DISCOVERING FUNCTIONAL ANNOTATION THROUGH DATA MINING OF LARGE SCALE PHENOMICS IN *ARABIDOPSIS THALIANA*

By

Shannon Marie Bell

A DISSERTATION

Submitted to
Michigan State University
In partial fulfillment of the requirements
For the degree of

DOCTOR OF PHILOSOPHY

Biochemistry and Molecular Biology

2012

ABSTRACT

DISCOVERING FUNCTIONAL ANNOTATION THROUGH DATA MINING OF LARGE SCALE PHENOMICS IN *ARABIDOPSIS THALIANA*

By

Shannon Marie Bell

To address society's biotechnology needs in agriculture, medicine, and beyond, a better understanding of the flow of information from gene to protein to phenotype is needed. However, despite the increasing amount of genome-scale (omic) data, the lack of annotation providing insight into gene function remains a challenge for researchers. The lack of functional annotation can hinder progress from targeted metabolic engineering to foundational biological research. Vague annotations coming from an expression profile or sequence similarity make it hard to design experiments to characterize the gene and can lead researchers down the wrong path. Using large-scale phenomics will provide more useful information to help guide researchers in the characterization of under-annotated genes. Unfortunately, many of the tools needed to carry out analyses of large-scale phenotypic data are lacking.

This work presents a suite of software tools developed to address this need. MIPHENO introduces a workflow to enable the post hoc analysis of screening data from quality control to normalization to prediction of individuals likely to show a response. The NetComp suite features an algorithm, SimMeasure, to calculate the similarity between individuals in the presence of missing data. SimMeasure also works with datasets that have

been thresholded to remove values under/above a given response value. It also features several additional functions aimed at data integration and network comparisons.

Results of these methods applied to a large phenotypic screen of gene disruption lines in *Arabidopsis thaliana* demonstrate the utility of these tools in the analysis of large-scale datasets. They show that phenotypic data can be successfully used in an analogous manner to other high throughput data to build models of gene function. This work presents a novel use of high throughput phenotypic data in higher organisms to build models for functional annotation. Together this work presents the next step in the analysis of omics data and moves the field closer to improving annotation quality.

ACKNOWLEDGMENTS

I would like to thank all those who supported me in the course of my education. To my committee, thank you for your time and expertise. To my advisor, thank you for your time, support, and flexibility to allow me to pursue the things that mattered to me. I wish to thank my friends for their patience, honesty, and support. I also thank my furry boys for their attention, supervision, lap warming, and comic relief. To my poor husband and best friend, I couldn't have made it without you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

Data explosion, a lack of functional annotation, and a role for high throughput screening

A better understanding of information flow from DNA to RNA to protein to phenotype is needed to address a whole host of challenges in the 21st century. Advances in personalized medicine, food security and nutrition, and mitigating the impact of industrial demands on the environment depend on our ability to gather, interpret, and anchor cellular data to biological processes. Ideally, this information will aid in predictive and targeted metabolic manipulation strategies, through genetic engineering of plants and microbes and, through gene and drug therapies. The post-genomic era of science and the ability to conduct big-data science have opened doors, providing data at a resolution not available before.

Unfortunately, the genome for many organisms is far from complete. While the DNA sequence of model organisms is available, the function of many genes is not. Genomics studies using the sequence information have improved with better structural and bioinformatic modeling tools and by using information from other species on structurally similar genes and proteins. However, some aspects of biology are not universal and using sequence similarity alone can lead to misannotation that complicates interpretation and experiments if they rely too heavily on these annotations (Furnham et al., 2009). Transcriptomics and co-expression studies using annotation enrichment have provided insight on the putative function of many transcripts (Horan et al., 2008). These expression-based studies lack the ability to reveal the definitive function of a gene and provide limited

information to aid in follow-up experiments. This is because the expression profile cannot necessarily be tied directly back to an effect (phenotype), making it challenging to design a follow-up experiment to probe the effect. Proteomics has been critical in assigning subcellular location and protein interaction networks provide information on interacting partners and protein complexes, with an expansion of resources making it easier to leverage these datasets (Wang et al., 2012). Unfortunately, neither helps with follow-up experiments, unless the protein of interest partners or clusters with other proteins that are well described functionally. What is missing is a biological phenotype that can provide insight as to the role of a gene within its system.

Phenotypic studies, particularly broad-spectrum metabolomics, which measure many different types of metabolites, have the potential to provide information needed to identify genes that affect the biochemistry within the organism. Even in the relatively simple plant model species *Arabidopsis thaliana,* researchers must look for traits in the context of whole genome duplication, large gene families, promiscuous enzymes, and complex metabolic feedback loops (Ober, 2010; The Arabidopsis Genome Initiative, 2000). Knockout mutants often have no phenotype due to genetic redundancy, lack of the right environmental conditions to observe a phenotype, or measuring for the wrong phenotype (Bouché and Bouchez, 2001). As the costs of high throughput technology decreases, metabolite screens are becoming more accessible to researchers.

The hypothesis that underlies this dissertation project is that metabolic data from less-targeted screens surveying multiple measurements may be used in an analogous fashion to transcriptomic or proteomics data. In particular, that data from gene disruption

lines (where a specific gene is knocked out or over expressed) can provide insight into that gene's role within the plant (Thorneycroft et al., 2001). By building from methods developed for analyzing other high-throughput datasets, for example using gene ontology or annotation enrichment, it will be possible to characterize phenotypically similar gene disruption lines by their common phenotype and by enriched annotation of the group. This approach should facilitate the design of follow-up experiments. If high-throughput metabolite screening data can be treated like other high-throughput datasets such as transcriptomics, data integration from disparate sources might be used to aid in functional gene annotation (Yuan et al., 2008).

Based on the hypotheses outlined above, there are three major challenges to this work. First, a dataset containing a large set of gene disruption lines from both annotated and unannotated genes as well as diverse metabolic measurements is needed. Measuring several different metabolite types, ideally in different tissues, is important for capturing tissue-specific metabolism and because members of gene families sometimes have preferential tissue expression. Having some level of annotation is key to testing a proof of concept (things we know are behaving as expected), as well as to help inform the generation of hypotheses for unannotated genes that behave similarly. Second, the dataset must be in a useful format for analysis, which is one where the information can be adequately compared across all samples and observed or predicted changes reflect the metabolic phenotype of the individual. Measurements that relationships are built upon must be connected to an underlying perturbation in the metabolic network (ideally resulting from a disruption in the gene of interest). If not, then the data will not be any

more informative for designing follow-up experiments than transcriptomic data. Lastly, the ability to build a correlation-type matrix or adjacency matrix is desiarable to facilitate the downstream processing typically used in transcriptomics (e.g. cluster enrichment calculations), as well as to allow data integration with other omics datasets. These matrices describe the relationship between one individual (for example, gene distribution line) to another and are a common tool for large data analysis, as will be described later.

High throughput phenomics dataset: Chloroplast 2010

The Chloroplast 2010 project is a high throughput screen of over five thousand *Arabidopsis thaliana* mutants, including T-DNA insertion lines (Alonso et al., 2003) and characterized mutants, most of which have been predicted to be chloroplast targeted (Lu et al., 2008; Lu et al., 2011b; Lu et al., 2011c). A wide range of primary metabolites were surveyed through this study: leaf and seed free amino acids, leaf free fatty acids, and the seed carbon and nitrogen levels. The major goal of the Chloroplast 2010 project is to further characterize the genes of the chloroplast; as the chloroplast carries out photosynthesis and is involved in *de novo* fatty acid biosynthesis as well as the production of many amino acids. These measurements provide a wealth of diverse information about the behavior of the putative knockout and are hypothesized to provide functional information about the function of the missing gene.

This dataset has the potential to provide insight into the function of under-annotated genes. Similar to most high-throughput screening studies, it has features that make it challenging to use the data for developing broader hypotheses about the individuals being screened, beyond simply identifying which individuals should be

prioritized for further study. These types of screening studies typically have data collected over the course of many years, and lack the replication and explicit controls necessary to carry out traditional variance-normalization methods used in small–scale experiments and some large-scale studies, such as microarrays. These aspects tend to be inherent to screening studies because the researcher is often interested in large changes that may be easily observable by comparing individuals within a group (Jander et al., 2004), and to keep the costs down while screening as many candidates as possible. However, these aspects make it difficult to perform cross-dataset comparisons of the data: hence a processing method is needed to overcome these issues.

Development of analysis methods

1.3.1 Normalization

Data normalization is a process that removes technical variance while preserving the biological variance of interest. Very few normalization methods for screening of high-throughput datasets exist, but there is a large body of literature focused on normalization of microarrays (Eckel et al., 2005; Quackenbush, 2002). The biggest limitation with many screening studies, including the Chloroplast 2010, is the lack of a common control line or replication between sample runs that would allow for an estimation of the technical variance. Fortunately, some concepts for high-throughput screening studies that facilitate their analysis may be utilized. Chiefly, most observations should be in a 'normal' range, such as the assumption in transcript studies that expression levels of most genes are constant. For the Chloroplast 2010 data, this means that most of the observed responses will be in the background or wild type range, which is supported by prior findings

(Barbaric et al., 2007; Bouché and Bouchez, 2001; Jander et al., 2004). Additionally, the same individuals are not in each sample set and many individuals die (plants fail to germinate), but algorithms dealing with some of these issues in expression sets have been published, which can be built upon to address the issues in screening datasets (Mar et al., 2009).

It is proposed that a method be developed, built on methods for expression data, facilitating the use of the Chloroplast 2010 dataset while being extendable to other high throughput datasets. Currently there is no published normalization method that addresses the needs of these large screening studies characterized by little/no replication, uneven sample sizes/missing data, and lack of controls. Further, a way to quantify the 'response' of an individual (for example, which ones are likely high metabolite accumulators versus those which are likely behaving as wild-type) is also necessary. Quantifying the assay response aids in prioritizing individuals for follow-up studies and for making comparisons between individuals in terms of the magnitude of assay response. The development of these tools will open up high-throughput screening datasets and drive integration with other omic data which is not currently possible using existing methodologies.

1.3.2 Correlation calculations with missing data

High-throughput screens are often carried out under the assumption that few measured responses are going to be changing or different from the bulk of the observed responses. The implication of this assumption is that relationships, for example correlations between individuals, may be based purely upon having a wild type or background level response. Relationships like these are counterproductive to the aim of

functional characterization of a gene as they do not highlight the traits that are altered when the gene is missing. Furthermore, building a relationship only based on the cases where both individuals have observations (i.e. pairwise complete observation) could bias results in which one individual had many responses above a threshold (pleiotropic) and the other only had one.

Another aspect that needs to be addressed is that these datasets may also be prone to missing data. Data in this case may be missing completely at random (MCAR; events leading to the missing data are statistically independent of the individual and the unobserved attribute), missing at random (MAR; statistically independent of the missing value itself, but after controlling for some external factor), or missing not at random (MNAR; lack of an observation depends on the value of that observation) (Schlomer et al., 2010). This is a hierarchy, where if conditions for MCAR are not met then MAR is considered and so on. If one considers the Chloroplast 2010 dataset, data may be missing because a sample was not available for that analysis (e.g. the plant died; MCAR), or it failed some quality control parameter (MAR). Additionally, if one were to remove data that was within the range for background signal, this would add missing values that are MNAR.

While methods concerning missing data for microarray analysis and other omics data do exist (Aittokallio, 2010), many are not aimed at handling missing data of the type described. Simple methods for dealing with missing data include omitting the missing pairs when calculating the value, referred to as using pair-wise complete observations, or replacing the missing value with a zero or the row/column mean. There are also more sophisticated methods that have been shown far superior and tend to fall into two general

categories (Liew et al., 2011) applicable to the discussion here: global and local. Global methods use information on the entire dataset and include methods like Bayesian principle component analysis (BPCA), which incorporates prior information (generally uninformative prior distribution) into the model and does not require model parameters to be specified by the user (Oba et al., 2003). Local methods, in contrast, use a subset of the data that is similar to the individual with missing data such as K-nearest neighbor and local least squares. K-nearest neighbor and similar clustering approaches use information from K-closely related genes to obtain the missing value (Liew et al., 2011). This approach works well if values in the dataset share a high amount of similarity (or correlation). It also requires some advanced determination of the parameter K. When employing local least squares, and other least square regression methods, a linear model is assumed between the gene with missing values and those with similar values (e.g., K most correlated genes). The least squares estimate can be calculated from each of the similar genes and combined for a final estimate (Liew et al., 2011; Stacklies et al., 2007). This method also requires the user to provide K and potentially the correlation parameter to use.

The methods described above all seek to impute a missing value such that the downstream analyses can be carried out. For gene expression these analyses are typically differential expression, clustering, or classification. In theory, imputing of the missing values is not needed provided the downstream product can be produced. In this case, the downstream process is clustering as the desired outcome is to identify what gene disruption lines behave similarly. Typically, clustering uses a correlation matrix. A correlation matrix is a matrix where the rows and columns represent a gene and the value

represents correlation between the two genes. These values will range from -1 (oppositely correlated) to +1 (perfectly correlated). A weighted adjacency matrix is a means of representing a graph (network) where the edges connecting the gene-nodes represent the values in the matrix. If one were to build a graph based on a correlation matrix, then those edges would be the correlation coefficient. Thus, a weighted adjacency matrix would enable the desired downstream analyses.

Because many correlations calculations (such as Pearson's product-moment correlation or Spearman's rank correlation) require a complete data set, one must either impute the missing values or use pair-wise complete observations. There are other methods besides correlation that describe a relationship between two sets of observations, such as similarity and distance measures. For numerical data these can range from a simple calculation of distance between the two sets of observations (for example, Euclidian distance), to calculation of the angle between the two vectors (cosine similarity). Because this type of measure makes direct comparisons between sets of values, it might be more amenable to control for the missing data without disregarding it completely. As long as the output is still between -1 and 1, it is in the same numerical range as a correlation coefficient. This implies that the value can be used to make an adjacency matrix and comparable to other omics data in a similar format. Thus a method is needed that would calculate the similarity between two observations, is tolerant to missing values, and takes into account instances where one individual has an observation while the other does not.

Moving from screening data to hypothesis generation

The main objective for this work is to provide a model for functional gene annotation. Methods for normalization and similarity calculations can be used to generate communities of individuals with shared phenotypes. The Chloroplast 2010 dataset can be used as a test case because of the diversity of phenotypic information. The definition of community here refers to a group of individuals that are more similar to each other, across the community, than they are to individuals outside of the community. Because the dataset includes individuals with some annotation, enrichment calculations such as gene ontology (Ashburner et al., 2000) enrichment, can be used to develop the hypotheses. Furthermore, as all communities are driven by a phenotypic signature or a set pattern of phenotypes, this information can be used in characterizing the insertion line. As the data is from a high-throughput screen, it is possible to have high levels of false positives (responses that appear significantly different but are not). Being able to compare the phenotypic signature of known genes in the cluster to literature-established phenotypes can provide an additional check. Additionally, the phenotypes may be attributable to a second insertion or mutation other than the gene initially thought to be disrupted (Ajjawi et al., 2010). Alternatively, these observed phenotypes, while they may not be previously published, could lead to novel discoveries of the role of the characterized genes (Lu et al., 2008).

The work presented in this dissertation aims at leveraging high-throughput data to understand the biological system. Chapter 2 presents MIPHENO, an open-source R (R Development Core Team, 2011) package normalization method for high-throughput screening data. This package includes a workflow enabling researchers to take advantage

of high throughput data to determine what individuals may be responsive to a treatment. It was used here to transform the data gathered in the Chloroplast 2010 project into something that could be analyzed on a cross-dataset basis. Chapter 3 introduces the R package NetComp and the SimMeasure algorithm. SimMeasure calculates the weighted adjacency matrix, tolerates missing data, and facilitates using thresholds to remove the impact of background responses in calculating the similarity between individuals. Other features of the NetComp package are aimed at facilitating network comparison such as intersections and unions, desirable to those seeking to integrate different omics datasets. Chapter 4 presents results from the analysis of the Chloroplast 2010 data. The final chapter discusses further directions for the research.

REFERENCES

References

Aittokallio, T. (2010). Dealing with missing values in large-scale studies: microarray data imputation and beyond. Briefings in Bioinformatics *11*, 253-264.

Ajjawi, I., Lu, Y., Savage, L.J., Bell, S.M., and Last, R.L. (2010). Large-scale reverse genetics in Arabidopsis: case studies from the Chloroplast 2010 Project. Plant Physiology *152*, 529-540.

Alonso, J.M., Stepanova, A.N., Leisse, T.J., Kim, C.J., Chen, H., Shinn, P., Stevenson, D.K., Zimmerman, J., Barajas, P., Cheuk, R., et al. (2003). Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. Science *301*, 653-657.

Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J., et al. (2000). Gene Ontology: tool for the unification of biology. Nature Genetics *25*, 25-29.

Ballman, K.V., Grill, D.E., Oberg, A.L., and Therneau, T.M. (2004). Faster cyclic loess: normalizing RNA arrays via linear models. Bioinformatics *20*, 2778-2786.

Barbaric, I., Miller, G., and Dear, T.N. (2007). Appearances can be deceiving: phenotypes of knockout mice. Briefings in Functional Genomics and Proteomics *6*, 91-103 .

Berardini, T.Z., Mundodi, S., Reiser, L., Huala, E., Garcia-Hernandez, M., Zhang, P., Mueller, L.A., Yoon, J., Doyle, A., Lander, G., et al. (2004). Functional annotation of the Arabidopsis genome using controlled vocabularies. Plant Physiology *135*, 745-755.

Bouché, N., and Bouchez, D. (2001). Arabidopsis gene knockout: phenotypes wanted. Current Opinion in Plant Biology *4*, 111-117.

Eckel, J.E., Gennings, C., Therneau, T.M., Burgoon, L.D., Boverhof, D.R., and Zacharewski, T.R. (2005). Normalization of two-channel microarray experiments: a semiparametric approach. Bioinformatics *21*, 1078-1083.

Ferri, C., Hernandez-Orallo, J., and Modroiu, R. (2009). An experimental comparison of performance measures for classification. Pattern Recognition Letters *30*, 27-38.

Fiehn, O., Kopka, J., Dörmann, P., Altmann, T., Trethewey, R.N., and Willmitzer, L. (2000). Metabolite profiling for plant functional genomics. Nature Biotechnology *18*, 1157-1161.

Furnham, N., Garavelli, J.S., Apweiler, R., and Thornton, J.M. (2009). Missing in action: enzyme functional annotations in biological databases. Nature Chemical Biology *5*, 521-525.

Horan, K., Jang, C., Bailey-Serres, J., Mittler, R., Shelton, C., Harper, J.F., Zhu, J.-K., Cushman, J.C., Gollery, M., and Girke, T. (2008). Annotating genes of known and unknown function by large-scale coexpression analysis. Plant Physiology *147*, 41-57.

Jander, G., Norris, S.R., Joshi, V., Fraga, M., Rugg, A., Yu, S., Li, L., and Last, R.L. (2004). Application of a high-throughput HPLC-MS/MS assay to Arabidopsis mutant screening; evidence that threonine aldolase plays a role in seed nutritional quality. The Plant Journal *39*, 465-475.

Last, R.L., Jones, A.D., and Shachar-Hill, Y. (2007). Towards the plant metabolome and beyond. Nature Reviews Molecular Cell Biology *8*, 167-174.

Liew, A.W.-C., Law, N.-F., and Yan, H. (2011). Missing value imputation for gene expression data: computational techniques to recover missing data from available information. Briefings in Bioinformatics *12*, 498-513.

Lu, Y., Savage, L.J., Ajjawi, I., Imre, K.M., Yoder, D.W., Benning, C., DellaPenna, D., Ohlrogge, J.B., Osteryoung, K.W., Weber, A.P., et al. (2008). New connections across pathways and cellular processes: industrialized mutant screening reveals Novel associations between diverse phenotypes in Arabidopsis. Plant Physiology *146*, 1482-1500.

Lu, Y., Savage, L.J., Larson, M.D., Wilkerson, C.G., and Last, R.L. (2011a). Chloroplast 2010: A Database for large-scale phenotypic screening of Arabidopsis mutants. Plant Physiology *155*, 1589-1600.

Lu, Y., Savage, L.J., and Last, R.L. (2011b). Chloroplast phenomics: systematic phenotypic screening of chloroplast protein mutants in Arabidopsis. In Chloroplast Research in Arabidopsis : Methods and Protocols, Volume II, R.P. Jarvis, ed. (NY: Humana Press), pp. 161-185.

Mar, J.C., Kimura, Y., Schroder, K., Irvine, K.M., Hayashizaki, Y., Suzuki, H., Hume, D., and Quackenbush, J. (2009). Data-driven normalization strategies for high-throughput quantitative RT-PCR. BMC Bioinformatics *10*, 110.

Miron, M., and Nadon, R. (2006). Inferential literacy for experimental high-throughput biology. Trends in Genetics *22*, 84-89.

Mueller, L.A., Zhang, P., and Rhee, S.Y. (2003). AraCyc: a biochemical pathway database for Arabidopsis. Plant Physiology *132*, 453-460.

Oba, S., Sato, M.-a., Takemasa, I., Monden, M., Matsubara, K.-i., and Ishii, S. (2003). A Bayesian missing value estimation method for gene expression profile data. Bioinformatics *19*, 2088-2096.

Ober, D. (2010). Gene duplications and the time thereafter - examples from plant secondary metabolism. Plant Biology *12*, 570-577.

Quackenbush, J. (2002). Microarray data normalization and transformation. Nature Genetics *32*, 496-501.

R Development Core Team (2011). R: A Language and Environment for Statistical Computing (http://www.R-project.org: R Foundation for Statistical Computing).

Rhee, S.Y., Beavis, W., Berardini, T.Z., Chen, G., Dixon, D., Doyle, A., Garcia-Hernandez, M., Huala, E., Lander, G., Montoya, M., et al. (2003). The Arabidopsis information resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. Nucleic Acids Research *31*, 224-228.

Rocke, D.M. (2004). Design and analysis of experiments with high throughput biological assay data. Seminars in Cell & Developmental Biology *15*, 703-713.

Schlomer, G.L., Bauman, S., and Card, N.A. (2010). Best practices for missing data management in counseling psychology. Journal of Counseling Psychology *57*, 1-10.

Stacklies, W., Redestig, H., Scholz, M., Walther, D., and Selbig, J. (2007). pcaMethods—a bioconductor package providing PCA methods for incomplete data. Bioinformatics *23*, 1164-1167.

The Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature (London) *408*, 796-815.

Thorneycroft, D., Sherson, S.M., and Smith, S.M. (2001). Using gene knockouts to investigate plant metabolism. Journal of Experimental Botany *52*, 1593-1601.

Van Eenennaam, A.L., Lincoln, K., Durrett, T.P., Valentin, H.E., Shewmaker, C.K., Thorne, G.M., Jiang, J., Baszis, S.R., Levering, C.K., Aasen, E.D., et al. (2003). Engineering vitamin E content: from Arabidopsis mutant to soy oil. Plant Cell *15*, 3007-3019.

Wang, C., Marshall, A., Zhang, D., and Wilson, Z.A. (2012). ANAP: an integrated knowledge base for Arabidopsis protein interaction network analysis. Plant Physiology *158*, 1523-1533.

Yang, Y.H., Dudoit, S., Luu, P., Lin, D.M., Peng, V., Ngai, J., and Speed, T.P. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Research *30*, e15.

Yuan, J.S., Galbraith, D.W., Dai, S.Y., Griffin, P., and Stewart, C.N. (2008). Plant systems biology comes of age. Trends in Plant Science *13*, 165-171.

Chapter 2

MIPHENO: data normalization for high throughput metabolite analysis

MIPHENO: data normalization for high throughput metabolite analysis

Bell, Shannon M, Burgoon, Lyle D, Last, Robert L. (2012). MIPHENO: data normalization for high throughput metabolic analysis. BMC Bioinformatics *13*, 10.

For those conducting *post hoc* data analysis (i.e., analyses of data after the experiment has been carried out, which were not specified *a priori*), there are few tools for normalizing data if the experiment was not conducted with the standard controls and replication most methods require. Controls and replication are often limited or omitted entirely from screening studies as the goal is to maximize the number of individuals screened, and there is typically an anticipation of very few individuals showing the attribute of interest. With screening studies it is expected that a follow-up will be carried out on individuals that show an interesting response. The goal is to identify and prioritize, versus quantify how different an individual is from the control.

In large-scale studies there are often multiple factors at play including the time over which the data were collected, how the samples were grouped, and who conducted the analyses. Furthermore, metadata describing the experimental details is typically limited so one may not have knowledge of how to model these issues appropriately. MIPHENO was developed to address these aspects of high throughput screening studies. Designed for use in high throughput screens, it uses the principle that the majority of the signals will be within the background range, and conducts normalization to remove technical variance based on scaling to a global median. The developed software package and workflow includes a quality control measure, which is important for removing groups that appear to

be behaving differently than others and could bias the normalization. There is also a method for identifying individuals that are likely to exhibit a response, useful for prioritizing individuals for follow-up.

While MIPHENO is an admittedly simplistic approach, it is demonstrated to outperform the standard approach of looking at individuals within a sample group. Additional features of the software package address several needs in the analysis of screening data. It can be used to go from the raw data to a prioritized list of candidates for follow-up or onto other analyses such as clustering. The real significance of this method is that it facilitates the use of poorly designed experiments and enables comparisons to be made over the course of a multi-year experiment.

Abstract

2.1.1 Background

High throughput methodologies such as microarrays, mass spectrometry and plate-based small molecule screens are increasingly used to facilitate discoveries from gene function to drug candidate identification. These large-scale experiments are typically carried out over the course of months and years, often without the controls needed to compare directly across the dataset. Few methods are available to facilitate comparisons of high throughput metabolic data generated in batches where explicit in-group controls for normalization are lacking.

2.1.2 Results

Here we describe MIPHENO (Mutant Identification by Probabilistic High throughput-Enabled Normalization), an approach for post-hoc normalization of quantitative first-pass screening data in the absence of explicit in-group controls. This approach includes a quality control step and facilitates cross-experiment comparisons that decrease the false non-discovery rates, while maintaining the high accuracy needed to limit false positives in first-pass screening. Results from simulation show an improvement in both accuracy and false non-discovery rate over a range of population parameters ($p < 2.2 \times 10^{-16}$) and a modest but significant ($p < 2.2 \times 10^{-16}$) improvement in area under the receiver operator characteristic curve of 0.955 for MIPHENO vs 0.923 for a group-based statistic (*z*-score). Analysis of the high throughput phenotypic data from the Arabidopsis Chloroplast 2010 Project (http://www.plastid.msu.edu/) showed ~ 4-fold increase in the

ability to detect previously described or expected phenotypes over the group based statistic.

