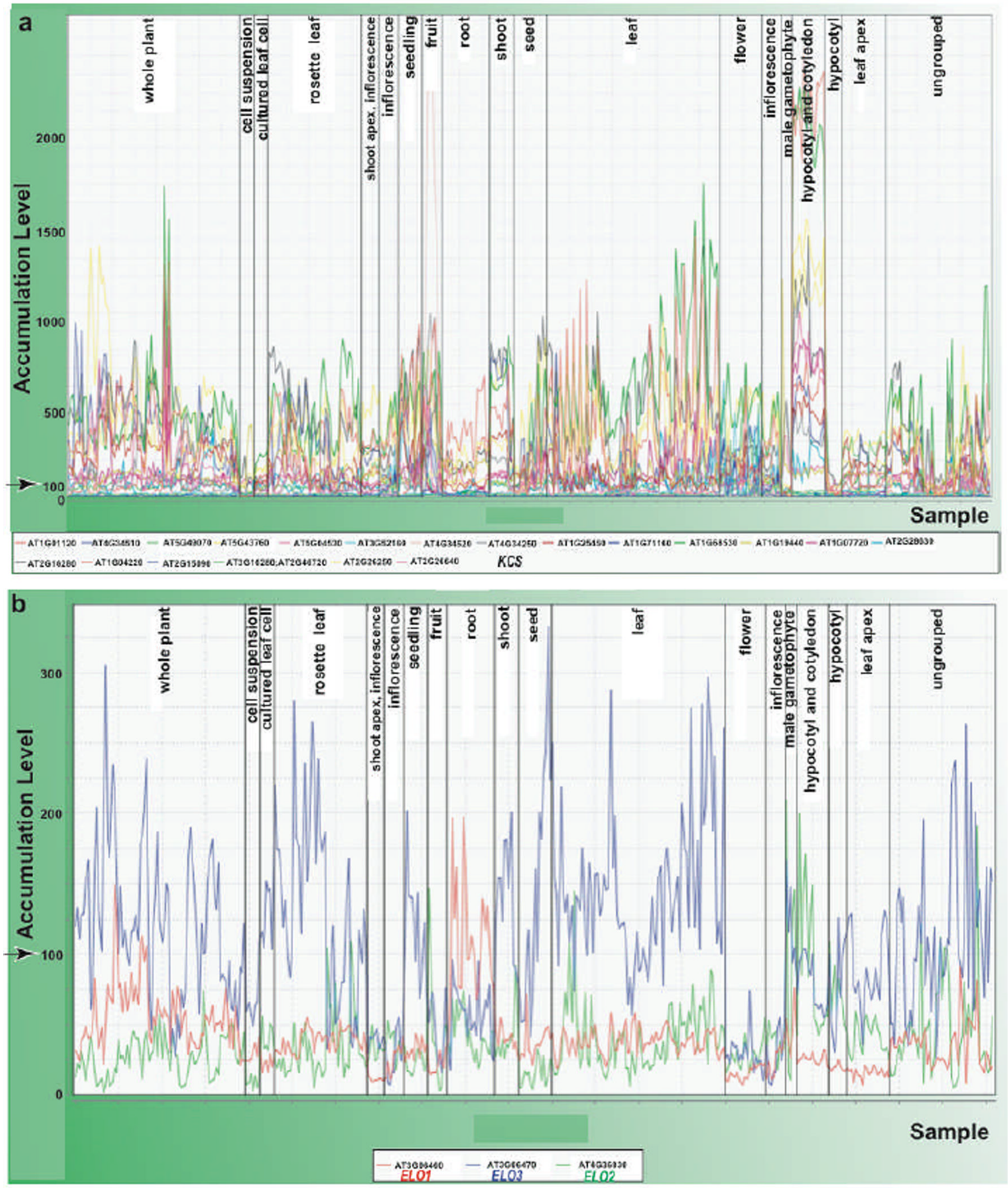


Fig. 10.

Accumulation of 21 *KCS* and three *ELO*-like transcripts (*ELO1*, *ELO2*, *ELO3*) across multiple conditions. Each point on the x-axis represents an mRNA transcriptomics profiling data for the mean of the replicates for a given experimental sample. Samples are derived from publicly available 956 Affymetrix ATH1 chip transcriptomic experiments (Mentzen and Wurtele, 2008). The y-axis represents the normalized expression level for the user-selected genes. The average transcript accumulation level for each chip is normalized to a value of 100 (arrow pointed). These data are visualized using MetaOmGraph software (Feng *et al*., 2012; http://www.metnetdb.org). a) Most of the 21 *KCS* transcripts are more abundant than the three *ELO*-like transcripts, and all are diversely expressed across different conditions. b) Three *ELO* genes have distinct expression patterns. *ELO4* (AT1G75000) transcript is not represented in the ATH1 chip.