2.1.3 Conclusions

Results demonstrate MIPHENO offers substantial benefit in improving the ability to detect putative mutant phenotypes from post-hoc analysis of large data sets. Additionally, it facilitates data interpretation and permits cross-dataset comparison where group-based controls are missing. MIPHENO is applicable to a wide range of high throughput screenings and the code is freely available through an R package in CRAN (http://cran.r-project.org/web/packages/MIPHENO/index.html).

Background

High-throughput screening studies in biology and other fields are increasingly popular due to ease of sample tracking and decreasing technology costs. These experimental setups enable researchers to obtain numerous measurements across multiple individuals in parallel (e.g. gene expression and diverse plate-based assays) or in series (e.g. metabolomics and proteomics platforms). The large number of measurements collected often comes at the cost of measurement precision or the overall power of detection. For many large-scale studies, the experimental design aims to maximize the number of compounds or individuals tested, resulting in limited replication and few to no controls. In the case of microarray studies, several methods for normalizing arrays have been developed (Ballman et al., 2004; Eckel et al., 2005; Quackenbush, 2002) with no universal method adopted as the standard. Quantitative PCR faces the same issues as it is

used more frequently in high throughput platforms, with analysis methodologies being developed paralleling those for expression arrays (Mar et al., 2009).

Metabolite profiling is a rapidly expanding area of high throughput measurements, where samples having large amounts of biological variability and diverse physical properties makes quantification of large numbers of structurally diverse metabolites challenging (Last et al., 2007). Few strategies exist for normalization in metabolite analysis to control for run-to-run variance other than to include negative and positive controls. For large-scale screens involving mutagenized populations (plant, bacteria) or crosses (plant breeding), the goal is to identify putative hits, or individuals that are likely to be different from the bulk of the samples for subsequent follow-up (e.g. (Jander et al., 2004)). In these conditions, properties of the sample cohort serve as controls with the measure of differences between an individual and its cohort used to identify samples differentially accumulating a metabolite (Jander et al., 2004). This strategy can streamline sample processing and maximize throughput when the expected effects are large and easily observable.

For studies where comparisons are sought across an experiment conducted over the course of several months or in different sample batches, normalizing factors are necessary, especially given typically high levels of biological and technical variability (Fiehn et al., 2000; Miron and Nadon, 2006; Rocke, 2004). Ideal experiments include technical and biological replication within each set as well as controls facilitating comparisons between sample batches, but these are often limited or omitted entirely due to likely increases in experimental costs or the negative impacts on throughput. However, absence of these

experimental controls limits the ability to handle variability between sample groups (e.g. remove batch effects) making it a greater challenge to identify individuals within the range between normal and aberrant phenotypes. Without the ability to normalize the data provided by experimental controls, some of the benefits of high throughput screens are lost, yet the desire to maximize throughput places constraints on the experimental design.

The motivation for algorithm development came from the *Arabidopsis thaliana* Chloroplast 2010 Project large-scale reverse genetic phenotypic screen [Chloroplast 2010, http://www.plastid.msu.edu/, (Ajjawi et al., 2010; Lu et al., 2008; Lu et al., 2011b; Lu et al., 2011c)]. This project leverages the collection of T-DNA insertion lines and genomic sequence for the plant model species *A. thaliana* to screen large numbers of putative gene knockouts with the aim of functionally characterizing chloroplast-targeted genes. The presence of a large T-DNA insertion can block or reduce expression of the gene it lands in, and altered phenotypes can provide insights into the normal function of the gene and its protein or RNA product(s).

In addition to qualitative and semi-quantitative measures of physiological and morphological characteristics, the levels of leaf fatty acids and leaf and seed free amino acids, important outputs of chloroplast metabolism. The pipeline assays were performed on groups of individual plants planted in units of up to thirty-two per tray and three trays of plants per assay group. Two assay groups were grown concurrently under controlled environment plant growth conditions. Individuals representing T-DNA insertion events in different locations within the same gene (alleles) are present in the dataset, and it is of interest to compare the assay responses of these individuals as well as to identify other

individuals with similar responses. Because the experimental design lacked cross-group controls (e.g. designated WT), the ability to make even semi-quantitative cross-dataset comparisons was not possible using existing methodology.

Developing phenotypic annotation for un- and under- annotated genes is a primary goal for the Chloroplast 2010 project and identification of individuals with like phenotypes (phenotypic clustering) is a way to achieve that goal. Thus, a method that would allow cross-dataset comparisons and identify putative mutants was needed to achieve the goal. The resulting method, MIPHENO (Mutant Identification by Probabilistic High throughput-Enabled Normalization), is aimed at improving first-pass screening capabilities for large datasets in the absence of defined controls. Algorithm performance was tested using a synthetic data set and the Chloroplast 2010 high throughput phenotypic dataset. The executable code and data for the Chloroplast 2010 analysis are available as a CRAN package (MIPHENO, http://cran.r-project.org/web/packages/MIPHENO/index.html).

The following describes a quality control process for identifying aberrant groups followed by a data normalization method, which aims to bring samples into the same distribution allowing for dataset-wide comparisons. Additionally, we describe a hit detection function based on the cumulative distribution function (CDF) to identify samples with putative, 'non-normal' phenotypes. For clarity, the terms normal and wild type (WT) are used to describe the typical response of the population. Generally, this could be the untreated (chemically or genetically) population or the base level of the system (e.g. background response). Non-wild type responses, a hit or mutant, refer to a response that is distinct from the normal response distribution, with a putative hit/putative mutant

referring to a sample that is predicted to have a response different from the normal response distribution but has yet been confirmed. In high throughput screens, the objective is to identify putative hits balancing the false positive rate (FPR), or the number of WT samples that are called hits, with the false non-discovery rate (FNDR), the number of true hits that are missed. Results are presented from analysis of the synthetic dataset and biological data.

Results

2.3.1 Input data characteristics and structure

MIPHENO is specifically designed for the analysis of first pass screening data where the majority of measured responses are from the WT or normal class and the number of responses not in this group (putative hits) is quite small. Examples of experiments yielding appropriate data are non-targeted protein binding/activator assays, reporter gene assays, or population screens, where there are either no defined classes or very unbalanced classes such that a large majority of responses fall in the WT class. Data coming from a treatment vs control experiment would not meet the criteria if there were large numbers of 'non-WT' responses expected. Additionally, the approach is tolerant to repetition of both individual samples and sample groups across the course of the experiment so long as the portion of individuals showing a WT response in any sample group is over 50%. As the portion of WT individuals in a sample group decreases, there will be a reduction in accuracy and a corresponding increase in false non-discovery rate (FNDR) due to the assumptions of the algorithm, as demonstrated in the Testing section below. Additionally, while some measured responses may not be independent (ex, metabolite measures of branch chain

amino acids), the method treats these attributes (e.g., metabolites) as independent to increase the flexibility of the analysis. For instance, the results for attribute 1 (including normalization and downstream analyses) do not impact the results for attribute 2. This is beneficial in post hoc analysis where the individual performing the analysis has limited knowledge of the relationship between measures.

Input data for analysis by MIPHENO assumes that multiple attributes are measured for each individual. The data structure treats each row as an individual sample, whose relationship to other samples can be described by one or multiple factor variables represented in columns (grouping factor). For example, the assay group representing the identification number for a 96-well plate containing up to 96 individuals. Subsequent columns describing the response of the individual to some assay (attribute response) are quantitative, continuous values. Information must be present that enables association of a grouping factor to the attribute responses, but a single data object may include the responses for different attributes as long as the appropriate grouping factor is present. For example, a 'LC_ID' column might provide the grouping factor for ten columns of LC-MS amino acid data, while 'HPLC_ID' might provide the grouping factor for five columns of HPLC-derived responses on the same set of samples. This structure is aimed at simplifying situations where multiple measurements are taken on the same individual.

2.3.2 Algorithm

MIPHENO is based on invariant set normalization with three key assumptions made of the input data. The first is that samples from the same genetic background should have a similar assay response over time. This implies that, given a population P, the distribution of

an observed response r from sample set p in set P should have the same distribution as the response R from population P as p approaches P. Following this logic, the second assumption of the data is that the observed differences between the distributions r and R are due to technical error as opposed to biological or genetic variance as p approaches P. The last assumption is that there will be limited observable effects of simple genetic manipulations to an organism for any random gene. This is based on empirical evidence from years of published studies (Barbaric et al., 2007; Bouché and Bouchez, 2001; Jander et al., 2004; Van Eenennaam et al., 2003). Specifically, due to genetic redundancy and metabolic flexibility, a given disruption in gene function will likely cause a response outside the WT distribution in only a limited number of measured responses.

These assumptions are similar to those for microarray analysis, specifically that for a random or large grouping of individuals (e.g. cDNAs), changes will be observed for a relatively small proportion (Yang et al., 2002). Other assumptions used to normalize the data (e.g. a balance in the total amount of transcript in quantile normalization (Quackenbush, 2002)) have the same effect of forcing the median value of a sample set across several experiments or arrays to be equal. Similar assumptions also apply to data from other high throughput screens, e.g. reporter gene-based assays and enzymatic assays.

An overview of the algorithm is presented in Figure 1. The algorithm requires that input data have a grouping factor that presents a batch or process group on which the normalization steps can be performed (see “Input Data Structure and Characteristics” above). If multiple grouping factors are present (e.g. different sample collection, processing, and analysis dates) it is recommended to use the factor representing the

highest level of technical (i.e. non-biological) error for normalization. This can be determined by familiarity with the methodology or by checking the grouping factors to see which factor has the largest interquartile range for group medians.

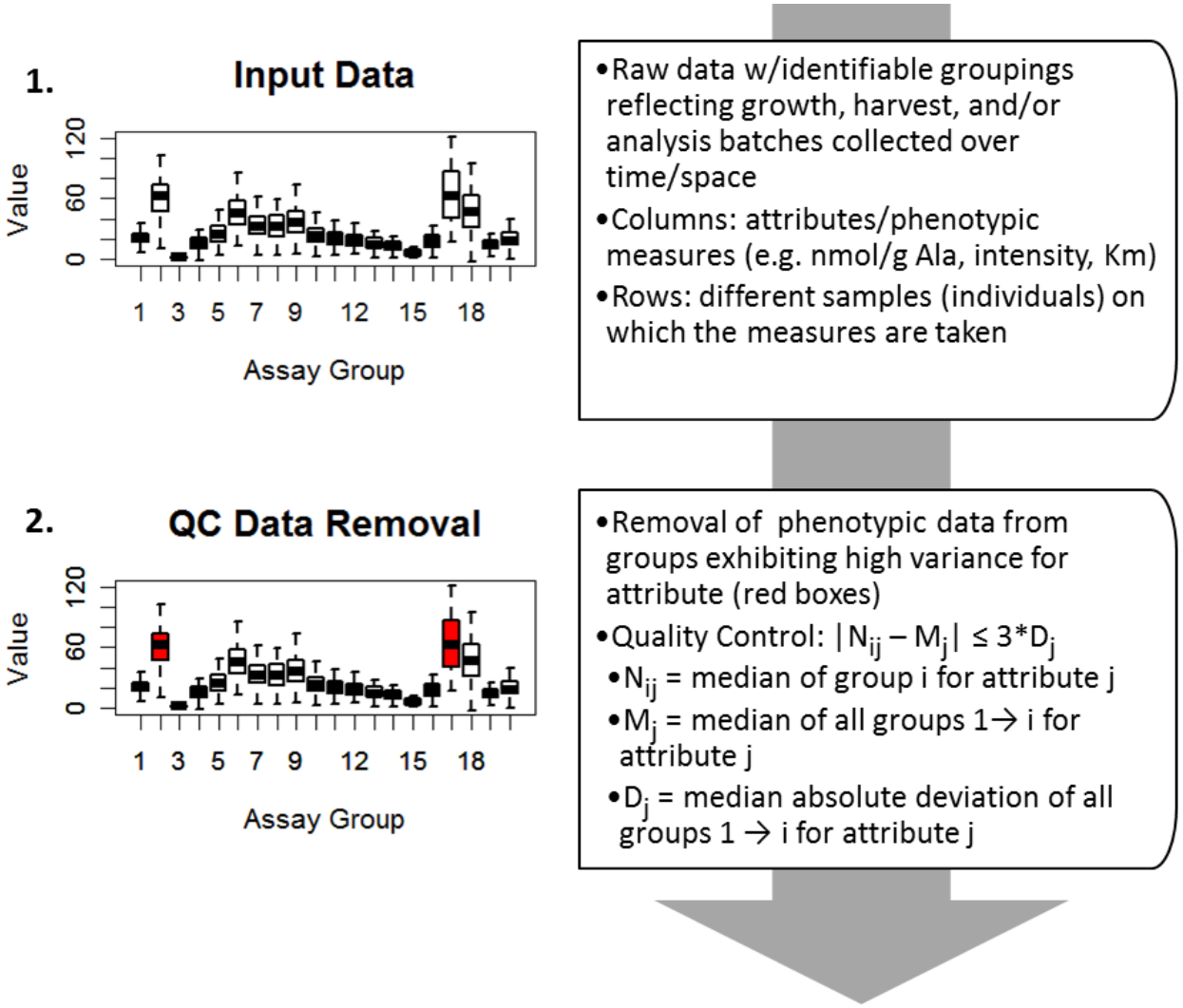
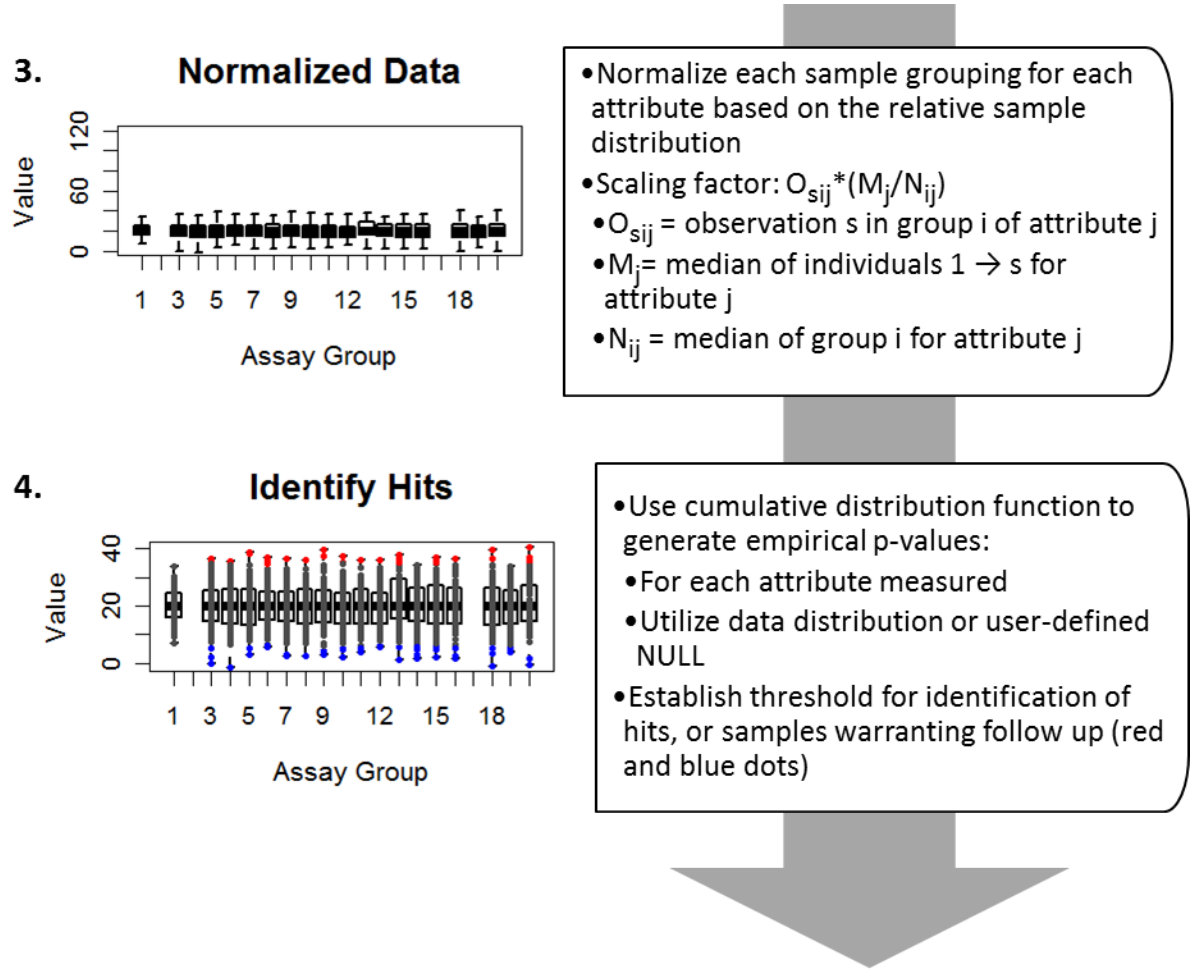


Figure 1 Flowchart of MIPHENO

"Input Data" (1) contains data with identifiable parameters for grouping/processing the data. The data pass through a quality control (QC) removal step (2), where groups not meeting the cut offs are identified and removed on an attribute-by-attribute basis. Data are normalized (3) using a scaling factor based on the data distribution. Putative hits are identified (4) using a CDF built from the data or user defined NULL distribution and an empirical p-value is assigned to each observation. Thresholds can be established based on follow-up capacity and prior knowledge (e.g. ability to detect known 'gold standard' mutant samples). For interpretation of the references to color in this and all other figures, the reader is referred to the electronic version of this dissertation.

Figure 1 (cont'd)



2.3.3 Quality control method

In performing post-hoc data analysis it is often unknown if on-line quality control (QC) was conducted or where process changes occurred that could negatively affect the outcome of analysis. To address these issues, a quality control (QC) step prior to analysis was included to identify samples with a high likelihood of assay or group-specific process error. Examples of sources of these types of error include instrument malfunction (for assay-specific error), abnormalities in growth or preparation of material (group-specific error), or improper sample handling affecting a group of samples exposed to the same conditions rather than an individual response. If an on-line QC step was already used to filter the dataset this step can be omitted. Thresholds for QC are determined from the overall distribution of the collected data with a user-defined cut off; for example groups with group median > 3 median adjusted deviations (MAD) from the global median. The amount of data removed will depend on the cut off used and the data distribution. A visual inspection of the data using box and whisker plots is advised to check the data for clear signs of drift or likely changes in protocol that may require manual QC. Examples would be group medians steadily increasing or decreasing across dataset or a switch to a new average median response corresponding with sample order, respectively. For post hoc analysis on datasets where the order in which samples were assayed or collected is unknown, it may be advisable to use a cut off of 3 MAD to permit more data passing on to the next stage.

Data quality is assessed on an attribute-by-attribute basis with the assumption that the measured traits are independent; with an attribute being any measured or observed

response. Thus, if multiple attributes are measured for a group (for example, numerous metabolites or promoter-reporter gene outputs), only attribute data for the trait that shows high deviation would be removed and the rest of the data for the group retained. For example, 'HPLC_ID' is the grouping factor for the response of metabolites, such as amino acids. The overall response distribution of each metabolite is assumed to be independent of the other metabolites; thus if the measured response of alanine is 10x the response of proline it will not impact the QC step (or subsequent steps). If the median response for alanine in HPLC_ID = 1 is greater than the QC cut off, all responses for alanine in HPLC_ID = 1 are removed but the other measured responses for HPLC_ID = 1 are retained, provided they too pass QC. While this does not control for drift, it provides a facile QC step for post-hoc data analysis where the order of data generation is unknown.

2.3.4 Normalization

The normalization process is done on an attribute-by-attribute basis using a user-defined grouping. A grouping factor should encompass the highest amount of non-biological variation and may be the same factor used in the QC step, but should include as many individuals as possible (e.g. n>10). A scaling factor is calculated to bring the median of each group to the global median, similar to invariant set normalization (Mar et al., 2009). The key difference from invariant set or quantile strategies is that just the median value is used, not an explicit individual or multiple quantiles to take into account lack of replication between groups and limited sample size. It is important that groupings represent a selection of individuals where the frequency of non-WT behaviors approaches that of the

overall population to avoid bias in cases when a particular group is enriched with non-WT behaviors for a given attribute.

2.3.5 Testing

To gauge the performance of the approach, a synthetic dataset was generated emulating characteristics of actual data (see Methods). This dataset was used initially since the true properties of the individuals could be known, allowing for observation classification (e.g. WT and mutant) and to evaluate the effect of population distribution on the performance of the method. Figure 2 illustrates the population distributions used to test the performance of MIPHENO.

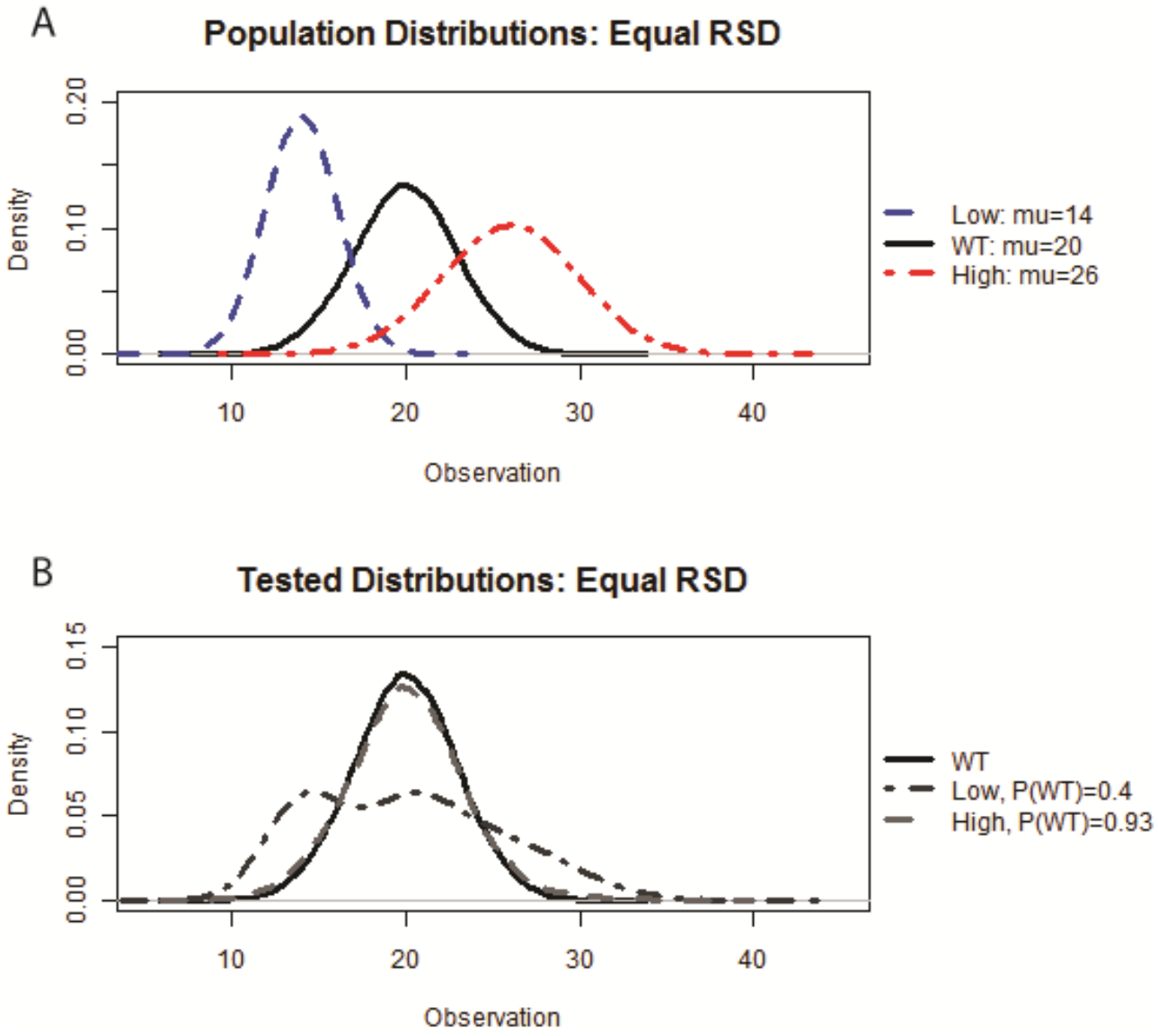
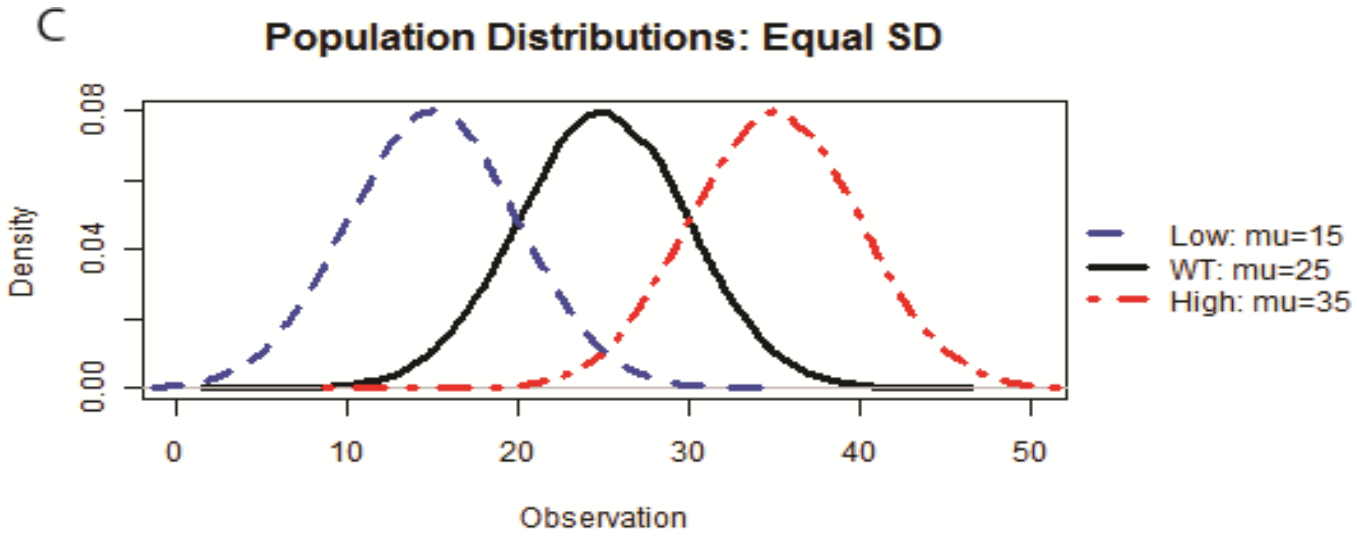
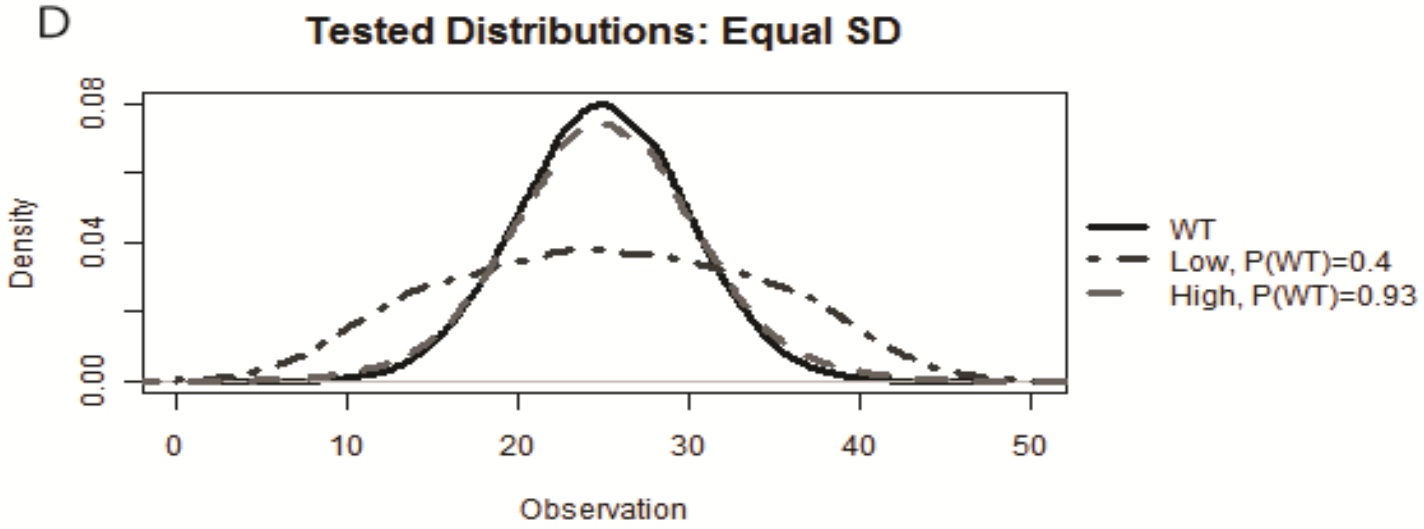


Figure 2 Synthetic Populations Used in Testing

Synthetic data were generated to measure the performance of the three different methods in a case where 'ground truth' is known. Samples were randomly drawn from a low abundance population (Low, blue line), high abundance population (High, red line) or a WT population (WT, black line) as shown in the upper panels (A, C). Two population structures were sampled, one with a low probability of WT, P(WT=0.4), and the other with a high probability of WT, P(WT)=0.93, shown in the lower panels (B, C). To test the effect of population shape, equal relative standard deviation (RSD=15%, A and B) or equal standard deviation (SD =5, C and D) were independently tested.

Figure 2 (cont'd)





Comparison of two different data analysis approaches was used to test 1) if pre-processing steps remove high amounts of real biological variation indicative of a putative hit and 2) whether an increased false non-discovery rate (FNDR) resulted from using MIPHENO verses a sample-group based method (results in Figures 3, 4, and 5). The first approach referred to as 'Raw', uses the raw, unprocessed data, but followed the same process as in MIPHENO to identify putative mutants. Differences between Raw and MIPHENO aid in illuminating the effectiveness of pre-processing in noise removal. The second approach, referred to as 'Z', also utilized the raw data but used a MAD score on a sample-group basis to identify putative mutants as described for the Chloroplast 2010 data (Lu et al., 2008). Comparison of MIPHENO to Z aids in determining potential loss of information due to normalizing across the data sets (e.g. whether true mutants were more severely scaled in normalization), or if the group-based error was controlled for without negatively impacting hit detection. In a review of performance metrics by Ferri et al., 2009, accuracy (ACC) was found to be a better metric than area under the receiver-operating curve (AUC) in the case of unbalanced sample size as well as misclassification noise, which are both properties of the data under analysis. Conversely, they found AUC outperformed ACC in probability and, to a lesser degree, ranking noise. False non-discovery rate is an important metric when considering first-pass screens as one seeks to limit the true positives missed, which is the situation described here.

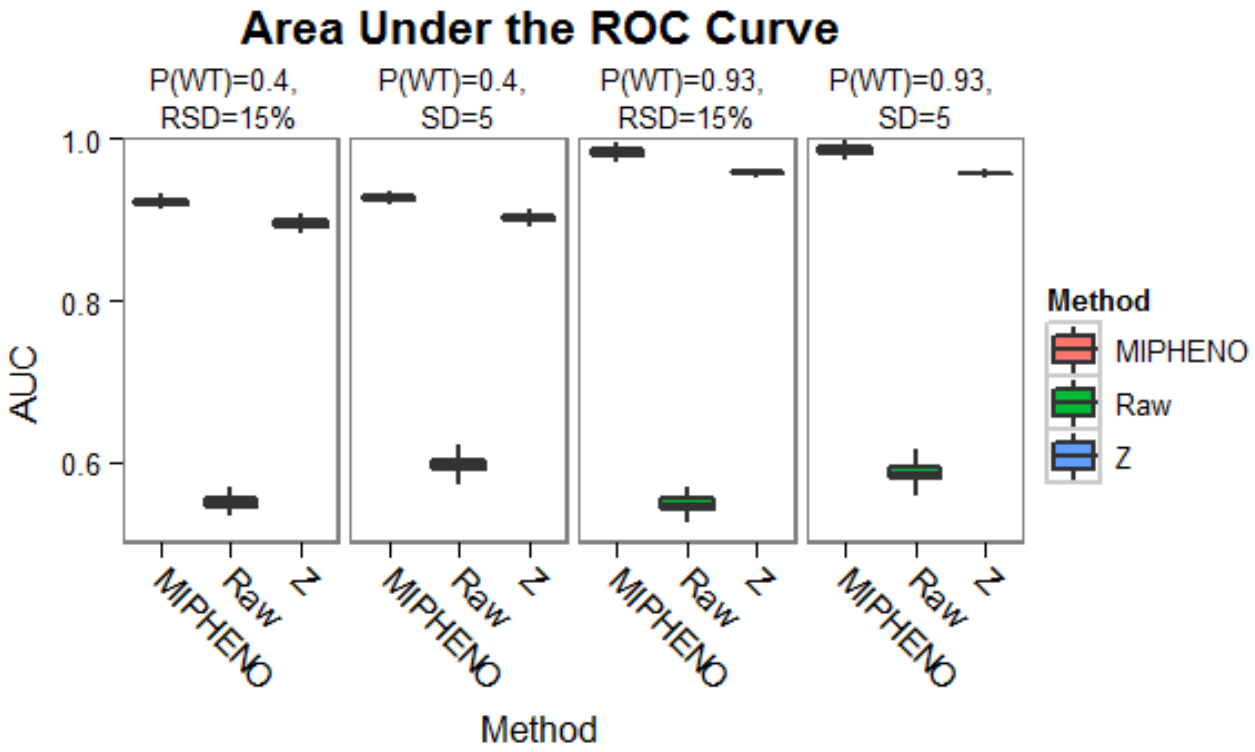


Figure 3 Performance of Methods on Synthetic Data: AUC

The AUC was used to evaluate classification performance of MIPHENO, the use of raw data followed by a CDF classifier (RAW), and a group-based metric (Z) on synthetic data described in Figure 2. MIPHENO (pink, first in set) outperforms both RAW (green, middle) and Z (blue, left in set) across the different population parameters.

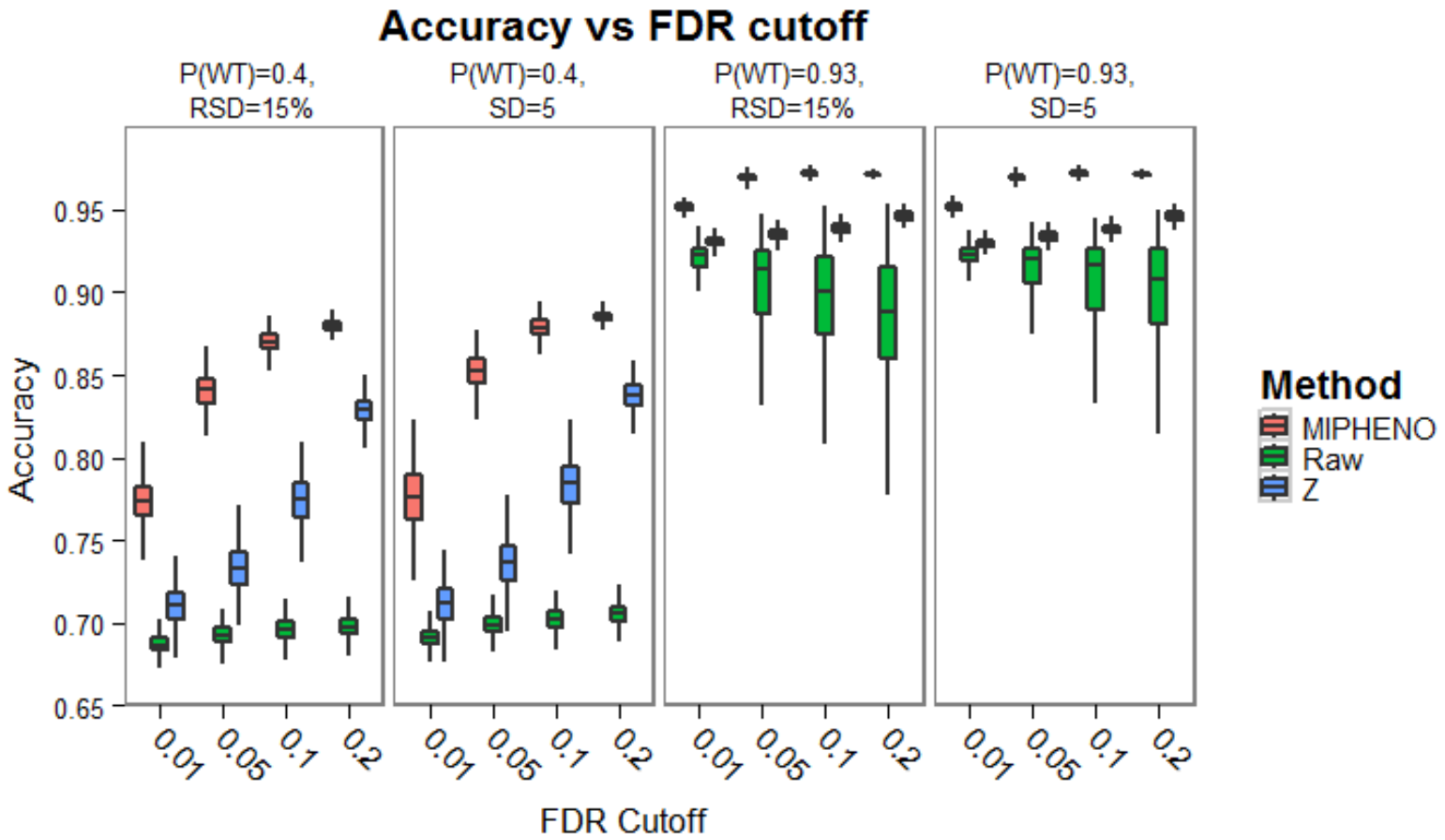


Figure 4 Performance of Methods on Synthetic Data: Accuracy

Accuracy of classification was used to compare the performance of MIPHENO, the use of raw data followed by a CDF classifier (RAW), and a group-based metric (Z) on synthetic data from populations described in Figure 2. The percent accuracy is plotted along the y-axis while the false discovery rate (FDR) cut off is along the x-axis. Each population distribution tested is shown in a separate panel. Note that MIPHENO (pink) achieved higher classification than Z (blue) ($p < 2.2e-15$, Wilcoxon sign rank) and both methods outperformed Raw (green) independent of the population parameters tested.

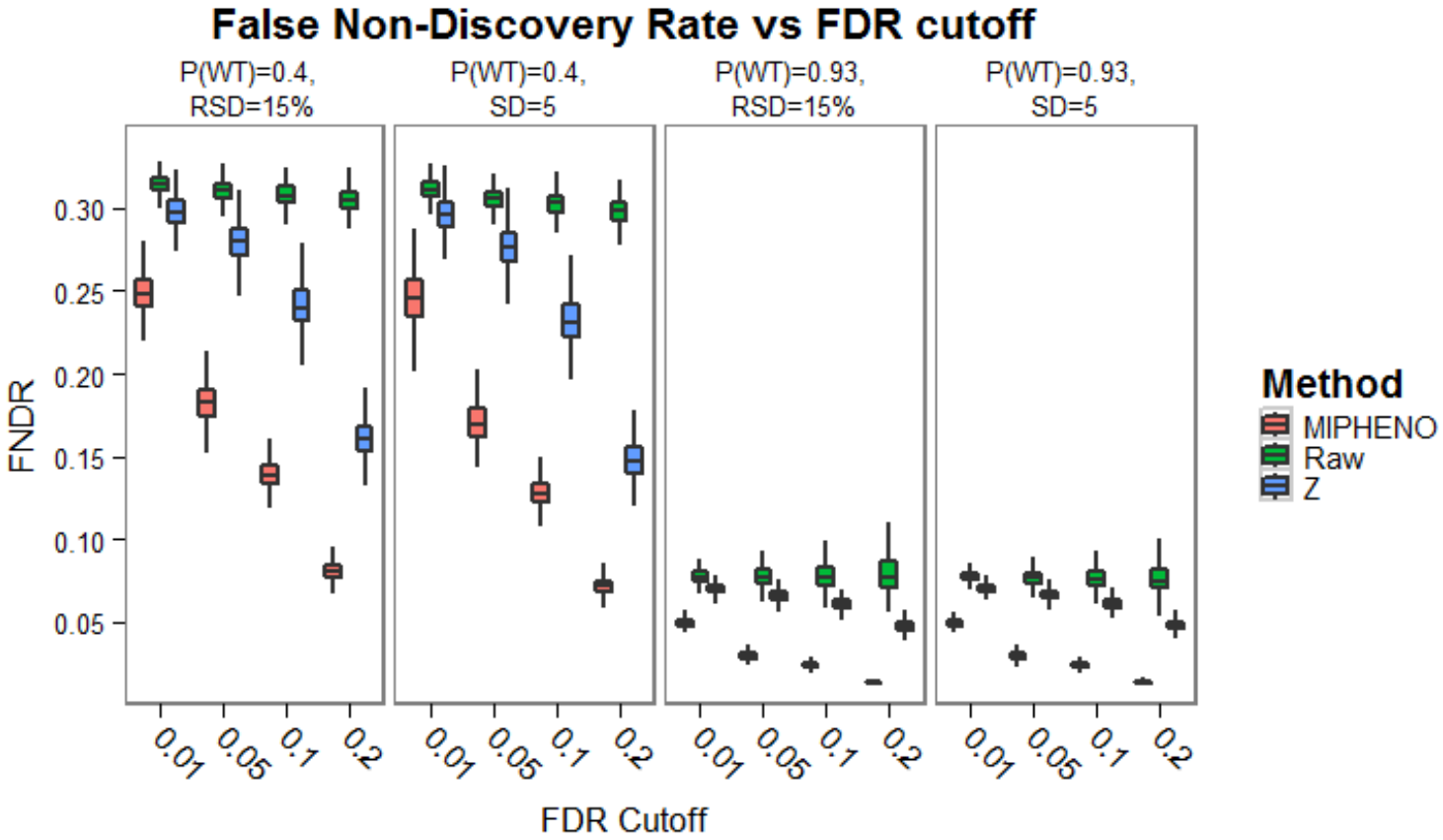


Figure 5 Performance of Methods on Synthetic Data: False Non-Discovery Rate

The false non-discovery rate (percent positive hits missed) was used to compare the performance of MIPHENO, the use of raw data followed by a CDF classifier (RAW), and a group-based metric (Z) on synthetic data from populations described in Figure 2. The FNDR is plotted along the y-axis with the different false discovery rate (FDR) cut offs along the x-axis. Each population distribution is shown in a different panel. Note that across all populations tested, MIPHENO has a lower FNDR than the other two method, suggesting that fewer putative hits missed with MIPHENO compared to Z-score (blue) or raw data (green).

Results of the performance trials using a combination of two population distributions that had a high frequency of WT (P(wt) = 0.93) and low WT frequency (P(wt) = 0.40), drawn from populations of equal standard deviation (SD) or relative standard deviation (RSD) (Figure 2), are shown in Figures 3, 4, and 5. These results suggest that the proportion of true WT in the sample had little effect on the performance of the methods relative to each other, regardless of the metric used; however, the accuracy is decreased and the false non-discovery rate is increased for all methods when the portion of data from the mutant class is increased (Figures 4 and 5). MIPHENO showed a higher accuracy and lower FNDR ($p < 2.2 \times 10^{-16,}$ Wilcoxon signed rank test) across a range of FDR cut offs compared to the other methods (Figure 5). Furthermore, the AUC of both MIPHENO and Z outperformed an analysis of Raw (Figure 3), which performed just above what is expected at random, highlighting the importance of controlling for group-based variability. In summary, MIPHENO outperformed both the Raw and Z-methods across all three metrics tested.

2.3.6 Implementation

Results from the Chloroplast 2010 Project (Ajjawi et al., 2010; Lu et al., 2008) were used to test the performance of MIPHENO on experimentally generated high throughput screening data. This dataset includes results for leaf protein amino acids and fatty acid methyl esters as well as seed protein amino acids for plants run through the Chloroplast 2010 pipeline. Multiple individuals representing the same seed stock or the same gene are present in the dataset although they were not assayed in the same group. Thus, it is of interest to look at the consistency between individuals representing the same gene to

identify Leaf and seed metabolite data from mutants in the Col-0 (CS60000, (Alonso et al., 2003)) ecotype genetic background were processed using MIPHENO and *z* score methods independently. Figure 6 outlines the methods for comparison. Briefly, both MIPHENO empirical *p*-values and *z* scores were calculated for the two data measurements available in the Chloroplast 2010 dataset (mol% and nmol/gFW). The average score per T-DNA insertion line was calculated for each data type to avoid overemphasizing lines that were analyzed multiple times. Aracyc (Mueller et al., 2003) and Gene Ontology (GO) (Berardini et al., 2004) information obtained from The Arabidopsis Information Resource (TAIR) (Rhee et al., 2003) were used to generate a list of loci previously demonstrated to have a biological function in Arabidopsis. Loci with phenotypes predicted by the methods were compared to the list of literature-documented loci. The biological role and/or phenotypes of the genes were compared to the published information to determine the accuracy of the prediction. Results are given in Table 1. While both methods had a similar frequency of correctly identifying mutant phenotypes at the initial level of Z cut off of 2.5, the Z method returned fewer lines than MIPHENO. It was necessary to adjust the Z threshold to 1.3 to recover these lines, which resulted in no additional mutants but an increase in false positives. Overall, there was ~four-fold improvement in the ability to detect previously described or expected phenotypes compared with the *z*-score.

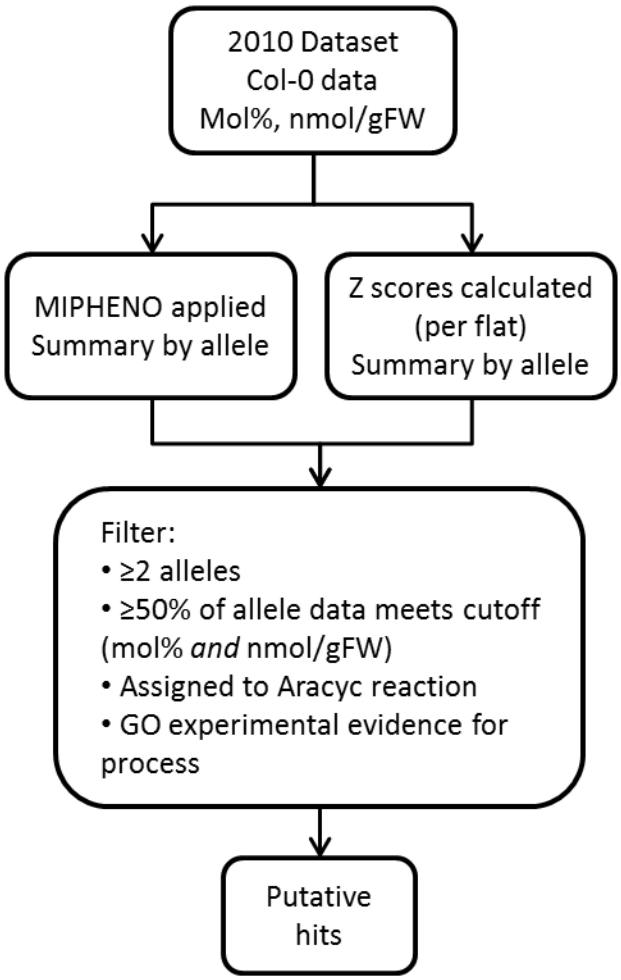


Figure 6 Flowchart of Performance Measures for Chloroplast 2010 Data

Metabolite data from wild-type Col-0 ecotype samples were taken from the Chloroplast 2010 dataset. MIPHENO empirical *p*-values and *z*-scores were calculated separately for metabolite values reported as mol % and nmol/g fresh weight (nmol/gFW) and results filtered according to criteria. Publicly available annotation (Aracyc and GO) for annotated genes provided a basis of comparison between the two metrics

Table 1 Lines identified by MIPHENO and Z methods

Locus	Description	Tissue	MIPHENO Cutoff= 0.1	Zscore Cutoff = 2.5	Zscore Cutoff = 1.3
At1g08250	ADT6: Plastid-localized arogenate dehydratase	Seed	**High**: GLN, TYR		**High**: GLN, TYR
		Leaf			
At1g09795	ATATP-PRT2: ATP phosphoribosyl transferase	Seed			
		Leaf	**High**: HIS		**High**: HIS, LEU
At1g11790	ADT1: Plastid-localized arogenate dehydratase	Seed			
		Leaf	**Low**: PHE		**Low**: PHE
At1g65960	GAD2: glutamate decarboxylase	Seed			**Low**: GABA
		Leaf	**Low**: GABA	**Low**: GABA	**Low**: GABA
At2g39800	P5CS1: delta1-pyrroline-5-carboxylate synthase	Seed	**Low**: HPRO		
		Leaf	**Low**: PRO		**Low**: PRO
At3g11170	FAD7: Responsible for the synthesis of 16:3 and 18:3 fatty acids	Seed			
		Leaf	**High**: 16:1D7, 16:2, 18:1D9, 18:2; **Low**: 16:3, 18:3	**High**: 16:2, 18:1D9, 18:2; **Low**: 16:3, 18:3	**High**: 16:1D7, 16:2, 18:1D9, 18:1D11, 18:2; **Low**: 16:3, 18:3
At3g45300	IVD: Isovaleryl-CoA Dehydrogenase	Seed	**High**: ARG, GABA, HIS, ILE, LEU, MET, TRP, VAL; **Low**: GLU	**High**: ARG, GABA, HIS, ILE, LEU, TRP, VAL, MET; **Low**: GLU	**High**: N, ARG, GABA, HIS, ILE, LEU, LYS, MET, PRO, SER, TRP, TYR, VAL; **Low**: GLU
		Leaf	**High**: 16:3; **Low**: 18:2		**High**: 16:3, GABA; **Low** 18:2

Table 1 (cont'd)

Locus	Description	Tissue	MIPHENO Cutoff= 0.1	Zscore Cutoff = 2.5	Zscore Cutoff = 1.3
At4g19710	AK-HSDK II: Bifunctional aspartate kinase, homoserine dehydrogenase.	Seed			
		Leaf	**High**: 18:1D11, CYS, HSER, ILE, THR	**High**: CYS, HSER, ILE, THR	**High**: 18:1D11, CYS, HSER, ILE, THR
At4g27030*	FAD4: Palmitate desaturase	Seed			
		Leaf	**High**: 16:0, ALA, GLN, L.ALA; **Low**: 16:1D3	**High**: ALA; **Low**: 16:1D3	**High**: 16:0, ALA, GLN, SER, TRP; **Low**: 16:1D3
At4g33150	LKR/SDH: Splice variant of a bifunctional enzyme for lysine catabolism	Seed	**High**: HIS, LYS; **Low**: GLU	**High**: HIS, LYS	**High**: HIS, LYS, PRO
		Leaf			
At5g05730	ASA1: Alpha subunit of anthranilate synthase	Seed			
		Leaf	**Low**: TRP	**Low**: TRP	**Low**: TRP
At5g53460	GLT1: NADH-dependent glutamate synthase	Seed	**High**: ASN; **Low**: ASP		**High**: ASN, CYS; Low: ASP
		Leaf			

*Aracyc information not updated, manually added

Discussion

MIPHENO offers a way to control for assay variability in high throughout mutant screening studies. It outperformed using raw data or the group-based Z method in mutant identification on the synthetic data set (Figures 3, 4, and 5). Comparison of population parameters including proportion of WT and the distribution shape suggest that the method is tolerant to uneven distributions (tailing) and to higher mutant frequencies within the population. When applied to a biological data set, MIPHENO led to identification of more true mutants than the Z method for the Chloroplast 2010 set (Table 1) based on literature reported phenotypes or pathways. This suggests that MIPHENO reduces the false positive rate by decreasing the variation due to batch effects but does not directly influence the false non-discovery rate. The method additionally offers the user the ability to utilize any *a priori* information on the WT population/NULL distribution available as well as customize a quality control step that is sensitive to the needs of their process.

One drawback of using the normalization strategy described here is that it fails to control for the within-group variance to the degree that a quantile normalization strategy might. Quantile normalization makes the assumption that both the median or mean and the standard deviation of the data are all equal and would require sample sizes to be more or less equal as well as large enough to start approximating the normal distribution. This assumption does not always apply to post-hoc analysis; for example, the size of the sample groups in the Chloroplast 2010 data set varied from 12 to 96. MIPHENO aims at addressing this type of use case.

Conclusions

The strong performance of MIPHENO on two different data sets and its ability to permit cross-dataset comparisons of individuals without explicit controls makes it an ideal method for processing large datasets prior to Meta analyses combining different data sets from high-throughput experiments. Because more researchers are making their primary data available and the number of large-scale, high-throughput experiments keeps increasing, MIPHENO will provide a valuable processing platform that can theoretically be applied to very diverse measurement types (e.g. gene expression, enzyme kinetics, metabolite amounts).

Methods

2.6.1 Data analysis

All calculations were performed in R (R Development Core Team, 2011) v 2.11.0 on 64-bit Windows 7 platform. Chloroplast 2010 Project data used in the reported analysis was obtained on 8/18/2010. GO and Aracyc pathway information were obtained from the TAIR FTP site, files dated 8/2/2010 and 6/21/2010 respectively.

2.6.2 Generation of synthetic test data

Synthetic data were generated by sampling from three random Gaussian distributions representing low abundance, high abundance, and wild type levels of 'metabolite' (Figure 2) using a set of sampling probabilities. Distributions were created to assess the effects of uniform variance (e.g. same standard deviation) and proportional variance given by a relative standard deviation of 15% based on prior observations of real

data from the Chloroplast 2010 study. Means for the distributions were set such that the means of the 'mutant' populations were two standard deviations away from that of the wild type, because this is a common cut off for identifying hits in screening assays. The proportion of individuals sampled from each population (low, wt, high) was set prior to generating sample groups to test how different population composition influenced algorithm performance. To mirror the biological population structure, data were assigned to a flat, assay, and planting group representing individuals grown in the same physical unit, processed and assayed together, or grown over the same time course, respectively. Classification of each observed value was done at this step, prior to adding random noise (described below), defining a 'low' mutant as one that was 2 standard deviations below the WT mean and a 'high' mutant as one that was 2 standard deviations above. For calculating performance metrics, only the WT and mutant class were considered.

To simulate the non-biological variance, random uniform noise was added first at the level of planting group then at the level of assay group as empirical evidence suggested a greater assay effect than planting group effect. The resulting synthetic dataset was defined as raw data for use in the Z and raw data methods.

2.6.3 Method performance using the Chloroplast 2010 data

An overview of the data analysis approach is depicted in Figure 6. Data from the Chloroplast 2010 for mol% and nmol/g FW fatty acid methyl esters and amino acids were used to calculate both MIPHENO empirical *p*-values and *z*-scores. Samples genotyped as wild type or heterozygous for the T-DNA insertion were removed. The average phenotypic score (*z*-score or empirical *p*-value) per T-DNA insertion line was calculated and this was

used to define the phenotype for that insertion line. Next, loci where there were ≥ insertion lines showing the same (putative) phenotype for any attribute were identified based on either the empirical *p*-value or *z*-score and data from these line was combined across the 'mol %' and 'nmol/g FW' datasets. Loci from this list were analyzed and loci where >50% of the sampled lines showed a phenotype at a given cut off are considered putative mutants. To identify lines out of the putative mutants where phenotypic information is known, loci were cross-referenced to information from Aracyc and Gene Ontology annotation on biological processes (for experimentally-derived evidence codes only). Phenotypes predicted for these loci was then compared to phenotypes or experimental evidence reported in the literature to see if the predicted phenotype had been reported or if there was evidence for the gene product to act in a pathway leading directly to or from the measured metabolites.

REFERENCES

References

Ajjawi, I., Lu, Y., Savage, L.J., Bell, S.M., and Last, R.L. (2010). Large-scale reverse genetics in Arabidopsis: case studies from the Chloroplast 2010 Project. Plant Physiology *152*, 529-540.

Alonso, J.M., Stepanova, A.N., Leisse, T.J., Kim, C.J., Chen, H., Shinn, P., Stevenson, D.K., Zimmerman, J., Barajas, P., Cheuk, R., et al. (2003). Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. Science *301*, 653-657.

Ballman, K.V., Grill, D.E., Oberg, A.L., and Therneau, T.M. (2004). Faster cyclic loess: normalizing RNA arrays via linear models. Bioinformatics *20*, 2778-2786.

Barbaric, I., Miller, G., and Dear, T.N. (2007). Appearances can be deceiving: phenotypes of knockout mice. Briefings in Functional Genomics and Proteomics *6*, 91-103.

Berardini, T.Z., Mundodi, S., Reiser, L., Huala, E., Garcia-Hernandez, M., Zhang, P., Mueller, L.A., Yoon, J., Doyle, A., Lander, G., et al. (2004). Functional annotation of the Arabidopsis genome using controlled vocabularies. Plant Physiology *135*, 745-755.

Bouché, N., and Bouchez, D. (2001). Arabidopsis gene knockout: phenotypes wanted. Current Opinion in Plant Biology *4*, 111-117.

Eckel, J.E., Gennings, C., Therneau, T.M., Burgoon, L.D., Boverhof, D.R., and Zacharewski, T.R. (2005). Normalization of two-channel microarray experiments: a semiparametric approach. Bioinformatics *21*, 1078-1083.

Ferri, C., Hernandez-Orallo, J., and Modroiu, R. (2009). An experimental comparison of performance measures for classification. Pattern Recognition Letters *30*, 27-38.

Fiehn, O., Kopka, J., Dörmann, P., Altmann, T., Trethewey, R.N., and Willmitzer, L. (2000). Metabolite profiling for plant functional genomics. Nature Biotechnology *18*, 1157-1161.

Jander, G., Norris, S.R., Joshi, V., Fraga, M., Rugg, A., Yu, S., Li, L., and Last, R.L. (2004). Application of a high-throughput HPLC-MS/MS assay to Arabidopsis mutant screening; evidence that threonine aldolase plays a role in seed nutritional quality. The Plant Journal *39*, 465-475.

Last, R.L., Jones, A.D., and Shachar-Hill, Y. (2007). Towards the plant metabolome and beyond. Nature Reviews Molecular Cell Biology *8*, 167-174.

Lu, Y., Savage, L.J., Ajjawi, I., Imre, K.M., Yoder, D.W., Benning, C., DellaPenna, D., Ohlrogge, J.B., Osteryoung, K.W., Weber, A.P., et al. (2008). New connections across pathways

and cellular processes: industrialized mutant screening reveals novel associations between diverse phenotypes in arabidopsis. Plant Physiology *146*, 1482-1500.

Lu, Y., Savage, L.J., Larson, M.D., Wilkerson, C.G., and Last, R.L. (2011a). Chloroplast 2010: a database for large-scale phenotypic screening of Arabidopsis mutants. Plant Physiology *155*, 1589-1600.

Lu, Y., Savage, L.J., and Last, R.L. (2011b). Chloroplast phenomics: systematic phenotypic screening of chloroplast protein mutants in Arabidopsis. In Chloroplast Research in Arabidopsis : Methods and Protocols, Volume II, R.P. Jarvis, ed. (NY: Humana Press), pp. 161-185.

Mar, J.C., Kimura, Y., Schroder, K., Irvine, K.M., Hayashizaki, Y., Suzuki, H., Hume, D., and Quackenbush, J. (2009). Data-driven normalization strategies for high-throughput quantitative RT-PCR. BMC Bioinformatics *10,* 110.

Miron, M., and Nadon, R. (2006). Inferential literacy for experimental high-throughput biology. Trends in Genetics *22*, 84-89.

Mueller, L.A., Zhang, P., and Rhee, S.Y. (2003). AraCyc: a biochemical pathway database for Arabidopsis. Plant Physiology *132*, 453-460.

Quackenbush, J. (2002). Microarray data normalization and transformation. Nature Genetics *32*, 496-501.

R Development Core Team (2011). R: A Language and Environment for Statistical Computing (http://www.R-project.org: R Foundation for Statistical Computing).

Rhee, S.Y., Beavis, W., Berardini, T.Z., Chen, G., Dixon, D., Doyle, A., Garcia-Hernandez, M., Huala, E., Lander, G., Montoya, M., et al. (2003). The arabidopsis information resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. Nucleic Acids Research *31*, 224-228.

Rocke, D.M. (2004). Design and analysis of experiments with high throughput biological assay data. Seminars in Cell & Developmental Biology *15*, 703-713.

Van Eenennaam, A.L., Lincoln, K., Durrett, T.P., Valentin, H.E., Shewmaker, C.K., Thorne, G.M., Jiang, J., Baszis, S.R., Levering, C.K., Aasen, E.D., et al. (2003). Engineering vitamin E content: from Arabidopsis mutant to soy oil. Plant Cell *15*, 3007-3019.

Yang, Y.H., Dudoit, S., Luu, P., Lin, D.M., Peng, V., Ngai, J., and Speed, T.P. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Research *30*, e15.

Chapter 3

SimMeasure: A non-imputing approach to analyzing missing data

SimMeasure: A non-imputing approach to analyzing missing data

For many techniques used in the analysis of omics data a correlation matrix or adjacency matrix is needed. This type of data presentation can be used to graph networks, generate heat maps, and for clustering to identify communities of similar individuals. Screening data and other phenotypic type data has a couple issues that make creating a correlation matrix a challenge. First, a lot of responses are simply uninformative. These responses are ones in the background (like wild type phenotype) or were below detection level and are coded as zeros from the machine. Second there is a lot of missing data. This could be because a sample died or no response was observed.

SimMeasure, a central function in the NetComp package, aims at addressing these issues. It uses a modification of an existing distance metric and some programmatic changes that allow it to capture both positive and opposite relations. In addition it keeps track of the number of times one individual has a response where the other individual is missing a value or has a response below a given threshold. This is used to penalize the score such that pleiotropic individuals aren't shown as highly related to individuals with minimal responses. The resulting score describes the relationship between the two individuals and can be used to form the adjacency matrix.

In addition to SimMeasure, NetComp includes several other functions for analysis of large scale data, particularly adjacency matrices (coming from SimMeasure or a correlation calculation). Three functions aimed at making network comparisons easier: netIntersect (intersection, or what both have in common), netUnion (union, or everything from both

datasets), and netDif (difference between the two graphs). These are useful for combining information from two datasets or for getting an estimate of what information is lost by altering some parameters in the upstream analysis. Additional functions are included to facilitate thresholding of a matrix and comparing the clustering results of two graphs.

Abstract

3.1.1 Motivation

High throughput datasets are often plagued by missing data, making it difficult to conduct large-scale analyses and compare across datasets without data removal or computationally-intensive imputation. Additionally, many high throughput measurements for screening studies (e.g. metabolomics, transcriptomics, enzyme kinetics) contain sample responses in the background range or are otherwise not of interest. Methods with a tolerance for missing data and uninformative responses are needed to conduct cross-dataset comparisons of high throughput data.

3.1.2 Results

SimMeasure is a method for calculating the similarity between two individuals with a high tolerance of large amounts of missing data. We show that SimMeasure is an effective algorithm for analyzing datasets with large amounts of missing data as it is robust to missing data, can handle data thresholding, and requires little *a priori* knowledge versus existing data imputation methods. SimMeasure is part of the NetComp R package, developed for the analysis of high throughput phenotypic and other large-scale quantitative data. Additional functions for adjacency matrices calculate the intersection, union, and difference between graphs. These functions aid the exploration of high throughput data and enable faster graph calculations to facilitate meta-analysis. Analysis of a complex screening dataset, ToxCast, using the methods in NetComp illustrates the utility in hypothesis generation and data integration.

3.1.3 Availability

NetComp is distributed through the Comprehensive R Archive Network (CRAN), http://cran.r-project.org/web/packages/NetComp/index.html

Introduction

Data from high throughput and large-scale experiments can provide a wealth of information on the relationship between individuals (e.g., tissue samples, compounds, mutants) and measured attributes (Ideker et al., 2001; Joyce and Palsson, 2006; Last et al., 2007). Ideally, researchers could leverage these large datasets in hypothesis development in much the same way that transcriptomics data are used to identify communities with similar properties. Unfortunately, high throughput data, especially from biological experiments, tend to be plagued by missing or uninformative data representing some basal response (Aittokallio, 2010; Bell et al., 2012). For phenotypic screens in particular, the researcher is often looking for a response signature composed of just one characteristic, differentiating a few individuals from the group for further analysis. In this type of situation, the small number of cases where there is a strong attribute response are more important in establishing the communities than missing data due to a failed assay or responses at background level.

Existing methods for calculating the correlation or similarity between two individuals (e.g., Pearson's correlation coefficient, Euclidean distance) do not handle these types of data well, especially as the proportion of missing/uninformative data increases. Imputation methods, designed to estimate the missing values have been developed for use in large datasets, including those based on *k*-nearest neighbors, variations of least squares

analyses, Bayesian PCA, and singular value decomposition (Boulesteix and Strimmer, 2007; Brock et al., 2008; Oh et al., 2011; Troyanskaya et al., 2001; Yang et al., 2006). Data imputation has been used with gene expression and other high throughput datasets (see Aittokallio, 2010; Liew et al., 2011 for review) with some success. Recently, several different data imputation methods were compared to evaluate their effectiveness on downstream transcriptomic analyses such as differential gene expression, clustering, and classification (Oh et al., 2011). It was found that Bayesian PCA and local least squares outperformed other methods with respect to differential expression and clustering analysis, but no clear imputation method stood out for classification.

While imputation may help alleviate some of the issues with missing data, its utility depends at some level on the structure of the data. If the data have no correlation or expected relationship between observations, these methods may not be appropriate. This is especially true when combining different datasets, such as proteomics and transcriptomics, to perform meta-analyses. An additional constraint is the required time and computational resources to appropriately model the missing values, which can be burdensome as the dataset grows.

SimMeasure is based on the Canberra distance (Lance and Williams, 1967) and calculates the similarity between two individuals, providing a result analogous to a correlation coefficient. It provides a way to carry out clustering and network-based analyses on high throughput datasets with missing values. Along with SimMeasure, the package NetComp contains a suite of adjacency-matrix based functions for meta-analysis to quickly compare or integrate data sets. Together, NetComp is aimed at facilitating the

generation and comparison of networks based on high throughput and large scale data providing a framework for analyzing sparse and low information content datasets.

System and methods

3.3.1 Algorithm

SimMeasure is based on a modified version of the Canberra distance metric and used to capture the similarity between two sets of observations. The SimMeasure algorithm, outlined in Figure 7, is implemented in C/R to shorten computational time. User inputs are a data matrix with rows describing individuals and columns containing the assays/measured responses, and an optional threshold value, *t*, to define the background level. Only values ≥ *t* are considered in the calculation, thus removing background signal along with missing values.

Consider individuals, *X* and *Y*, and their responses i → N. The algorithm first considers each response pair and evaluates if one individual had a response that was greater than the threshold while the other was missing. If this is the case, the counter *nm* (no match) is increased. In the event that both responses are greater than *t* and of the same sign, the value *pm* (positive match) is increased by the percent similarity of the two responses. If the value *t* is not provided and both individuals have a response of 0, *pm* is increased by 1. In the event that the responses are of opposite sign (i.e. one is a positive value and the other negative), the value *om* (opposite match) is increased instead. Once all responses are evaluated, the similarity score is calculated using a weighted approach. This approach uses *nm* to penalize for the number of instances where one individual had a

response of interest and the other did not. Responses where both individuals have missing values are omitted.

Consider samples **X, Y**

For each response ***i***

If one response is missing and the other is above threshold

Increase **NM**

If both responses have the same sign

If both are **ZERO, pm=pm+1,** increase **pcnt**

Else $\mathbf{pm = pm + 1 - \frac{\left||X_i| - |Y_i|\right|}{|X_i| + |Y_i|}}$**;** increase **pcnt**

If responses have opposite sign

$\mathbf{om = om + 1 - \frac{\left||X_i| - |Y_i|\right|}{|X_i| + |Y_i|}}$**;** increase **ocnt**

$$\mathbf{SimMeasure} = \frac{\mathbf{pm} - \mathbf{om}}{\mathbf{pcnt} + \mathbf{ocnt} + \frac{\mathbf{nm}}{\mathbf{pcnt} + \mathbf{ocnt} + \mathbf{nm}}}$$

Figure 7 SimMeasure Algorithm

3.3.2 Datasets

Three datasets were used to evaluate SimMeasure: a yeast gene expression dataset (Causton et al., 2001), ToxCast, a high throughput chemical screening dataset (Judson et al., 2010), and ToxRef, a physiological-based throughput chemical screening dataset (Knudsen et al., 2009; Martin et al., 2009). The yeast data were preprocessed by removing any values below the Affymetric detection call. ToxCast data consists of quantitative high throughput screening *in vitro* assay responses to a given chemical, including gene expression, cell-based and cell free, receptor, and cytotoxicity assays. ToxRef is a complementary dataset to ToxCast, and provides *in vivo* toxicological and pathophysiological measurements such as tumor counts and developmental abnormalities corresponding to the chemicals assayed in ToxCast. Values in ToxRef represent the lowest dose of chemical at which the endpoint such as tumor was observed, with 'no value' recorded when no effect was observed in the study. Chemicals (based on Chemical Abstracts registry number; CASRN) that were replicated in the study had a letter amended to the CASRN such that this could serve as a unique identifier. This had the additional benefit of being able to identify internal consistency as replicates of the same chemical are expected to have highly similar responses.

3.3.3 Method evaluation

To evaluate the performance of SimMeasure (SM) against existing data imputation techniques, missing values were introduced at random into the yeast dataset (from 0.1% to 90% missing values) and the ability of the downstream method to generate a correlation

coefficient or clusters as the original dataset was evaluated. Bayesian principle component analysis (BPCA) and Local Least Squares (LLS) were used in the comparison based on their performance across a wide range of datasets (Brock et al., 2008; Oh et al., 2011). Pearson's Correlation Coefficient (PC) with pairwise complete observations was used to represent the case of no intervention as it simply evaluates instances where observations from both individuals are present. Analyses were all performed in R statistical software (R Development Core Team, 2011) with missForest (Stekhoven and Bühlmann, 2011) used to generate the test datasets and pcaMethods (Stacklies et al., 2007) for the imputation functions.

For each level of missing values (0.1 to 90%), 200 iterations were run comparing the missing value (MV) data matrix to the complete (CV) data matrix. As the percentage of missing values increased, data sparsity increased such that there were whole rows with missing data in the MV matrix. These rows were removed from both MV and CV to carry out the downstream analyses. After the MV imputation step (for BPCA and LLS only), the weighted adjacency matrix for each of the methods tested (BPCA, LLS, PC, and SM) were evaluated for rows containing all missing values. Again, only individuals present in the MV and CV matrix were used for method evaluation.

To evaluate the ability of the methods to reconstruct the appropriate adjacency matrix (and thereby clusters) from the MV matrix, the root mean square error (RMSE), Adjusted Rand Index (ARI) and balanced accuracy (BA) were measured (Oh et al., 2011). Equations are detailed in Figure 8.

$$RMSE = \sqrt{\frac{(observed - predicted)^2}{2}}$$

$$BA = \frac{0.5 * TP}{TP + FP} + \frac{0.5 * TN}{TN + FP}$$

$$ARI = \frac{TP - EI}{\frac{M + M_2}{2} - EI}$$

$$EI = \frac{M_1 + M_2}{TP = FP = TN = FN}$$

$$M_1 = TP + FP$$

$$M_2 = TP + FN$$

Figure 8 Equations used for Evaluating Method Performance

TP= True Postive, TN= True Negative, FP= False Postive, FN=False Negative

Balanced accuracy represents the arithmetic mean of sensitivity and specificity and is useful for imbalance classes, as expected in the clustering. The adjacency matrices were used for the RMSE to generate an estimate of how different the correlations derived from missing data compared to those from the complete dataset. If thresholding is used when generating the communities (e.g., removing correlations below 0.5), then the accuracy of the correlation coefficient becomes an important consideration. Balanced accuracy and ARI were used to evaluate the communities to obtain an estimate of the ability to obtain the same structure from the MV data as the CV. Communities, or groups of individuals with shared properties, were built from hierarchical clustering using Ward's minimum variance. Ward's minimum variance minimizes the within-cluster variance. A cutoff of 500 (500 communities) was chosen to keep the average cluster size small.

3.3.4 Application to complex dataset

The ToxCast ('Cast') and ToxRef ('Ref') datasets were analyzed using the SimMeasure and netIntersect functions in NetComp to illustrate NetComp's utility in analysis of datasets with large numbers of missing data (as few as one observation noted per chemical). Note that missing data in this set is often due to no response measured at the levels of chemical assayed, in which case imputing the missing value would be inappropriate. Rows and columns were removed from both datasets (Ref and Cast) where the number of observations fell below the median for that dataset; this had the effect of reducing the dataset by almost half. This was done to remove data-poor assays and physiological measures. Datasets were converted into adjacency matrixes using

SimMeasure with a threshold = 0 (no data excluded), and edge weights represent the similarity measure.

To find instances where evidence supported a community from each graph (Ref and Cast), the network intersection was determined using the NetComp function netIntersect with an edge weight threshold of 0.5. This enables capture of strong edges from the Ref network, i.e. those with a weight of 1; this is desirable given the high false-negative rate of the ToxCast screening study. Ten communities were designated from the intersection graph using hierarchical clustering with Ward's minimum variance. Communities were analyzed to identify the driving assays from both input datasets (ToxCast and ToxRef).

Results and discussion

3.4.1 Network generation with missing values

The four methods were evaluated over a wide range of missing data for their ability to generate a network/community structure similar to the complete dataset. The RMSE describes the error in correlation values (e.g. edge weights) between CV matrix and the MV matrix (Figure 9, A). All methods perform well at low missing values (0.1-10%), but SimMeasure consistently outperforms the other methods at all levels of missing data and appears stable up to 50% missing data. Interestingly, Local Least Squares shows a sharp increase in RMSE between 20-40% missing values with correspondingly large interquartile ranges.

The quality of clustering based was measured by the Adjusted Rand Index (ARI, Figure 9, B) (Oh et al., 2011)and Balanced Accuracy (BA, Figure 9, C) clustering which

measures the similarity between clusters (adjusted for chance) and the averaged accuracy of each class, respectively. Note that the methods perform similarly across both performance measures, with the SimMeasure outperforming the other methods. Of interest is the point at which the methods begin to give cluster results at random (BA=0.5). Local Least Squares performance fails with over 20% missing values while Bayesian PCA and Pearson's fail with over 40% missing values and SimMeasure with over 50%. These results demonstrate that SimMeasure is able to reconstruct the network structure of a dataset with a high portion of missing values.

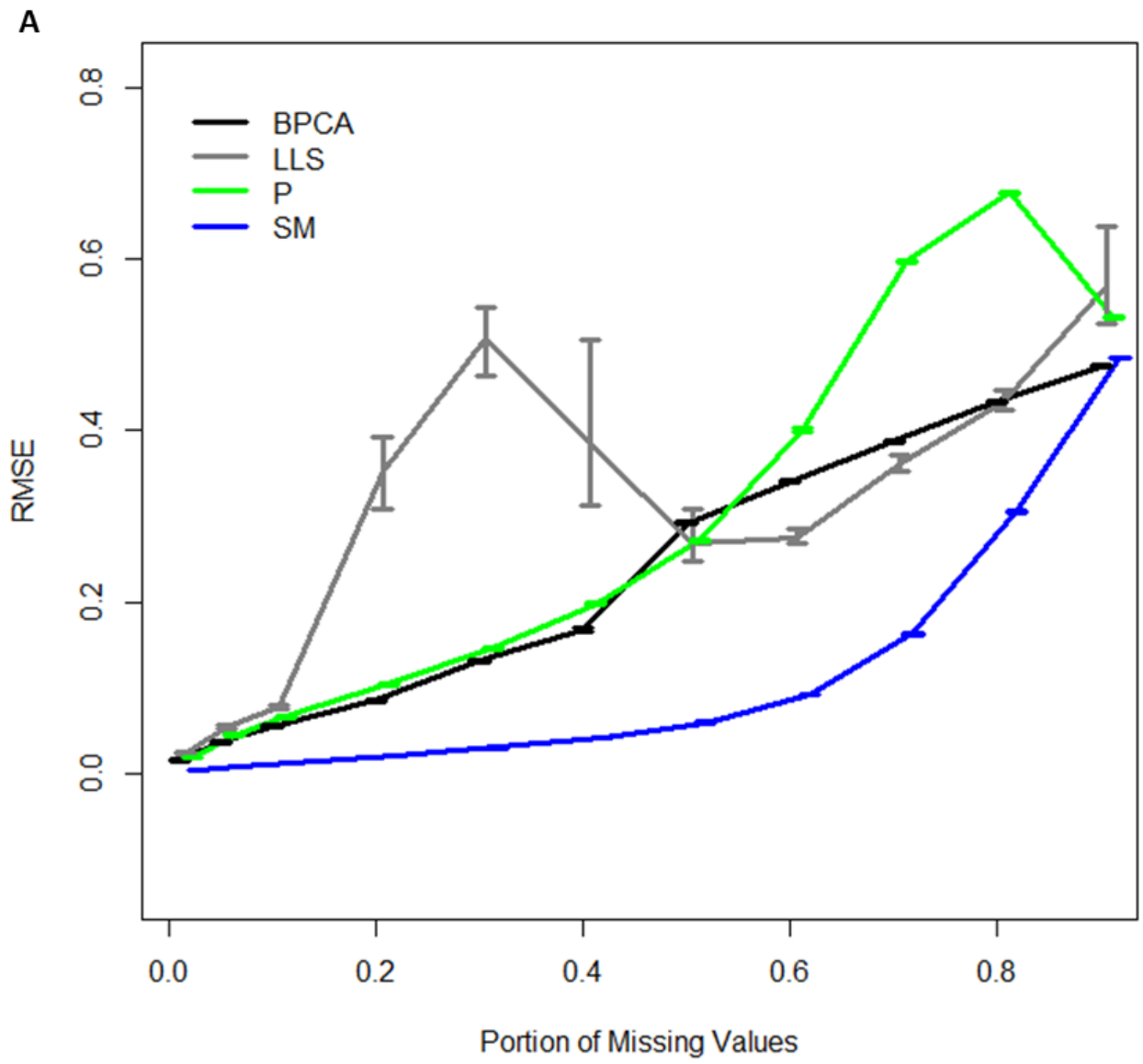


Figure 9 Method Performance verses Missing Values

The performance (y-axis) is shown across the portion of missing values (x-axis). BPCA= Bayesian Principle Component Analysis, LLS= Local Least Squares, P= Pearson's Correlation with pairwise-complete observations, SM= SimMeaure. Graphs represent the median value and inter quartile range of 200 trials. (A) Root Mean Square Error measures, (B) Adjusted Rand Index, and (C) Balanced Accuracy.

Figure 9 (cont'd)

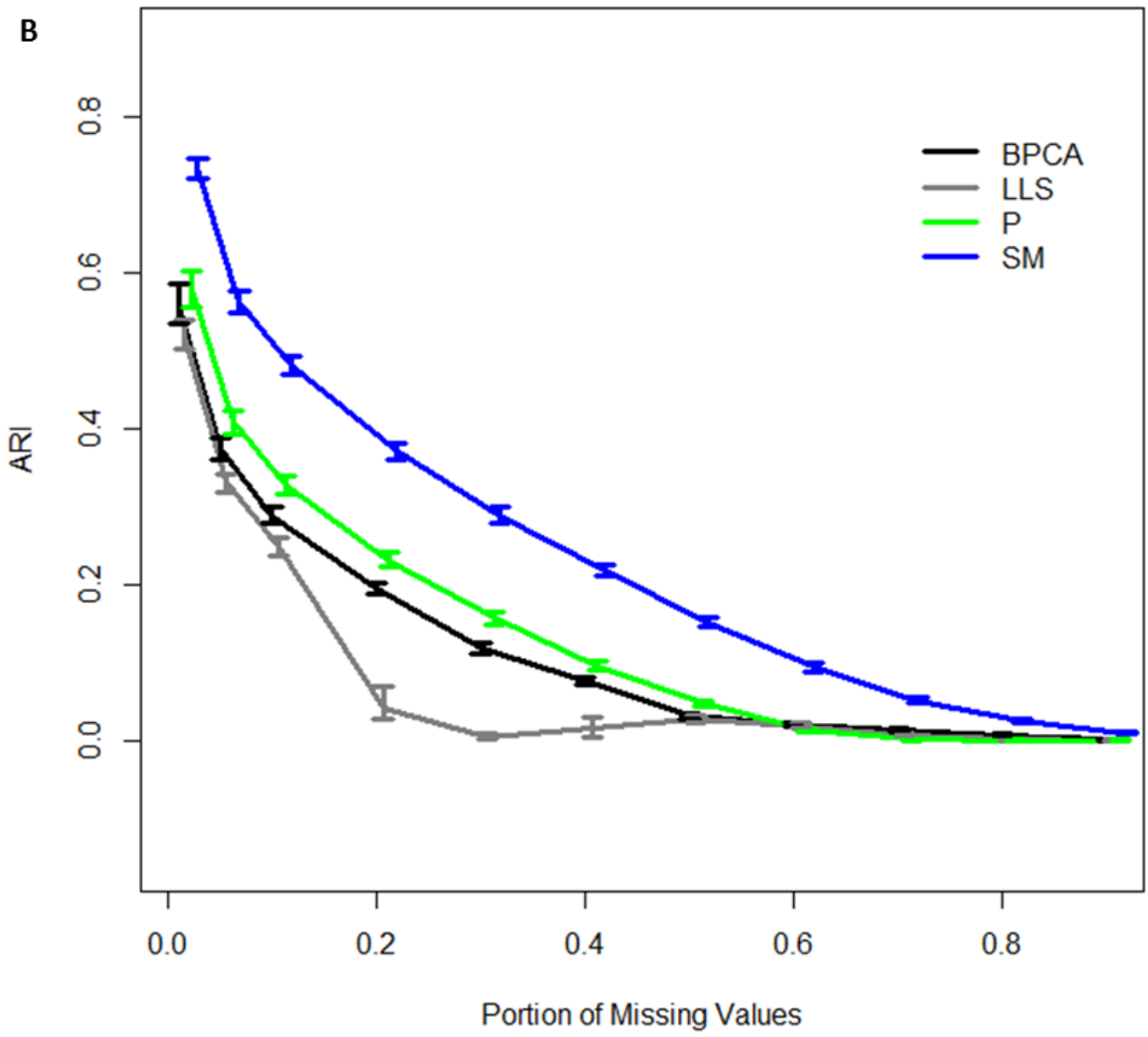
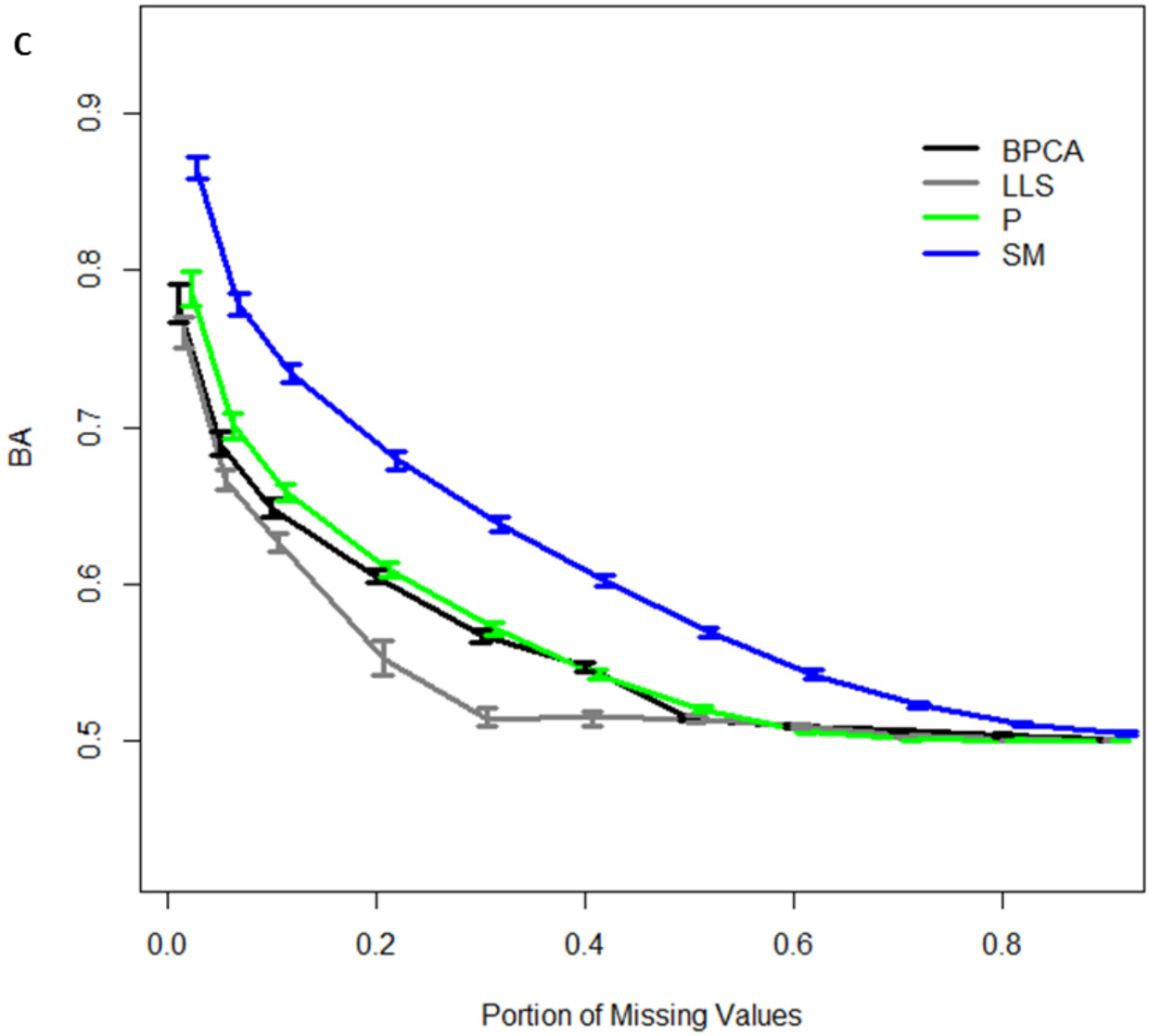


Figure 9 (cont'd)



3.4.2 NetComp workflow: complex dataset

ToxCast is a complex high throughput chemical screen with a high portion of missing data. The goal with the analyses is to see how well the results in ToxCast match up with the physiological endpoints for the same chemicals in ToxRef. Because of the data sparsity, none of the other methods tested in the evaluation could work without error on the dataset (data not shown). Cluster analysis of the intersection between the ToxCast and ToxRef datasets (clustering Figure 10, results Table 2) shows the ability of this approach to categorize chemicals based on toxicity. The yellow cluster, for example, contains a group of chemicals across a range of pesticide categories, many of which are organophosphates and organochlorines, although they are structurally diverse. The signature for this group is driven by liver assays (based on both ToxCast and ToxRef), but the ToxCast fails to highlight the potential reproductive effects of these chemicals noted in the ToxRef dataset (a false negative result). The red cluster, on the other hand, shows high levels of maternal and reproductive toxicity in addition to liver toxicity. These factors are well captured in the ToxCast data by the activation of receptors for progesterone and androgen as well as androgen and testosterone metabolizing enzymes. The hypoxia response could be hypothesized to contribute to both the liver toxicity and the reproductive toxicity in this case.

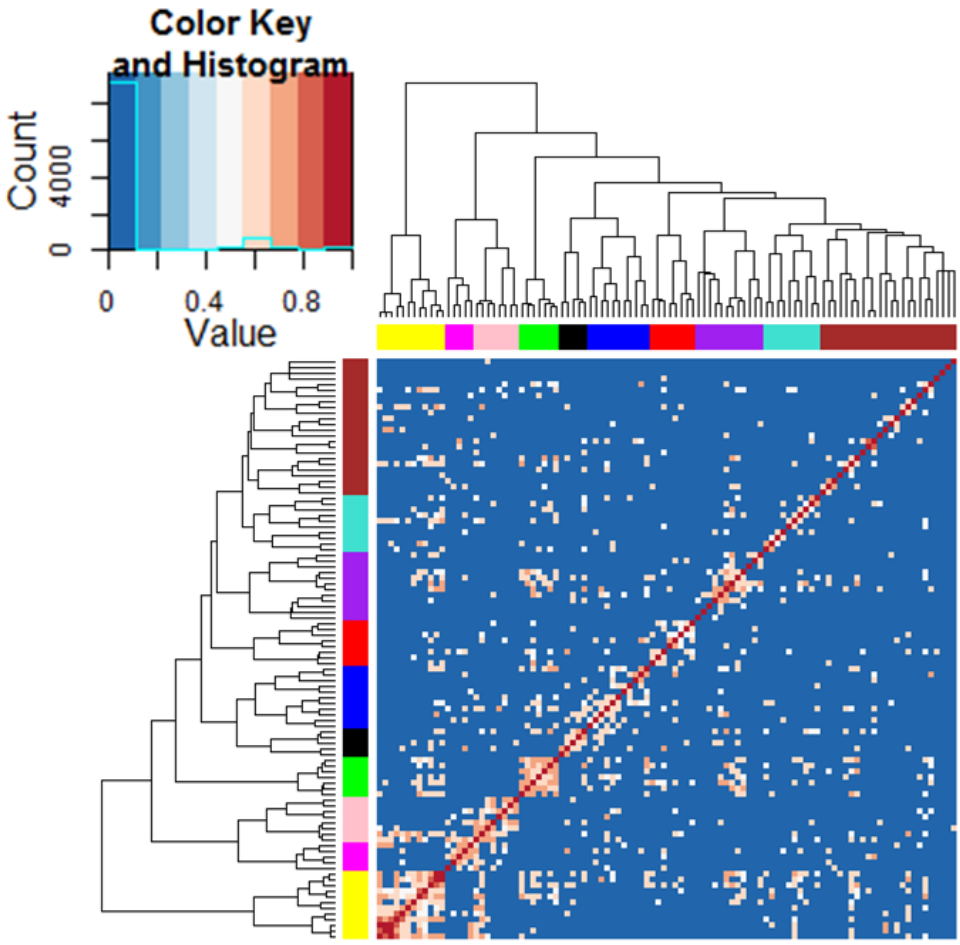


Figure 10 Heatmap of the Intersection between ToxCast and ToxRef

Row and column colors represent the clusters described in Table 2. Color reflects edge weights.

Table 2 Results of intersection between ToxRef and ToxCast Networks

Cluster	ToxRef	ToxCast
Yellow	Reproductive, Liver	Liver
Magenta	Liver, Maternal	Nuclear Receptor, Hormone Receptor, Genotoxic
Pink	Maternal, Developmental, Thyroid, Adrenal, Liver, Kidney	Androgen Receptor, Cytotoxic
Green	Liver, Tumor, Maternal	Cytotoxic
Black	Liver, Tumor, Reproductive, Development	Serotonin, Opiate, Pregnane X Receptor
Blue	Liver, Maternal	Hypoxia Response, PXR Activation, CYP2C19 Activation
Red	Liver, Maternal, Reproductive	Progesterone/Androgen Receptor, Hypoxia Response, Steroid and Drug Metabolism
Purple	Maternal	Steroid Metabolism, CYP2C19 Activation
Turquoise	Tumor, Maternal, Developmental	CYP2C19 Activation
Brown	Liver Tumor, Maternal	CYP2C19 Activation

Conclusions

NetComp provides a critical set of tools for those dealing with high throughput or other large numerical datasets with missing or uninformative vales within the R statistical software suite (R Development Core Team, 2011). SimMeasure enables researchers to calculate the similarity between observations in their dataset and leverage the resulting adjacency matrix to perform Meta analyses with other datasets. SimMeasure was able to outperform other missing-value imputation methods, requires minimal parameter optimization, and corrects the output for the number of missing observations. Results of applying functions in the NetComp package to a complex dataset of low density (ToxCast) suggest that the approach is useful for analysis of high throughput data and can aid in hypothesis development. Further, these tools will facilitate analyses and enable the integration of diverse datasets to overcome lack low information content.

REFERENCES

References

Aittokallio, T. (2010). Dealing with missing values in large-scale studies: microarray data imputation and beyond. Briefings in Bioinformatics *11*, 253-264.

Ajjawi, I., Lu, Y., Savage, L.J., Bell, S.M., and Last, R.L. (2010). Large-scale reverse genetics in Arabidopsis: case studies from the Chloroplast 2010 Project. Plant Physiology *152*, 529-540.

Alonso, J.M., Stepanova, A.N., Leisse, T.J., Kim, C.J., Chen, H., Shinn, P., Stevenson, D.K., Zimmerman, J., Barajas, P., Cheuk, R., et al. (2003). Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. Science *301*, 653-657.

Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J., et al. (2000). Gene Ontology: tool for the unification of biology. Nature Genetics *25*, 25-29.

Ballman, K.V., Grill, D.E., Oberg, A.L., and Therneau, T.M. (2004). Faster cyclic loess: normalizing RNA arrays via linear models. Bioinformatics *20*, 2778-2786.

Barbaric, I., Miller, G., and Dear, T.N. (2007). Appearances can be deceiving: phenotypes of knockout mice. Briefings in Functional Genomics and Proteomics *6*, 91-103.

Bell, S.M., Burgoon, L.D., and Last, R.L. (2012). MIPHENO: data normalization for high throughput metabolite analysis. BMC bioinformatics *13*, 10.

Berardini, T.Z., Mundodi, S., Reiser, L., Huala, E., Garcia-Hernandez, M., Zhang, P., Mueller, L.A., Yoon, J., Doyle, A., Lander, G., et al. (2004). Functional annotation of the Arabidopsis genome using controlled vocabularies. Plant Physiology *135*, 745-755.

Bouché, N., and Bouchez, D. (2001). Arabidopsis gene knockout: phenotypes wanted. Current Opinion in Plant Biology *4*, 111-117.

Boulesteix, A.-L., and Strimmer, K. (2007). Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. Briefings in Bioinformatics *8*, 32-44.

Brock, G., Shaffer, J., Blakesley, R., Lotz, M., and Tseng, G. (2008). Which missing value imputation method to use in expression profiles: a comparative study and two selection schemes. BMC Bioinformatics *9*, 12.

Causton, H.C., Ren, B., Koh, S.S., Harbison, C.T., Kanin, E., Jennings, E.G., Lee, T.I., True, H.L., Lander, E.S., and Young, R.A. (2001). Remodeling of yeast genome expression in response to environmental changes. Molecular Biology of the Cell *12*, 323-337.

Eckel, J.E., Gennings, C., Therneau, T.M., Burgoon, L.D., Boverhof, D.R., and Zacharewski, T.R. (2005). Normalization of two-channel microarray experiments: a semiparametric approach. Bioinformatics *21*, 1078-1083.

Ferri, C., Hernandez-Orallo, J., and Modroiu, R. (2009). An experimental comparison of performance measures for classification. Pattern Recognition Letters *30*, 27-38.

Fiehn, O., Kopka, J., Dörmann, P., Altmann, T., Trethewey, R.N., and Willmitzer, L. (2000). Metabolite profiling for plant functional genomics. Nature Biotechnology *18*, 1157-1161.

Furnham, N., Garavelli, J.S., Apweiler, R., and Thornton, J.M. (2009). Missing in action: enzyme functional annotations in biological databases. Nature Chemical Biology *5*, 521-525.

Horan, K., Jang, C., Bailey-Serres, J., Mittler, R., Shelton, C., Harper, J.F., Zhu, J.-K., Cushman, J.C., Gollery, M., and Girke, T. (2008). Annotating genes of known and unknown function by large-scale coexpression analysis. Plant Physiology *147*, 41-57.

Ideker, T., Thorsson, V., Ranish, J.A., Christmas, R., Buhler, J., Eng, J.K., Bumgarner, R., Goodlett, D.R., Aebersold, R., and Hood, L. (2001). Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science *292*, 929-934.

Jander, G., Norris, S.R., Joshi, V., Fraga, M., Rugg, A., Yu, S., Li, L., and Last, R.L. (2004). Application of a high-throughput HPLC-MS/MS assay to Arabidopsis mutant screening; evidence that threonine aldolase plays a role in seed nutritional quality. The Plant Journal *39*, 465-475.

Joyce, A.R., and Palsson, B.O. (2006). The model organism as a system: integrating 'omics' data sets. Nature Reviews Molecular Cell Biology *7*, 198-210.

Judson, R.S., Houck, K.A., Kavlock, R.J., Knudsen, T.B., Martin, M.T., Mortensen, H.M., Reif, D.M., Rotroff, D.M., Shah, I., Richard, A.M., et al. (2010). *In vitro* screening of environmental chemicals for targeted testing prioritization: the ToxCast Project. Environmental Health Perspective *118,* 485-492.

Knudsen, T.B., Martin, M.T., Kavlock, R.J., Judson, R.S., Dix, D.J., and Singh, A.V. (2009). Profiling the activity of environmental chemicals in prenatal developmental toxicity studies using the U.S. EPA's ToxRefDB. Reproductive Toxicology *28*, 209-219.

Lance, G., and Williams, W. (1967). Mixed-data classificatory programs I - agglomerative systems. Australian Computer Journal *1*, 15-20.

Last, R.L., Jones, A.D., and Shachar-Hill, Y. (2007). Towards the plant metabolome and beyond. Nature Reviews Molocular Cell Biology *8*, 167-174.

Liew, A.W.-C., Law, N.-F., and Yan, H. (2011). Missing value imputation for gene expression data: computational techniques to recover missing data from available information. Briefings in Bioinformatics *12*, 498-513.

Lu, Y., Savage, L.J., Ajjawi, I., Imre, K.M., Yoder, D.W., Benning, C., DellaPenna, D., Ohlrogge, J.B., Osteryoung, K.W., Weber, A.P., et al. (2008). New connections across pathways and cellular processes: industrialized mutant screening reveals novel associations between diverse phenotypes in Arabidopsis. Plant Physiology *146*, 1482-1500.

Lu, Y., Savage, L.J., Larson, M.D., Wilkerson, C.G., and Last, R.L. (2011a). Chloroplast 2010: a database for large-scale phenotypic screening of Arabidopsis mutants. Plant Physiology *155*, 1589-1600.

Lu, Y., Savage, L.J., and Last, R.L. (2011b). Chloroplast phenomics: systematic phenotypic screening of chloroplast protein mutants in Arabidopsis. In Chloroplast Research in Arabidopsis : Methods and Protocols, Volume II, R.P. Jarvis, ed. (NY: Humana Press), pp. 161-185.

Mar, J.C., Kimura, Y., Schroder, K., Irvine, K.M., Hayashizaki, Y., Suzuki, H., Hume, D., and Quackenbush, J. (2009). Data-driven normalization strategies for high-throughput quantitative RT-PCR. BMC Bioinformatics *10*, 110.

Martin, M.T., Judson, R.S., Reif, D.M., Kavlock, R.J., and Dix, D.J. (2009). Profiling chemicals based on chronic toxicity results from the U.S. EPA ToxRef Database. Environmental Health Perspectives *117*, 392-399.

Miron, M., and Nadon, R. (2006). Inferential literacy for experimental high-throughput biology. Trends in Genetics *22*, 84-89.

Mueller, L.A., Zhang, P., and Rhee, S.Y. (2003). AraCyc: a biochemical pathway database for Arabidopsis. Plant Physiol *132*, 453-460.

Oba, S., Sato, M.-a., Takemasa, I., Monden, M., Matsubara, K.-i., and Ishii, S. (2003). A Bayesian missing value estimation method for gene expression profile data. Bioinformatics *19*, 2088-2096.

Ober, D. (2010). Gene duplications and the time thereafter - examples from plant secondary metabolism. Plant Biology *12*, 570-577.

Oh, S., Kang, D.D., Brock, G.N., and Tseng, G.C. (2011). Biological impact of missing-value imputation on downstream analyses of gene expression profiles. Bioinformatics *27*, 78-86.

Quackenbush, J. (2002). Microarray data normalization and transformation. Nature Genetics *32*, 496-501.

R Development Core Team (2011). R: A Language and Environment for Statistical Computing (http://www.R-project.org: R Foundation for Statistical Computing).

Rhee, S.Y., Beavis, W., Berardini, T.Z., Chen, G., Dixon, D., Doyle, A., Garcia-Hernandez, M., Huala, E., Lander, G., Montoya, M., et al. (2003). The arabidopsis information resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. Nucleic Acids Research *31*, 224-228.

Rocke, D.M. (2004). Design and analysis of experiments with high throughput biological assay data. Seminars in Cell & Developmental Biology *15*, 703-713.

Schlomer, G.L., Bauman, S., and Card, N.A. (2010). Best practices for missing data management in counseling psychology. Journal of Counseling Psychology *57*, 1-10.

Stacklies, W., Redestig, H., Scholz, M., Walther, D., and Selbig, J. (2007). pcaMethods—a bioconductor package providing PCA methods for incomplete data. Bioinformatics *23*, 1164-1167.

Stekhoven, D.J., and Bühlmann, P. (2011). MissForest - nonparametric missing value imputation for mixed-type data. Bioinformatics *28*, 112-118.

The Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature (London) *408*, 796-815.

Thorneycroft, D., Sherson, S.M., and Smith, S.M. (2001). Using gene knockouts to investigate plant metabolism. Journal of Experimental Botany *52*, 1593-1601.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R.B. (2001). Missing value estimation methods for DNA microarrays. Bioinformatics *17*, 520-525.

Van Eenennaam, A.L., Lincoln, K., Durrett, T.P., Valentin, H.E., Shewmaker, C.K., Thorne, G.M., Jiang, J., Baszis, S.R., Levering, C.K., Aasen, E.D., et al. (2003). Engineering vitamin E content: from Arabidopsis mutant to soy oil. Plant Cell *15*, 3007-3019.

Wang, C., Marshall, A., Zhang, D., and Wilson, Z.A. (2012). ANAP: an integrated knowledge base for Arabidopsis protein interaction network analysis. Plant Physiology *158*, 1523-1533.

Yang, K., Li, J., and Wang, C. (2006). Missing values estimation in microarray data with partial least squares regression. In Proceedings of the 6th international conference on Computational Science - Volume Part II (ICCS'06), V. Alexandrov, G. Albada, P. Sloot, and J. Dongarra, eds. (Springer Berlin / Heidelberg), pp 662-669.

Yang, Y.H., Dudoit, S., Luu, P., Lin, D.M., Peng, V., Ngai, J., and Speed, T.P. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Research *30*, e15.

Yuan, J.S., Galbraith, D.W., Dai, S.Y., Griffin, P., and Stewart, C.N. (2008). Plant systems biology comes of age. Trends in Plant Science *13*, 165-171.

Chapter 4

Phenotypic enrichment from large scale phenomics of *Arabidopsis thaliana*

Phenotypic enrichment from large scale phenomics of *Arabidopsis thaliana*

The main objective of the dissertation was to test the hypothesis that high throughput phenotypic data could be leveraged in building models of gene function. This chapter applies the methods developed in the first two chapters to the Chloroplast 2010 dataset with that aim in mind. Using the normalization strategies and other tools from the MIPHENO package, the Chloroplast 2010 data was converted to values comparable across the dataset and reflective of their likelihood of being a mutant phenotype. These values were zero centered such that values approaching 0 had a high likelihood of being wild type, while those approaching -1 or 1 had a high likelihood of being mutant. The reflexive distribution could be used with SimMeasure to calculate the similarity score between individuals and generate the adjacency matrix facilitating community analysis.

Community analysis is a tool commonly used in large-scale data analysis. The purpose is to look at the members of the community to see if there are any common features. For example, enrichment in a specific gene ontology term or metabolic pathway (relative to the background) can serve as the basis for developing the hypothesis that the under annotated genes might also participate in that process or pathway. A community is a group of individuals that are more closely related to each other than to those outside the community, either by the number of edges (if defining graphically) or some similarity value. Communities can be defined based on many different things, but generally they are defined by using a clustering method (such as k means or hierarchical clustering) and specifying the number of communities or the distance on the tree (for the hierarchical

methods). This analysis uses a hierarchical clustering method that minimizes the variance between individuals in the branches of the tree called Ward's minimum variance. It is important to note that with many methods, including this method, not all individuals in a community are highly correlated with each other. It is the overall cohesion of the community that is actually considered.

These results support the hypothesis that phenomics data will improve models of gene function. They show that with minor modification approaches used for analyzing gene expression data can work with screening data and we believe that the results are more useful in the design of follow-up experiments because they are tied to a metabolic phenotype. The overall workflow presented in this paper is one of the first to leverage phenotypic data of this type for such a purpose and can serve as an example for future analyses.

Abstract

Functional gene annotation provides insight to the role a gene plays within the organism. Despite the growing amount of genome sequence available, gene annotation lags behind, even in model species. Large-scale high throughput experiments such as transcriptomic studies have helped develop hypotheses for under-annotated genes; however, transcript information is often lacking in its ability to describe the function of a gene. Technological advances have made using metabolic profiling for large scale phenomics increasingly possible and opens up a new data source from which genes might be characterized. Hypotheses of gene function and relationships between metabolic sub-networks were built using a large-scale high throughput phenomic screening study of *Arabidopsis thaliana* mutants. Methods explored in this work pave the way for using other high throughput datasets to build models of gene function.

Introduction

High-throughput experiments such as transcriptomics and proteomics have provided important information that enable researchers to better understand gene function and the effect of environment on cellular physiology. Furthermore, they have advanced our ability to develop hypotheses of a gene's role in the biology of an organism based on the response profile of transcripts or proteins. Even with all the terabytes of sequence, transcript and proteome data available, many model species have un- or under-annotated genomes and a small fraction of the genome is well annotated even for the best studied organisms. In the plant model species, *Arabidopsis thaliana*, almost half of the genes have no known or inferred function (The Arabidopsis Information Resource, 2010). These

missing pieces of information pose a challenge for metabolic engineering and developing a holistic understanding of biology in general. Unfortunately, sharing a similar transcript profile does not always mean that genes are involved in related physiological processes or the relationship may be driven by unclear patterns (false positives). Conversely, genes involved in related processes do not always have similar transcription profiles (Vandepoele et al., 2009; Williams and Bowles, 2004). Thus it is important to test for function by a variety of different approaches.

Genetic mutants have the ability to shed light on the function of altered genes, but phenotypes can be masked due to functional redundancy or the limited number of phenotypes being measured (Ajjawi et al., 2010; Bouché and Bouchez, 2001). Knockout collections, such as sequence-indexed T-DNA insertion lines (Alonso et al., 2003), provide starting material for surveying the function of a gene by measuring the state of the plant in the absence or reduced expression of the gene product. Unfortunately, the problem of knowing what to assay to describe the gene's function still exists. The Chloroplast 2010 project is a high throughput phenotypic screen of >5,000 *Arabidopsis thaliana* T-DNA mutant lines that analyzed metabolite, physiological and morphological data from leaf and seed defective in >3000 different loci (Lu et al., 2008; Lu et al., 2011b; Lu et al., 2011c). Because it is enriched for mutants in chloroplast targeted genes, this dataset may provide insight on the genes involved in key chloroplast processes, such as photosynthesis, de novo fatty acid biosynthesis, and the metabolism of some amino acids.

While it includes a diverse number of measurements, the Chloroplast 2010 dataset poses some challenges to traditional analyses for cross-dataset comparisons and

community analysis. As with many high throughput screening datasets, it lacks the replication and controls needed to minimize the impact of technical variance using more traditional analysis methods that would enable using the data to address broader questions. Furthermore, the dataset contains a fair amount of missing data (due to plants being unavailable for analysis or assays that failed to pass quality control), and large amounts of uninformative data where the mutant lines have no discernible phenotype in the assays measured. MIPHENO (Bell et al., 2012), a method developed for the analysis of large scale screening data, has been shown to minimize the technical variance in datasets with no controls or replication.

The MIPHENO workflow transformed this dataset into one that enables cross dataset comparison (Chapter 2). To address issues of missing and high background data, SimMeasure (Chapter 3) can be used to calculate the similarity between individual mutant lines. With proper processing, the Chloroplast 2010 dataset enables an omics approach to functional gene annotation similar to the ways in which transcriptomics have been used except that the observed characteristics (altered metabolites) are more closely related to the changes in physiology.

Materials and methods

4.3.1 Data preparation

Data corresponding to the Col-0 ecotype from the Chloroplast 2010 project were processed using MIPHENO (Bell et al., 2012) as described. Briefly, a *post hoc* quality control step was used to identify assay groups with a median response that was three

median adjusted deviations from the global response for that metabolite. Groups were then normalized, based on the group median, to the global median. A cumulative distribution function was applied to each metabolite measured to determine the probability of a response being as or more extreme than the observation. The resulting score indicates the probability that the observation represents a 'mutant' phenotype. Leaf free fatty acid and free amino acid data along with seed free amino acid and seed percent carbon and nitrogen data were processed separately by tissue source, treating each plant as an individual.

After calculation of the MIPHENO score (Chapter 2; Bell et al., 2012), individuals with genotyping information indicating they were wild-type or heterozygous for the insert were removed and remaining samples used to determine the score for that allele (e.g. SALK line), by taking the median MIPHENO score. Metabolite data from the Chloroplast 2010 dataset were available in two forms. Mole percent (MP) is calculated as the quotient of moles of a given attribute (e.g. a specific leaf amino acid or fatty acid methyl ester) over the total moles for the individual in that assay set (e.g. the sum of all leaf amino acids or fatty acid methyl esters for that individual) times one hundred. Fresh weight (FW), is calculates as the number of moles of that metabolite per gram tissue fresh weight (Lu et al., 2008; Lu et al., 2011b). Data from each representation (MP or FW) were combined into separate data frames. Seed carbon, nitrogen, and carbon to nitrogen ratio were only available as percentage values and not as MP or FW and thus the same value is present in each dataset.

4.3.2 Community identification and characterization

MIPHENO scores were transformed from the 0 to 1 range to a -1 to 1 range to facilitate processing; values approaching -1 are highly likely to have a decreased amount of

metabolite compared with the median and those approaching +1 are highly likely to have an increased amount of the metabolite (see Appendix 1, Chapter 4 supplementary files for details). Across both datasets (MP, FW), scores between -0.6 to 0.6 made up 80% of the data, and scores between -0.75 to 0.75 correspond to 90% of the data. Similarity scores were then calculated using SimMeasure (Chapter 3) to identify lines with similar responses across the measured metabolites using thresholds of 0.6 or 0.75 to remove background responses. SimMeasure calculates the similarity between two lines, controlling for instances where one individual has few responses while the other has many (i.e. lines that have a simple vs. pleiotropic phenotype). Values from SimMeasure are in the same range as correlation coefficients (-1 to 1), and can be interpreted the same way for clustering purposes. As graphical-based clustering approaches resulted in a complex network with loss of nodes at an edge cutoff of 0.75 (Figure 11), a hierarchical approach was used to create more well-defined communities. Hierarchical clustering of the adjacency matrix with Ward's minimum variance was used to define 400 communities. Four hundred was chosen to maximize the number of clusters having fewer than twenty individuals while minimizing those with fewer than four members.

Gene ontology (GO) enrichment for biological processes was used to help identify communities having characteristics that would make them more amenable to building hypotheses on under-annotated loci. The enrichment of GO terms in communities was calculated using a modified version of the GO enrichment functions (Horan et al., 2008). For the Chloroplast 2010 dataset, the alleles in the dataset were used as the background in creating the GO reference to account for cases where multiple alleles of a locus were

present. This has the effect of controlling the expectation of a locus that has four alleles verses the locus that has only one. Ontology libraries were built using the GO file from The Arabidopsis Information Resource (TAIR; www.arabidopsis.org) dated January 17, 2012 (Rhee et al., 2003). Phenotypes identified based on the MIPHENO scores were then compared for clusters with significant GO term enrichment (p value > 0.5).

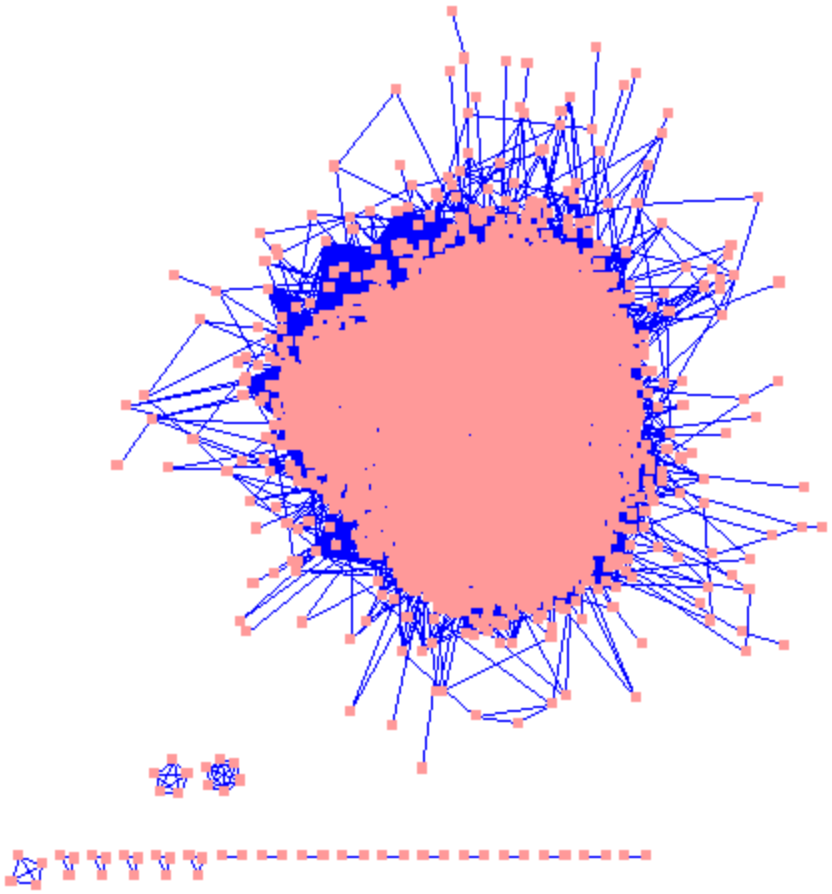


Figure 11 Graphical clustering of FW data

The adjacency matrix was converted to a graph object using a threshold of 0.75. A total of 98 loci were lost as unconnected nodes. While a majority of individuals belong to the main cluster, a few segregated out into smaller communities. These smaller communities contained both positively and negatively connected individuals, but with little information to drive hypothetical models of unknowns.

Results

Comparison of the phenotypic clustering between the MP and FW datasets highlights differences in the nature of the measurements (Figures 12, 13). Clustering based on MP (Figure 12) results in the assays being mixed with no clear separation between tissue (leaf or seed) on which the assay was performed. Additionally, there is no appearance of biosynthetically-related metabolites, for example the branched chain amino acids, clustering closer together than those metabolically distant. These two observations suggest that the MP data may not reflect underlying metabolic relationships. One would expect that biosynthetically-related metabolites would respond more similarly due to metabolic control, which should be reflected in at the tissue/assay level as well as at the metabolite level. This suggests that the MP data may not be representing the metabolite relationships adequately enough to allow inferences on the impact of a gene knockout. By contrast, the FW results (Figure 13) resulted in different assays being grouped together and a clear separation between the tissues, with some minor exceptions (seed carbon and the carbon: nitrogen ratio). Additionally, biosynthetically-related compounds are clustered close to each other.

Based on these initial results and the goal of developing functional hypotheses based on metabolic phenotypes, only the FW data were used for further analysis. The clusters based on a 0.60 threshold (corresponding to 20% of the data, Figure 13) and a 0.75 threshold (corresponding to 10% of the data, Figure 14) were used in the enrichment analyses. Terms relating to biological process were the focus as those were most easily reconcilable with metabolic phenotypes. Ninety-one unique clusters had enrichment in GO

biological process using a threshold of 0.60, while 94 unique clusters were identified at a threshold of 0.75. Selected results from each of these thresholds are shown in Tables 3 through 6.

Figure 12 Clustering of MP Data using a threshold of 0.6

Mole Percent data was processed using SimMeasure then clustered using hierarchical clustering with Wards minimum variance by individual (y-axis) or by assay (x-axis). Color labels on y-axis correspond to the 400 different communities. Colors within the plot reflect the accumulation of metabolites (darkest blue indicates those observations approaching -1 while darkest red indicates observations approaching 1. White indicates values in the designated wild type range of -0.6 to 0.6), with metabolites having an MIPHENO score between -0.6 and 0.6 removed. Observations closer to -1 have a high probability of being a low accumulator while those closer to 1 have a higher probability of being a high accumulator of the metabolite. Note that the similar assays (denoted with X, LF or SD) do not cluster together completely. LF=leaf, SD=seed, X=free fatty acid.

Figure 12 (cont'd)

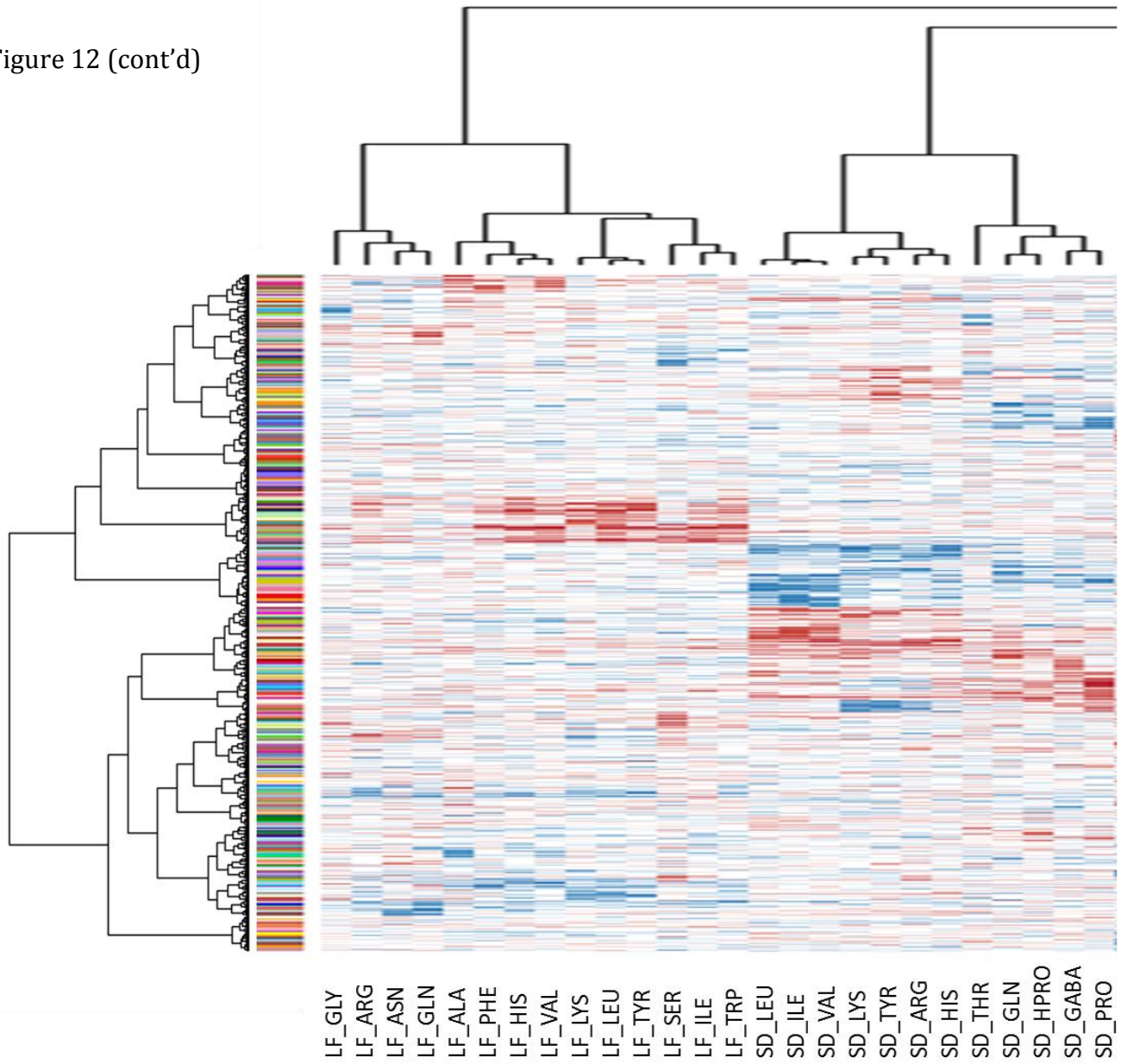


Figure 12 (cont'd)

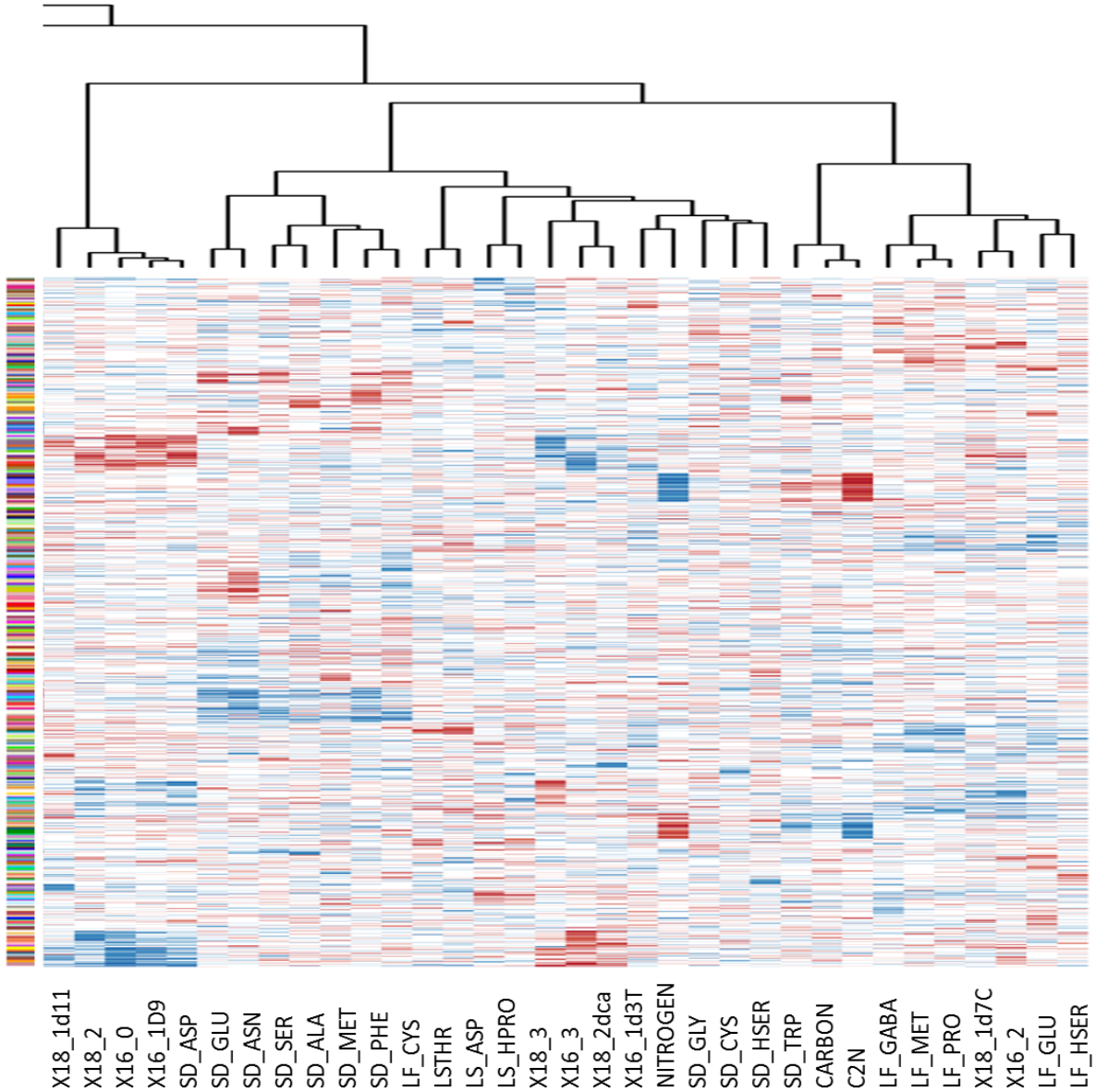


Figure 13 Clustering of FW Data using a threshold of 0.6

Fresh weight data was processed using SimMeasure then clustered using hierarchical clustering with Wards minim variance by individual (y-axis) or by assay (x-axis). Color labels on y-axis correspond to the 400 different communities. Colors within the plot reflect the accumulation of metabolites (darkest blue indicates those observations approaching -1 while darkest red indicates observations approaching 1. White indicates values in the designated wild type range of -0.6 to 0.6), with metabolites having an MIPHENO score between -.06 and 0.6 removed. Observations closer to -1 have a high probability of being a low accumulator while those closer to 1 have a higher probability of being a high accumulator of the metabolite. Note that similar assays appear to cluster together, which contrasts to results in Figure 11. LF=leaf, SD=seed, X=free fatty acid.

Figure 13 (cont'd)

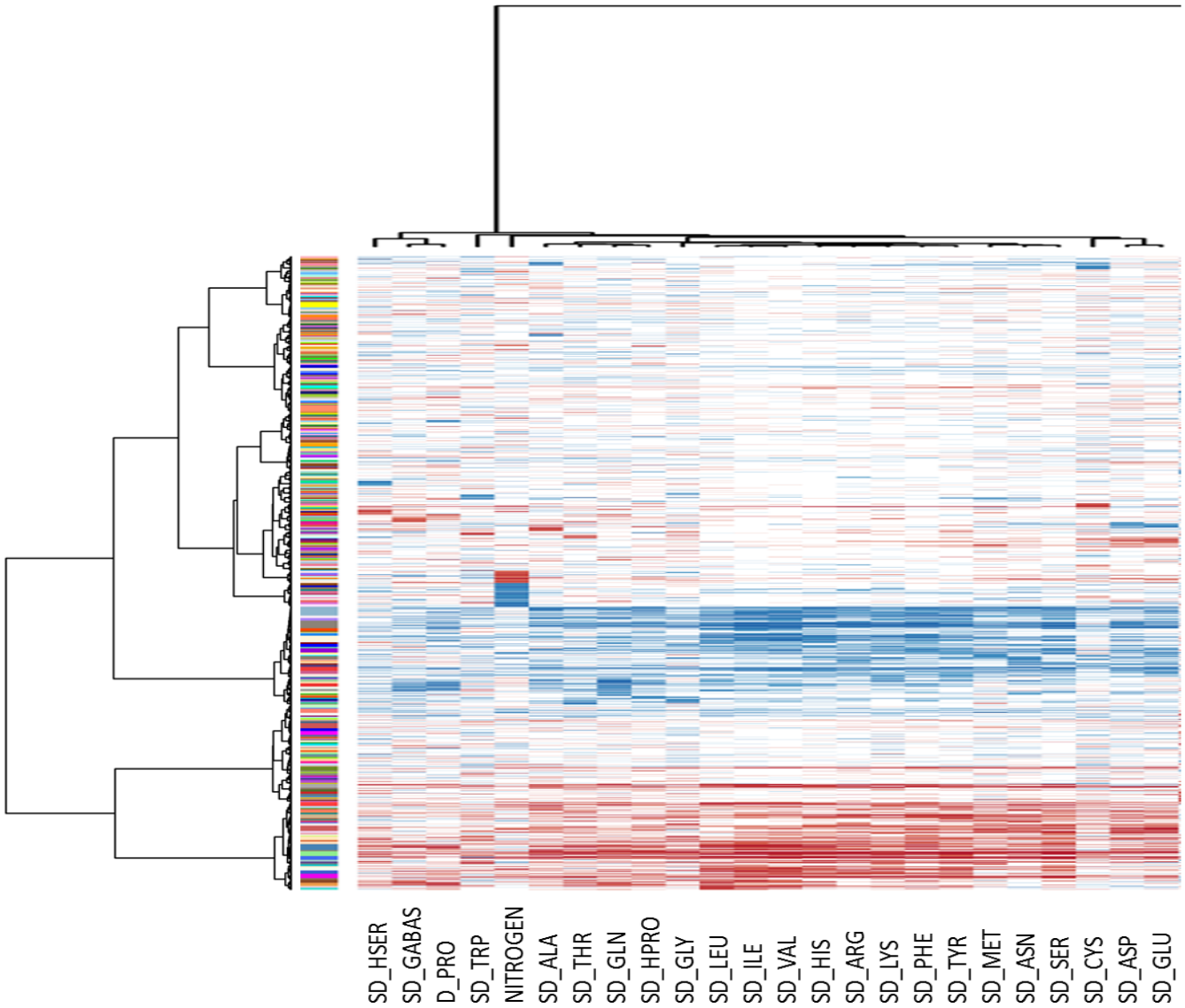


Figure 13 (cont'd)

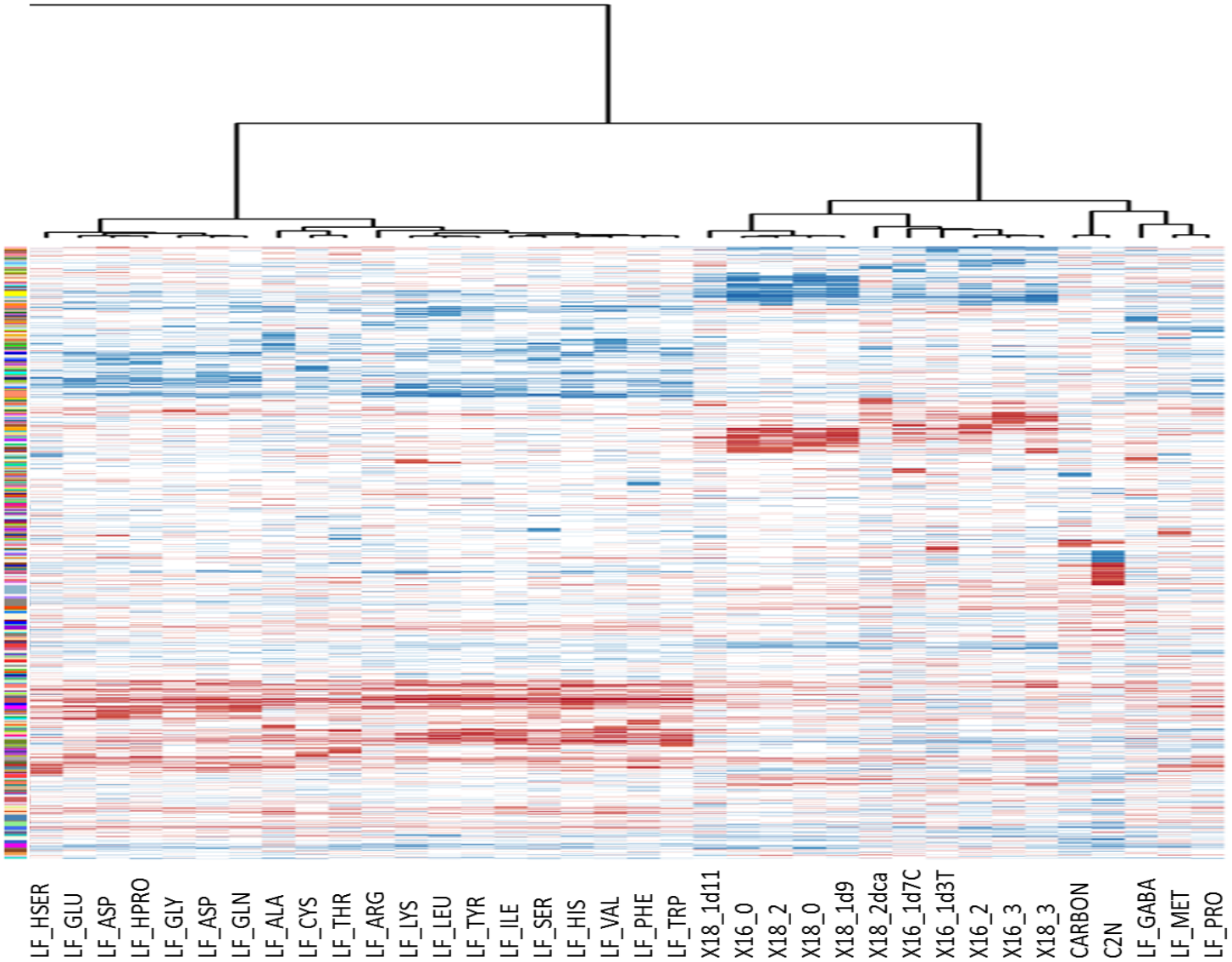


Figure 14 Clustering of FW Data using a threshold of 0.75

Fresh weight data was processed using SimMeasure then clustered using hierarchical clustering with Wards minim variance by individual (y-axis) or by assay (x-axis). Colors within the plot reflect the accumulation of metabolites (darkest blue indicates those observations approaching -1 while darkest red indicates observations approaching 1. White indicates values in the designated wild type range of -0.6 to 0.6) with metabolites having an MIPHENO score between -0.75 and 0.75 removed. Observations closer to -1 have a high probability of being a low accumulator while those closer to 1 have a higher probability of being a high accumulator of the metabolite. LF=leaf, SD=seed, X=free fatty acid.

Figure 14 (cont'd)

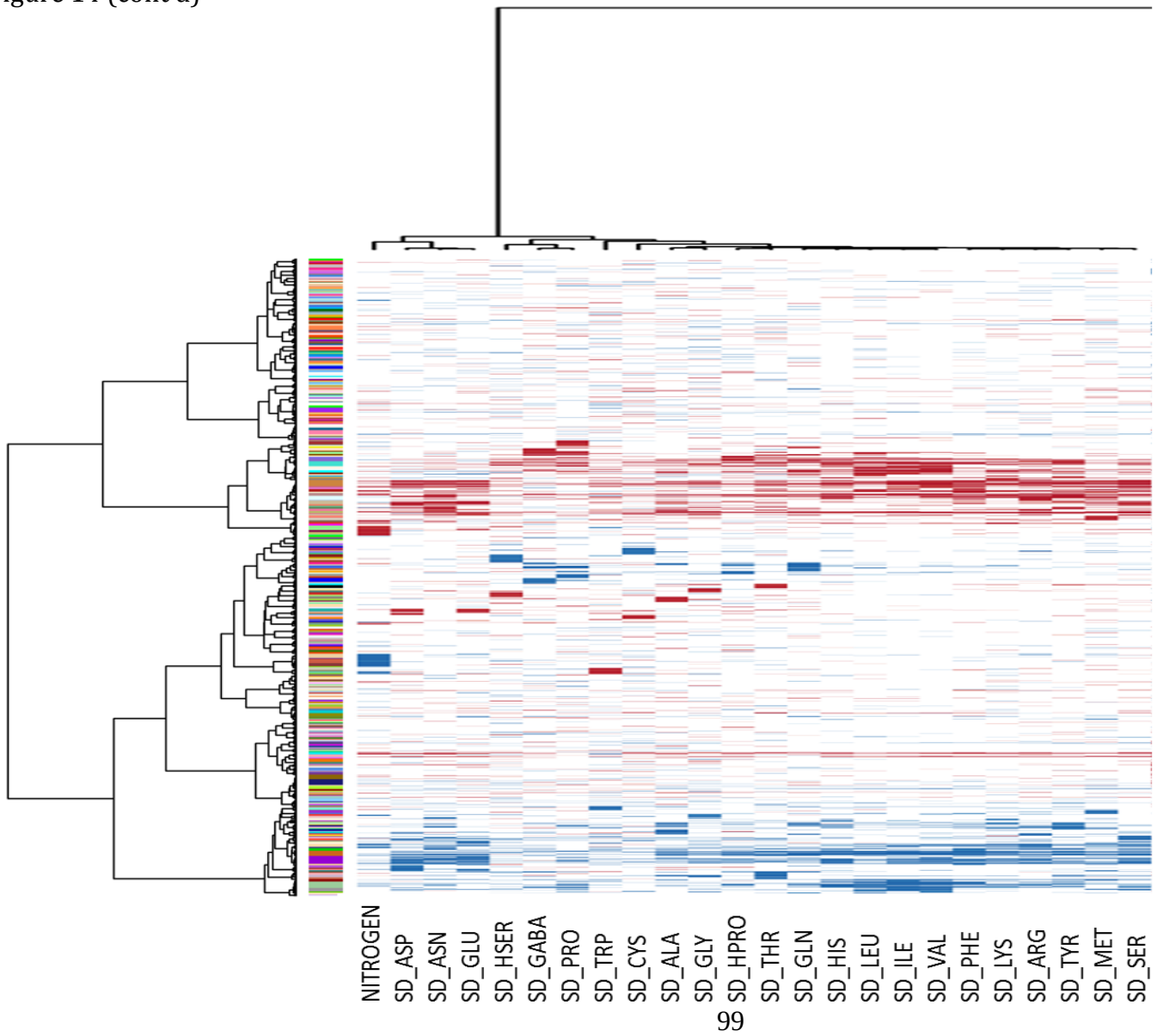


Figure 14 (cont'd)

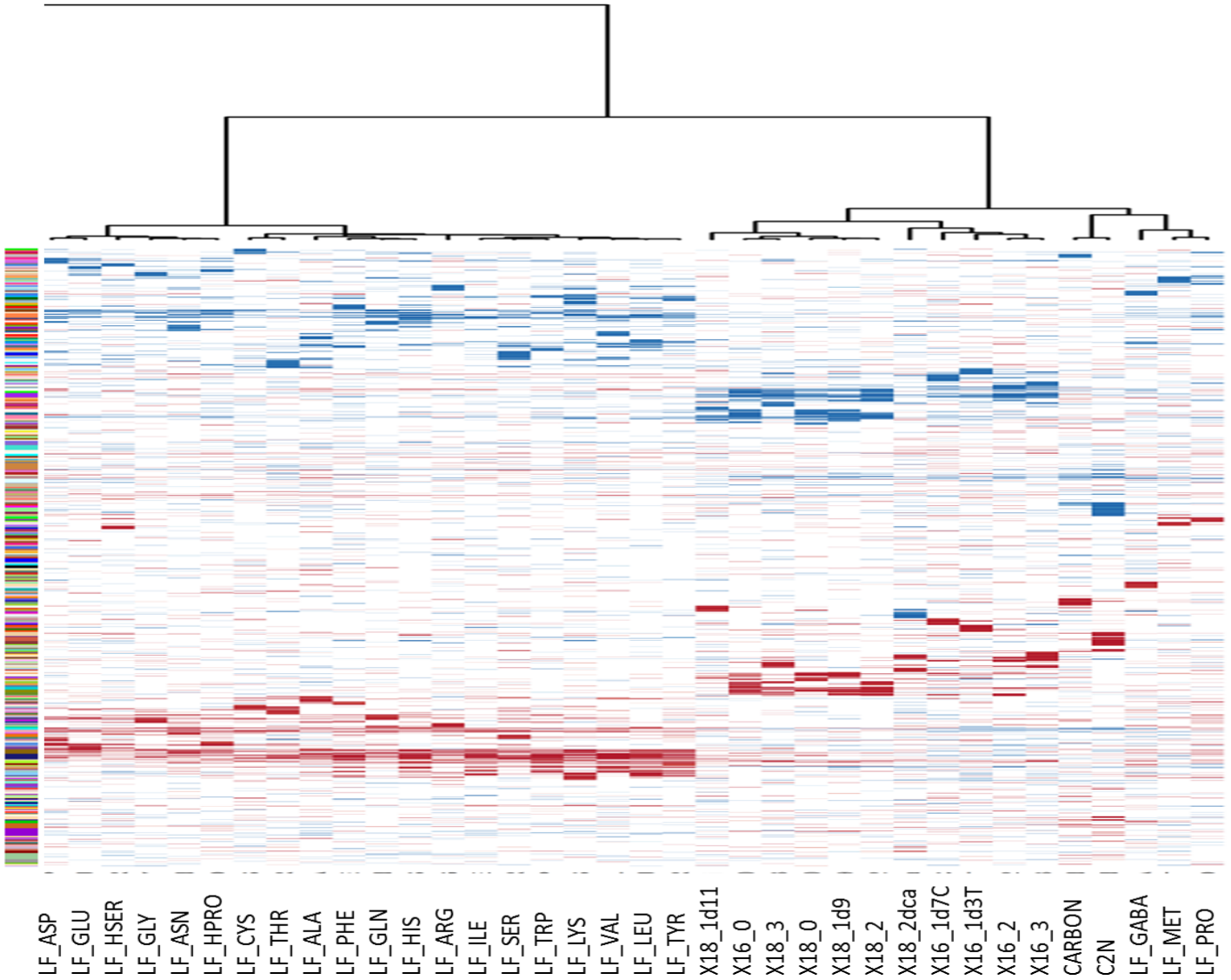


Table 3 Cluster: roseybrown4, Threshold=0.6, GO: photosynthesis

Locus	Insertion Line	TAIR Annotation
AT1G14150	SALK_006106	Encodes a subunit of the NAD(P)H dehydrogenase complex located in the chloroplast thylakoid lumen
AT1G48280	SALK_080607	Hydroxyproline-rich glycoprotein
AT1G48950	SALK_046110	C3HC zinc finger-like
AT1G75010	EMS16	Encodes ARC3 (Accumulation and Replication of Chloroplast 3), a chloroplast division factor functioning in the initiation of chloroplast division
AT2G26340	SALK_048391	Unknown protein
AT2G39470	SALK_063049	PsbP-like protein 2 (PPL2)
AT3G10470	SALK_106077	C2H2-type zinc finger family protein
AT4G29670	SALK_028498	Encodes a member of the thioredoxin family protein
AT5G13310	SALK_047869	Unknown protein

Table 4 Cluster: gold, Threshold=0.6, GO: branch chain amino acid family process

Locus	Insertion Line	TAIR Annotation
AT3G16890	SALK_019082	Encodes a mitochondrial pentatricopeptide repeat (PPR) domain protein
AT3G45300	ivd1-2	Encodes isovaleryl-coenzyme a dehydrogenase
AT5G05740	SALK_001991	S2P-like putative metalloprotease
AT5G23010	SALK_116223	Encodes a methylthioalkylmalate synthase
AT5G65770	SALK_097945	LITTLE NUCLEI4 (LINC4)
AT5G65780	SALK_071486	Encodes a chloroplast branched-chain amino acid aminotransferase
AT5G66380	SALK_011184	Encodes a folate transporter that is located in the chloroplast envelope

Table 5 Cluster: blue3, Threshold=0.75, GO: fatty acid metabolic process

Locus	Insertion Line	TAIR Annotation
AT1G08640	SALK_032130	Chloroplast J-like domain 1 (CJD1)
AT1G08640	SALK_039694	Chloroplast J-like domain 1 (CJD1)
AT1G78110	SALK_003000	Unknown protein
AT2G28540	SALK_152456	RNA binding (RRM/RBD/RNP motifs) family protein
AT4G13590	SALK_129037	Uncharacterized protein family (UPF0016)
AT4G30950	SALK_027548	Chloroplastic enzyme responsible for the synthesis of 16:2 and 18:2 fatty acids from galactolipids, sulpholipids and phosphatidylglycerol

Table 6 Cluster: violetred, Threshold=0.75, GO: amino acid biosynthesis

Locus	Insertion Line	TAIR Annotation
AT1G47510	SALK_108673	Encodes a phosphatidylinositol polyphosphate 5-phosphatase
AT3G04940	SALK_092696	Encodes cysteine synthase CysD1
AT4G19710	SALK_019023	Encodes a bifunctional aspartate kinase/homoserine dehydrogenase
AT4G19710	SALK_082155	Encodes a bifunctional aspartate kinase/homoserine dehydrogenase
AT5G57940	SALK_149893	Member of cyclic nucleotide gated channel family
AT5G62140	SALK_113654	Unknown protein

Discussion

The major aim of this work was to test the proof of concept that high throughput phenotypic studies could be used to develop hypotheses regarding the function of un- or under annotated genes. Because of the diversity of metabolites and genes under investigation, a secondary aim was to see if additional hypotheses could be built regarding new roles for genes with existing annotation or relationships between the metabolic networks in which they function. The results indicate that for these purposes, it is best to use measurements that are directly reflective of metabolite quantity and not a relative quantity such as mole percent that might be artificially changed due to the change in another measured value. While in large-scale studies there may be some quality issues in using an absolute measurement such as fresh weight due to the amount of cellular water and metabolite stability, these are generally small and the effects can be minimized with normalization to remove technical error. In contrast, differences based on a percent measurement can vary wildly due to the relative impact when a low -abundance metabolite increases many fold or the amount of a usually high accumulating metabolite is reduced by a large fraction. The following examples from the analysis reflect the FW data.

The first example cluster at a threshold of 0.60, 'roseybrown4', is enriched in mutants of genes associated with GO annotation related to photosynthesis with a phenotypic signature of high leaf glycine and glutamine (Table 3). The GO annotation enrichment is based on two genes, At1g14150 (PnsL2) and At2g39470 (PnsL1), which are components of the chloroplast NADH dehydrogenase-like complex (Ifuku et al., 2011; Suorsa et al., 2010). The other seven genes in the cluster include ARC3, which is involved in

chloroplast morphology, and an unknown gene, At2g26340. Other data from the Chloroplast 2010 project shows that the insertion line corresponding to At2g26340 is annotated as having a positive before-high-light Fv/Fm, indicative of a change in the quantum efficiency of photosystem II. This protein is also annotated as present in the thylakoid lumen (Friso et al., 2004; Peltier et al., 2004), which supports a possible involvement with photosynthetic complexes. The lack of a known catalytic domain in At2g26340 (Rhee et al, 2003) and the mild phenotype under normal conditions indicates it may have a regulatory or structural role with the photosystem complexes (Figure 15), interacting transiently or on the periphery.

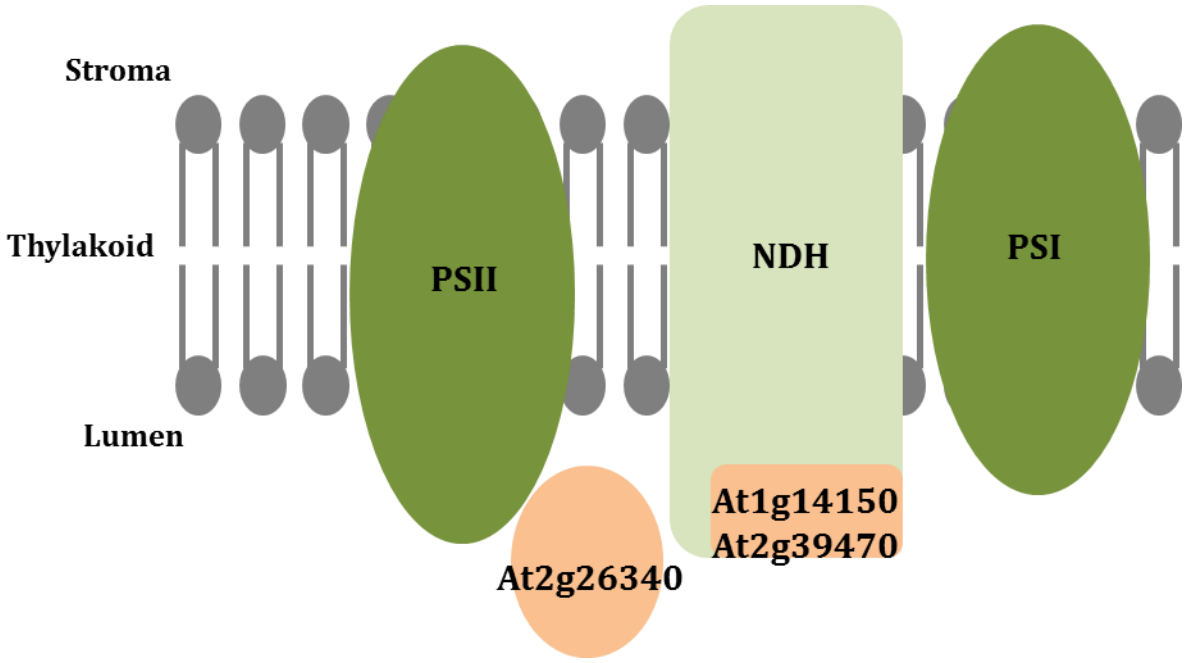


Figure 15 Hypothetical model for At2g26340

The unknown protein, At2g26340, is proposed to play a structural or regulatory role with respect to the photosystem complexes. The similarity of At2g26340's phenotype to two of the proteins in the NDH complex and the weak high before high light FV/FM measurements could indicate that it interacts with PSII or with NDH. PSII = photosystem II, NDH = chloroplast NADH dehydrogenase-like complex, PSI = photosystem I

Gold, the second cluster example using a threshold of 0.60, is enriched in genes involved in branched-chain amino acid family processes (Table 4). Two genes, At5g65780 (ATBCAT-5) and At3g45300 (IVD1) are involved in the degradation of branched-chain amino acids (Diebold et al., 2002; Gu et al., 2010). The T-DNA insertion line Salk_097945 is annotated to a third gene in the cluster, At5g65770. This insertion line actually lies in between At5g6770 and At5g65780, suggesting that it is likely disrupting the function of the downstream locus. Another member of this community is At5g23010 (MAM1) a methylthioalkylmalate synthase, which is involved in methionine chain elongation and is similar to 2-isopropylmalate synthase involved in leucine biosynthesis (Field et al., 2004; Kroymann et al., 2001). Possible hypotheses for the apparent phenotype is altered activity of MAM3 (Textor et al., 2007) caused by the disruption of MAM1, or a role for MAM1 in the biosynthesis of branched-chain amino acid derived glucosinolates.

Using the more stringent threshold of 0.75, the cluster blue3 stands out as an example of a community driven by a free fatty acid phenotype (Table 5). The GO annotation of fatty acid metabolic process is driven by At4g30950 (Fad6) and two alleles of At1g08640 (CJD1), both localized to the chloroplast envelope (Ajjawi et al., 2011; Nandi et al., 2003; Schmidt et al., 1994). The cluster phenotype of increased 16:1d7C and decreased 16:2, 16:3, and 18:3 fatty acids are consistent with a defect in fatty acid metabolism and published phenotypes for knockouts of these genes. The other three members of this cluster are largely unknown. At4g13590, is located to the inner chloroplast membrane (Ferro et al., 2003). Its location within the cell puts it near At1g08640 and At4g30950

suggesting that it may have a role in fatty acid metabolism, likely that of a regulatory or structural protein (Figure 16) due to the lack of evidence for a catalytic domain but support for a transmembrane domain (Rhee et al, 2003).

Violetred, Table 6, is a cluster also found using the threshold of 0.75, is enriched in genes involved in amino acid biosynthesis. It includes two alleles of At4g19710 (encoding the bifunctional aspartate kinase-homoserine dehydrogenase II) which is involved in the synthesis of threonine, isoleucine and methionine from homoserine (Curien et al., 2005; Ghislain et al., 1994) and one of At3g04940 (CYSD1), which catalyzes the synthesis of cysteine (Hatzfeld et al., 2000; Yamaguchi et al., 2000). This cluster has the phenotypic signature of high leaf cysteine, glutamate, homoserine, and threonine. Part of the phenotype seen may be due to the allosteric regulation of the proteins (or those upstream in the case of CYSD1) and genetic redundancy compensating for the loss of a dominate enzyme. The other three loci in the cluster have no evidence for their involvement in amino acids (phosphatidylinositol polyphosphate 5-phosphatase, At1g47510; cyclic-nucleotide-gated-channel-family protein, At5g57940; unknown protein, At5g62140). For a completely unknown protein such as At5g62140, analyses such as this may serve as a starting point in trying to characterize the locus, for example by carrying out a more precise measurement of leaf amino acids.

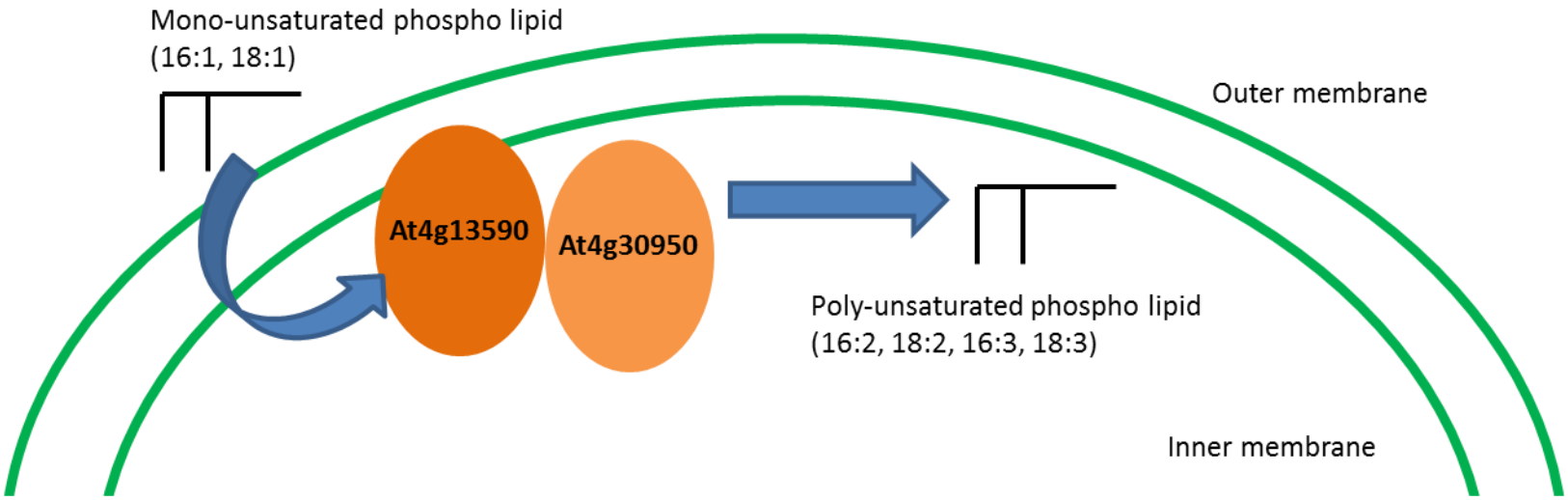


Figure 16 Hypothetical model for At4g13590

Given the lack of known domain for catalysis, it is hypothesized that At4g13590 serves as a regulatory role either in the transport of mono-unsaturated lipids or in conjuncture with At4g30950 (FAD6). It is likely positioned at or within the inner membrane based on predictions for transmembrane domains. From this position it is hypothesized to interact with other proteins involved in the biosynthesis of desaturated fatty acids.

Multiple insertion lines targeting genes involved in branched-chain amino acid degradation are present in the Chloroplast 2010 dataset. These include two alleles for a hydroxymethylglutaryl-CoA lyase (At2g26800), an isovaleryl-CoA dehydrogenase (At3g45300), a member of the branched-chain amino acid transferase family, BCAT5 (At5g65780), and MCCA/MCCB (At1g03090 and At4g34030, respectively), which are the subunits of the methylcrotonoyl-CoA carboxylase. While they all show a similarity in their phenotype, specifically increases in branched-chain amino acids, they cluster to different groups (phenotypes using a 0.75 threshold shown in Table 7). This difference is likely due to the pleiotropic nature of the phenotype and the complex nature of the metabolism (Figure 17).

This work presents one of the first examples of using high throughput phenomics to develop hypotheses about gene function. Initial results provide examples of how large-scale analyses of knockouts can help develop hypotheses about gene function (roseybrown4, blue3) or suggest relationships between metabolic sub networks (gold, violetred). As the prevalence of high throughput experiments increases, it may be possible to combine phenotypic information across several datasets to help answer questions about biology in a way that researchers have been able to using transcriptomics data. Additionally, integration of transcriptomics data with the phenotypic data may help expand the communities and provide information on redundant genes or genes that were not assayed in one of the datasets.

Table 7 Phenotype and cluster assignment for branched-chain amino acid degradation loci*

Locus	At5g65780	At3g45300	At1g03090	At4g34030	At2g26800(A)	At2g26800(B)
Cluster	cyan1	cyan1	orangered 4	orangered 4	lightcyan	lightcyan
Step	1	2	3	3	4	4
Nitrogen	0.94	0.89	-0.67	NA	-0.91	NA
Alanine	NA	0.9	NA	-0.96	0.97	0.86
Arginine	0.99	1	0.66	NA	0.99	1
Asparagine	NA	-0.96	-0.88	-0.92	0.95	NA
Aspartate	-0.76	NA	NA	NA	0.82	NA
Cysteine	0.83	0.97	NA	-0.91	0.99	0.64
GABA	0.84	0.96	NA	NA	0.85	0.97
Glutamine	0.99	0.98	NA	NA	0.9	0.86
Glutamate	-0.69	-0.99	-1	-1	NA	0.72
Glycine	NA	0.99	NA	-0.93	0.98	0.82
Hydroxy proline	0.98	NA	NA	NA	0.87	NA
Homo serine	1	0.89	NA	NA	NA	0.68

*Cluster refers to the clustering based on a threshold of 0.75. Step refers to the step highlighted in Figure 17. Values are the MIPHENO score, shaded to project the magnitude of the phenotype (red= high probability of a high accumulator phenotype, blue= high probability of a low accumulator phenotype).

Table 7, (cont'd)

Locus	At5g65780	At3g45300	At1g03090	At4g34030	At2g26800(A)	At2g26800(B)
Cluster	cyan1	cyan1	orangered 4	orangered 4	lightcyan	lightcyan
Step	1	2	3	3	4	4
Histidine	1	1	0.99	0.98	1	1
Isoleucine	1	1	1	1	1	1
Leucine	1	1	1	1	1	1
Lysine	0.99	1	0.71	NA	0.98	0.95
Methionine	0.99	1	0.95	NA	1	0.98
Phenyl-alanine	0.91	1	0.88	NA	0.84	NA
Proline	1	0.95	NA	NA	0.61	0.66
Serine	0.99	0.99	0.95	0.95	0.99	0.97
Threonine	0.74	0.92	NA	-0.91	NA	NA
Tryptophan	0.73	1	NA	NA	1	0.94
Tyrosine	0.94	1	0.6	NA	1	0.97
Valine	1	1	1	0.99	1	1

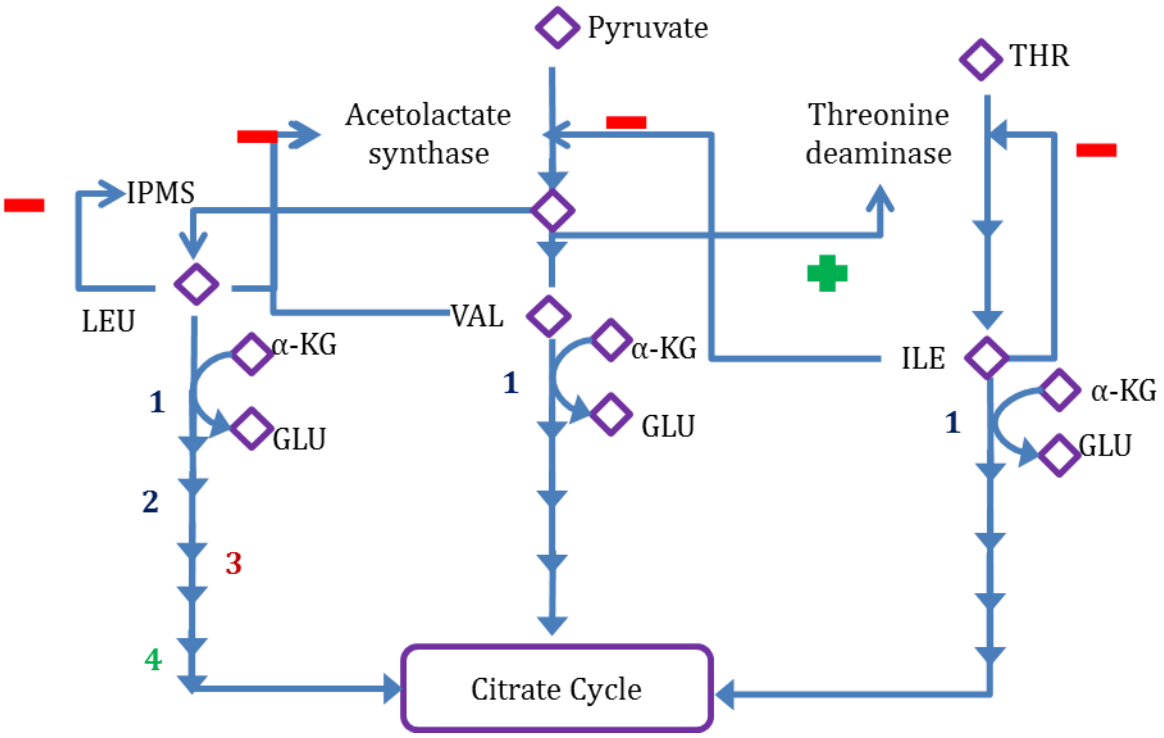


Figure 17 Schematic of branched-chain amino acid metabolism

Steps 1-4 correspond to the loci in Table 7. Minus (-) indicates feedback inhibition whereas positive (+) indicates feedback activation. α-KG= alpha ketoglutarate, GLU=glutamate ILE= isoleucine, IPMS=isopropylmalate synthase, LEU=leucine, THR=threonine, VAL= valine

REFERENCES

References

Ajjawi, I., Coku, A., Froehlich, J.E., Yang, Y., Osteryoung, K.W., Benning, C., and Last, R.L. (2011). A J-like protein influences fatty acid composition of chloroplast lipids in Arabidopsis. PLoS ONE *6*, e25368.

Ajjawi, I., Lu, Y., Savage, L.J., Bell, S.M., and Last, R.L. (2010). Large-scale reverse genetics in Arabidopsis: case atudies from the Chloroplast 2010 project. Plant Physiology *152*, 529-540.

Alonso, J.M., Stepanova, A.N., Leisse, T.J., Kim, C.J., Chen, H., Shinn, P., Stevenson, D.K., Zimmerman, J., Barajas, P., Cheuk, R., et al. (2003). Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. Science *301*, 653-657.

Bell, S.M., Burgoon, L.D., and Last, R.L. (2012). MIPHENO: data normalization for high throughput metabolite analysis. BMC Bioinformatics *13*, 10.

Bouché, N., and Bouchez, D. (2001). Arabidopsis gene knockout: phenotypes wanted. Current Opinion in Plant Biology *4*, 111-117.

Curien, G., Ravanel, S., Robert, M., and Dumas, R. (2005). Identification of six novel allosteric effectors of *Arabidopsis thaliana* aspartate kinase-homoserine dehydrogenase isoforms. Journal of Biological Chemistry *280*, 41178-41183.

Diebold, R., Schuster, J., Däschner, K., and Binder, S. (2002). The branched-chain amino acid transaminase gene family in Arabidopsis encodes plastid and mitochondrial proteins. Plant Physiology *129*, 540-550.

Ferro, M., Salvi, D., Brugière, S., Miras, S., Kowalski, S., Louwagie, M., Garin, J., Joyard, J., and Rolland, N. (2003). Proteomics of the chloroplast envelope membranes from *Arabidopsis thaliana*. Molecular & Cellular Proteomics *2*, 325-345.

Field, B., Cardon, G., Traka, M., Botterman, J., Vancanneyt, G., and Mithen, R. (2004). Glucosinolate and amino acid biosynthesis in Arabidopsis. Plant Physiology *135*, 828-839.

Friso, G., Giacomelli, L., Ytterberg, A.J., Peltier, J.-B., Rudella, A., Sun, Q., and Wijk, K.J.v. (2004). In-depth analysis of the thylakoid membrane proteome of *Arabidopsis thaliana* chloroplasts: new proteins, new functions, and a plastid proteome database. The Plant Cell Online *16*, 478-499.

Ghislain, M., Frankard, V., Vandenbossche, D., Matthews, B.F., and Jacobs, M. (1994). Molecular analysis of the aspartate kinase-homoserine dehydrogenase gene from *Arabidopsis thaliana*. Plant Molecular Biology *24*, 835-851.

Gu, L., Jones, A.D., and Last, R.L. (2010). Broad connections in the Arabidopsis seed metabolic network revealed by metabolite profiling of an amino acid catabolism mutant. The Plant Journal *61*, 579-590.

Hatzfeld, Y., Maruyama, A., Schmidt, A., Noji, M., Ishizawa, K., and Saito, K. (2000). β-cyanoalanine synthase is a mitochondrial cysteine synthase-like protein in spinach and Arabidopsis. Plant Physiology *123*, 1163-1172.

Horan, K., Jang, C., Bailey-Serres, J., Mittler, R., Shelton, C., Harper, J.F., Zhu, J.-K., Cushman, J.C., Gollery, M., and Girke, T. (2008). Annotating genes of known and unknown function by large-scale coexpression analysis. Plant Physiology *147*, 41-57.

Ifuku, K., Endo, T., Shikanai, T., and Aro, E.-M. (2011). Structure of the chloroplast NADH dehydrogenase-like complex: nomenclature for nuclear-encoded subunits. Plant and Cell Physiology *52*, 1560-1568.

Kroymann, J., Textor, S., Tokuhisa, J.G., Falk, K.L., Bartram, S., Gershenzon, J., and Mitchell-Olds, T. (2001). A gene controlling variation in Arabidopsis glucosinolate Composition Is part of the methionine chain elongation pathway. Plant Physiology *127*, 1077-1088.

Lu, Y., Savage, L.J., Ajjawi, I., Imre, K.M., Yoder, D.W., Benning, C., DellaPenna, D., Ohlrogge, J.B., Osteryoung, K.W., Weber, A.P., et al. (2008). New connections across pathways and cellular processes: industrialized mutant screening reveals novel associations between diverse phenotypes in Arabidopsis. Plant Physiology *146*, 1482-1500.

Lu, Y., Savage, L.J., Larson, M.D., Wilkerson, C.G., and Last, R.L. (2011a). Chloroplast 2010: a database for large-scale phenotypic screening of Arabidopsis mutants. Plant Physiology *155*, 1589-1600.

Lu, Y., Savage, L.J., and Last, R.L. (2011b). Chloroplast phenomics: systematic phenotypic screening of chloroplast protein mutants in Arabidopsis. In Chloroplast Research in Arabidopsis : Methods and Protocols, Volume II, R.P. Jarvis, ed. (NY: Humana Press), pp. 161-185.

Nandi, A., Krothapalli, K., Buseman, C.M., Li, M., Welti, R., Enyedi, A., and Shah, J. (2003). Arabidopsis sfd mutants affect plastidic lipid composition and suppress dwarfing, cell death, and the enhanced disease resistance phenotypes resulting from the deficiency of a fatty acid desaturase. The Plant Cell Online *15*, 2383-2398.

Peltier, J.-B., Ytterberg, A.J., Sun, Q., and van Wijk, K.J. (2004). New functions of the thylakoid membrane proteome of *Arabidopsis thaliana* revealed by a simple, fast, and versatile fractionation strategy. Journal of Biological Chemistry *279*, 49367-49383.

Rhee, S.Y., Beavis, W., Berardini, T.Z., Chen, G., Dixon, D., Doyle, A., Garcia-Hernandez, M., Huala, E., Lander, G., Montoya, M., et al. (2003). The Arabidopsis information resource (TAIR): a model organism database providing a centralized, curated

gateway to Arabidopsis biology, research materials and community. Nucleic Acids Research *31*, 224-228.

Schmidt, H., Dresselhaus, T., Buck, F., and Heinz, E. (1994). Purification and PCR-based cDNA cloning of a plastidial n-6 desaturase. Plant Molecular Biology *26*, 631-642.

Suorsa, M., Sirpiö, S., Paakkarinen, V., Kumari, N., Holmström, M., and Aro, E.-M. (2010). Two proteins homologous to PsbQ are novel subunits of the chloroplast NAD(P)H dehydrogenase. Plant and Cell Physiology *51*, 877-883.

Textor, S., de Kraker, J.-W., Hause, B., Gershenzon, J., and Tokuhisa, J.G. (2007). MAM3 catalyzes the formation of all aliphatic glucosinolate chain lengths in Arabidopsis. Plant Physiology *144*, 60-71.

The Arabidopsis Information Resource (2010). TAIR - Portals - Genome Snapshot. [http://www.arabidopsis.org/portals/genAnnotation/genome_snapshot.jsp]

Vandepoele, K., Quimbaya, M., Casneuf, T., De Veylder, L., and Van de Peer, Y. (2009). Unraveling transcriptional control in Arabidopsis using cis-regulatory elements and coexpression networks. Plant Physiology *150*, 535-546.

Williams, E.J.B., and Bowles, D.J. (2004). Coexpression of neighboring genes in the genome of *Arabidopsis thaliana*. Genome Research *14*, 1060-1067.

Yamaguchi, Y., Nakamura, T., Kusano, T., and Sano, H. (2000). Three Arabidopsis genes encoding proteins with differential activities for cysteine synthase and β-cyanoalanine synthase. Plant and Cell Physiology *41*, 465-476.

Chapter 5

Conclusions

Despite the information available for the plant model species *Arabidopsis thaliana*, a larger portion of the genome is un- or under annotated. This creates challenges for researchers as they try to understand processes within the organism and engineer or breed plants to address new needs. One possible source for annotation is leveraging the information from large-scale studies, such as mutant screens. To date, there have been no such studies in any organism that have looked at the use of phenotypic mutant screens to build models for gene annotation.

The key objective of the work presented in the dissertation was to test the hypothesis that large-scale, high-throughput phenotypic screening data could be used in developing functional gene annotation. To achieve the objective, novel methods were needed to use screening data for community analysis. This is because large-scale screens tend to not meet the requirements for standard normalization and cross-dataset comparison methods.

Major accomplishments

The key objective of the dissertation was to test the hypothesis that large-scale, high-throughput phenotypic screening data could be used in developing functional gene annotation. Several methods were developed in support of this hypothesis. These methods include the normalization approach presented as part of the MIPHENO software package and the ability to calculate similarity from missing/sparse data included in the NetComp software package. Additionally, these two software suites offer tools to the data analysis community that can be used with diverse datasets, as demonstrated in the preceding chapters.

5.1.1 MIPHENO

To address the need for a normalization procedure that could be extended to datasets with no replication or controls, I developed the MIPHENO software package. MIPHENO uses the properties of the dataset and the assumption that the majority of responses will be in the wild type or background range to conduct normalization in the absence of more traditional controls. This software package was developed to be extendable to a variety of high throughput datasets, includes a post-hoc quality control step, a normalization step, and a method for giving the probability an observed response is not wild-type. This software package provides much needed community resources that facilitate cross-dataset comparisons of high throughput screening data.

5.1.2 NetComp

While many methods are available to tackle datasets with missing data, most require parameter optimization, do not allow for thresholding, and are computationally intensive. SimMeasure addresses these issues and calculates an adjacency matrix that is analogous to a correlation matrix, an important feature for conducting downstream analyses. It deals with the issue of missing data (either randomly occurring or do to thresholding) by calculating the similarity using pairwise complete observations and penalizing the final similarity value by the number of instances where data was missing in only one individual. The similarity calculation used is based on the Canberra distance metric, providing a straightforward calculation that considers each pair of observations independently. The output structure of an adjacency matrix makes the result useable for many other downstream processes, such as community enrichment, and makes the result

accessible to existing graph-based analysis strategies that are commonly used in systems biology. SimMeasure is part of the NetComp software package. NetComp was created as a way to easily integrate and compare large scale datasets. Functions created as part of NetComp are designed to work with large scale datasets such as microarray and screening data. A set of functions that utilize the adjacency matrix structure are useful for performing data integration and network comparisons quickly while maintaining edge values. Currently, there are no software packages available for R, a commonly used and free platform for large data analysis, that incorporate these types of functions in one cohesive package that can easily work with other packages given the input and output structures. As a whole, this software package can be integrated into existing workflows to facilitate analyses within the R analysis environment.

5.1.3 Analysis of high throughput data and hypothesis generation

Using the developed methods on the data generated through the Chloroplast 2010 project, I was able to perform several analyses that were not possible before. Data is now comparable across the datasets, rather than just being comparable within an assay group. MIPHENO enables interpretations of the data beyond one assay group, which is demonstrated in Chapter 2, leading to a lower false negative rate. The cumulative distribution function used in MIPHENO means that individual insertion lines can be prioritized for follow-up based on the likelihood of phenotype, without the need for explicit controls. The ability to compare across the dataset means that lines with similar mutant phenotypes are identifiable and can now be clustered.

The method I developed for calculating similarity in the presence of missing and uninformative data, SimMeasure, enables the use community enrichment methods commonly used in gene expression data. This community enrichment information, when combined with the phenotypic information driving the community formation, is useful to researchers in guiding experiments to determine gene function. The workflow I developed made it possible to develop hypotheses of gene function based on phenotypic information with low statistical power that previously would not be possible using existing methods.

The workflow I created brings all of these software and analysis pieces into one cohesive whole. Altogether, this information creates a useful launch pad for future investigators that was not possible without the creation of the software developed as part of the dissertation research.

Research presented in this dissertation supports the hypothesis that high throughput phenotypic data enable gene function predictions using gene disruption lines. Using gene disruption lines is an important feature in this work even though it creates challenges regarding the quality of the disruption and confidence in the genes link to the phenotype. While there may be some false associations, the strategy of screening insertion lines offered novel data that made it possible to identify phenotypic associations.

Questions to be addressed

Several questions still need to be addressed, chiefly the accuracy of the predictions made and if the predictions are more helpful than those with other omic resources. An easy way to explore the comparison of the phenotypic-derived annotation and, for example,

transcriptomics would be to compare the networks generated with expression data to those from the phenotypic data. Measuring the accuracy of the predictions is a bit harder, given the high level of noise coming from the input data, but beginning the follow-up process with verification of the phenotype is a start. Some suggestions for follow-up on the predictions mentioned in the fourth chapter are detailed below.

Aside from the obvious biological questions regarding the workflow, a few computational questions remain. For example, are there limitations to the types or ranges of data suitable for SimMeasure? While it was inappropriate to do data imputation on the Chloroplast 2010 data given the nature of the missing data, it would be interesting to measure the success of using some data imputation methods on the unthresholded data. For individuals that had missing data due to quality control it might be useful, especially as the values could be compared to any replicates of that individual or for other lines annotated to the same locus. Thresholded values could then be used with SimMeasure to calculate the adjacency matrix to overcome the bias that occurred using the whole dataset.

Future work

5.3.1 Verification and follow-up on biological predications

A few examples were given in Chapter 4 to support the use of phenomics in building hypotheses of gene function. Two of the genes described were unannotated except for proteomics information to suggest localization. These are At2g26340 (Salk_048391), suggested to be involved with the photosynthetic complexes and At4g13590 (Salk_129037), an inner chloroplast membrane protein that exhibits a fatty acid phenotype.

Initial characterization of the two genes is similar and has to do with verification of the phenotype and the phenotype's link to the gene. Pipeline experiments that showed a phenotype need to be repeated with a higher biological replication, and include additional alleles as well as a couple members of their community.

The first gene, At2g26340 has another allele in the Chloroplast 2010 dataset, Salk_099844, which is in approximately the same position and direction as the allele noted and exhibits a more severe phenotype than Salk_048391. It is likely that it did not cluster with the other allele due to other phenotypes (similarity score between alleles is 0.66). Quantifying both seed and leaf amino acid changes using a higher statistical powered design is suggested changes in both datasets are observed. Both alleles show a weak positive response in the before high light Fv/Fm and an inconsistent weak positive in recovery. This gene is highly expressed in photosynthetic tissues (Winter et al., 2007) , consistent with its localization and proposed role in photosynthesis. If the phenotypes seem stable, then checking to see if this unknown protein was directly involved with any of the photosystem complexes would be a logical next step. The lack of a catalytic domain and mild phenotype suggest that the function is a regulatory or structural one, which depending on the strength of interactions could pose a challenge and methods need to be chosen that minimize disturbance of fragile interactions. Unfortunately, most localization methods require an antibody or protein tag, which might inhibit protein-protein interactions. One could consider the stability of the different photosystems in the two insertion lines relative to the wild type similar to the work done by (Lu et al., 2011) to

characterize LQY1's association with photosystem II using blue native gels. These experiments should provide enough starting material to move forward.

For the second gene, At4g13590, there is also another allele, SALK_011783, where the insertion maps to the 300 base pair region in the 5' UTR of the gene (Salk_129037 insertion maps to an intron). Both show a fatty acid phenotype, but the second allele has a decrease in 18:2, 18:3, and 16:1d7C, among others. The alleles have a similarity score of 0.6 and the similarity between SALK_011783 and the other members of the blue3 cluster are below 0.2. Aside from fatty acid phenotypes, this second allele has a few other amino acid phenotypes likely contributing to the low similarity scores. The only other SALK line available for this particular locus is SALK_148315, located in the promoter region. Due to the lack of a strong candidate for a second allele, it might be worthwhile first step to obtain homozygous lines from all three alleles and look at the expression of the gene to see if it is actually decreased. The higher powered fatty acid profiling can be carried out on along with other fatty acid biosynthetic mutants (possibly more than Fad6 and CDJ1). If At4g13590 expression is not significantly decreased in the other two alleles then complementation of the SALK_129037 is needed to be sure that the phenotype is tied to the At4g13590 locus.

The lack of a known catalytic domain in the At4g13590 gene and the presence of a computationally predicted transmembrane domain suggest a structural or regulatory role for the gene. In the event that the phenotype can be confirmed, identification of the interacting partners is the next logical step. Tagging or development of an antibody is needed, with epitopes likely/predicted to be located on the stromal side. Comparisons to

other proteins known to affect the transport of phospholipids in and out of the chloroplast may be good candidates for partners as well as the biosynthetic proteins.

5.3.2 Data integration using NetComp

One advantage of having data in the adjacency matrix format is that it facilitates data integration. An initial hypothesis going into the dissertation work was that data integration would help to build better models of gene function because of the additional information and a potentially larger dataset as some individuals are not included in both. This was tried using the Arabidopsis arrays from MetNetDB (Mentzen and Wurtele, 2008). This set of arrays covers the developmental series, biotic and abiotic stresses and is thus quite diverse. The intersection between the Pearson correlation of the MetNetDB data and the phenotypic matrix was used to identify communities with strong support for a relationship in each. The results were worse in quality than using the phenotypic data alone, based on the number of clusters with enriched ontology terms that were related to the phenotype of the community. Increasing the stringency of the phenotype did not resolve this.

Based on these results, a better approach to the data integration may be to use tissue-specific gene expression. For example, using transcriptomic data that is seed specific and combining it with seed-specific phenotypic data. It might be worth comparing the whole dataset to the seed-specific dataset as well, just to better detect traits that have relationships in both seed and leaf material. If the overlap between the two datasets shows promise, then the union might be considered to bring in those connections that are specific to one data type. Consideration will be needed in considering the edge weights (correlation or similarity score) in the integration. Based on earlier trials with a different dataset, taking

the intersection (which uses the average edge weight) and then incorporating the additional edges using the weight from that dataset seems to give favorable results.

One last consideration for integration with transcript data is anchoring data to the genome. The insertion lines were anchored to the locus using the information provided in the Chloroplast 2010 dataset. There are many examples for which this information isn't the most accurate because the insertion lays between two genes or possible targets a single splice form. With the transcript data, there might be many probes for a single gene (corresponding to different parts of the transcript as well as different isoforms) in addition to probes mapping to multiple genes. All these factors complicate the data integration and several different approaches should to be tested to see how to minimize the impact on overall data interpretation.

Final comments

This work represents a first step in using large-scale data for building annotation of gene function. The methods developed are aimed at facilitating analysis of sparse datasets and screening studies, but are extendable to other data types. As more datasets are made publically available, these types of tools should facilitate additional post hoc analyses and data integration, hopefully lending itself to better design of follow-up experiments leading to improved gene annotation.

Improvements in data quality could advance this type of work. Small improvements, like expanding the experimental annotation to give details about what was studied, why it was studied, and how the measurements were taken, can go a long way in orientating

someone to a dataset. Even when conducting high throughput experiments, including small levels of replication (for example, two trays with the same individuals, or one cell devoted to a control) can go a long way in improving the power of an analysis but don't add considerably to the overall cost. Finally, making data available, freely, completely, and without restrictions, is important to moving the field forward as a whole. If data is kept in silos, it cannot be used in further analyses and new knowledge cannot be realized.

REFERENCES

References

Lu, Y., Hall, D.A., and Last, R.L. (2011). A small zinc finger thylakoid protein plays a role in maintenance of photosystem II in *Arabidopsis thaliana*. The Plant Cell Online *23*, 1861-1875.

Mentzen, W., and Wurtele, E. (2008). Regulon organization of Arabidopsis. BMC Plant Biology *8*, 99.

Winter, D., Vinegar, B., Nahal, H., Ammar, R., Wilson, G.V., and Provart, N.J. (2007). An "electronic fluorescent pictograph" browser for exploring and analyzing large-scale biological data sets. PLoS ONE *2*, e718.

APPENDICES

Appendix A

Software is available on the Comprehensive R Archival Network at the following locations:

MIPHENO: http://cran.r-project.org/web/packages/MIPHENO/index.html

NetComp: http://cran.r-project.org/web/packages/NetComp/index.html

Data and methods used to carry out the analyses as well as the results for Chapter 2 can be found at:

http://www.biomedcentral.com/1471-2105/13/10/additional

Data and methods used to carry out the analyses as well as the results for Chapter 3 can be found at:

http://www.plastid.msu.edu/links/Dissertation%20Supplemental%20Materials/Chapter%203%20supplementary%20materials/

Data and methods used to carry out the analyses as well as the results for Chapter 4 can be found at:

http://www.plastid.msu.edu/links/Dissertation%20Supplemental%20Materials/Chapter%204%20supplementary%20materials/

Appendix B

The code for the SimMeasure algorithm is included for reference. Code for all software and analyses can be found in the links provided in Appendix A.

From SimMeasure.R, the wrapper function that calls SimMeasure from the R environment

```
SimMeasure<-function(data, threshold=NULL, ...){
    x<-.Call("SimMeasure",data, threshold, pkg="NetComp" )
    if(!is.null(row.names(data))){
          row.names(x)<-colnames(data); colnames(x)<-
          colnames(data)
    }
    x
}
```

From SimMeasure.c

```
#include "Rdefines.h"
#include "Rinternals.h"
#include "R_ext/Rdynload.h"
#include "math.h"

SEXP SimMeasure(SEXP data_matrix, SEXP thresh){
    // data_matrix is a matrix
    // thresh is a double

    int num_cols, num_rows;
    double *rx = REAL(data_matrix), *rans, t;
    SEXP retval;
```

```
//Check to make sure that everything is of the proper type
//before going further...

if (isMatrix(data_matrix)){

     num_cols = ncols(data_matrix);

     num_rows = nrows(data_matrix);

}

else{

     Rprintf("invalid matrix.\n");

     return R_NilValue;

}

if (isNull(thresh)){

     Rprintf("warning, setting threshold to 0 by
     default.\n");

     t = 0;

}

else{

     t = REAL(thresh)[0];

}


//Check to see if the matrix has any null values

if (isNull(data_matrix)){

     Rprintf("matrix must not be NULL.\n");

     return R_NilValue;

}


PROTECT(retval = allocMatrix(REALSXP, num_cols, num_cols));

rans = REAL(retval);
```

```
for(int i = 0; i < num_cols; i++){

     for(int q = 0; q < num_cols; q++){

          double cor_val = 0.0;

          int count_row_nas = 0;

          for(int wi = 0; wi < num_rows; wi++){

               //Rprintf("row_nas: %d\n", row_nas[wi]);

               //Check to see if we need to skip this row
               //b/c BOTH of the elements are NA

               if(ISNAN(rx[i * num_rows + wi]) &&
               ISNAN(rx[q * num_rows + wi])){

                    count_row_nas++;

                    continue;

               }

               //Check to see if we need to skip this row
               //b/c BOTH of the elements are <hit

               //this is for cases where the 'non-hits'
               //werent removed

               if(fabs(rx[i * num_rows + wi]) < t &&
               fabs(rx[q * num_rows + wi]) < t){

                    count_row_nas++;

                    continue;

               }

          }

          //Calculate the parts of the similarity function

          double nm = 0.0; double pm = 0.0; double om =
          0.0;

          double pcnt = 0.0; double ocnt = 0.0;

          for(int wi = 0; wi < num_rows; wi++){
```

```
//Check to see if one of the elements are NA

if(ISNAN(rx[i * num_rows + wi]) &&
ISNAN(rx[q * num_rows + wi])){

     continue;

}

else if(fabs(rx[i * num_rows + wi]) < t &&
fabs(rx[q * num_rows + wi]) < t){

     continue;

}

//one value missing, the other above
//threshold

else if((ISNAN(rx[i * num_rows + wi]) &&
fabs(rx[q * num_rows + wi]) >= t) ||
(ISNAN(rx[q * num_rows + wi]) && fabs(rx[i *
num_rows + wi]) >= t)){

     nm++;

}

//one value below threshold, other value
//above

else if((fabs(rx[i * num_rows + wi]) < t &&
fabs(rx[q * num_rows + wi]) >= t) ||
(fabs(rx[q * num_rows + wi]) < t &&
fabs(rx[i * num_rows + wi]) >= t)){

     nm++;

}

//both are duds

else if((ISNAN(rx[i * num_rows + wi]) &&
fabs(rx[q * num_rows + wi]) < t) ||
(ISNAN(rx[q * num_rows + wi]) && fabs(rx[i *
num_rows + wi]) < t)){

     continue;

}
```

```
        else{

            if(rx[i * num_rows + wi] * rx[q *
            num_rows + wi] >= 0){

                if(rx[i * num_rows + wi]==0 &&
                rx[q * num_rows + wi]==0){

                    pm = pm +1;

                    pcnt++;

                }

                else{

                    pm = pm + 1-(fabs(fabs(rx[i *
                    num_rows + wi]) - fabs(rx[q *
                    num_rows + wi]))/(fabs(rx[i *
                    num_rows + wi]) + fabs(rx[q *
                    num_rows + wi])));

                    pcnt++;

                }

            }

            else{

                om = om + 1-(fabs(fabs(rx[i *
                num_rows + wi]) - fabs(rx[q *
                num_rows + wi]))/(fabs(rx[i *
                num_rows + wi]) + fabs(rx[q *
                num_rows + wi])));

                ocnt++;

            }

        }

    }


    //Calculate the correlation, and the correlation
    //matrix

    if(num_rows - count_row_nas < 1){
```

```
                    cor_val = NA_REAL;
                }
                else{
                    cor_val = (pm - om) / (pcnt + ocnt +
                    nm/(pcnt+ocnt+nm));
                }
                rans[i * num_cols + q] = cor_val;
            }
        }
        UNPROTECT(1);
        return retval;          //return the correlation matrix
}
```