QUANTITATIVE GENETICS AT THE MOLECULAR LEVEL: A ROOT METABOLOMICS CASE STUDY

Andrew Hauck and Martin Bohn

Department of Crop Sciences, University of Illinois at Urbana-Champaign

Introduction and Background

Over the last several years, a number of plant studies have utilized high throughput analysis methods, such as expression profiling, protein quantization, and metabolic profiling, on plants to obtain large data sets of molecular phenotypes on a variety of genetic backgrounds (see Young and Udvardi, 2008). For example, metabolic profiling has been used to observe the effects of mutated genes (Feihn et al., 2000) and genetically map quantitative metabolite traits in a population of introgression lines (Schauer et al., 2006). Despite such prevalence and interest in these approaches, there has been relatively little work to reconcile the kind of data obtained with these methods to plant breeding theory. There are many examples of breeding for traits identified with a rapid assay, such as spectroscopy or UV-chromatography, from the last several decades, but adaptation of the newer profiling approaches has not been intuitive. The molecular traits that should be improved are not initially defined, as they are generally not the desired end product themselves. In other words, the final overall desired phenotypes are known, but the quantitative molecular phenotypes that contribute to them are not, requiring some association between them to be made. The compensating advantage of profiling techniques is the ability to accurately determine the abundance and identity of a large variety of molecules, on the order of one hundred to tens of thousands, depending on the methods and molecules. The challenge is to discover ways to coherently interpret and apply these kinds of data, which are now a highly available resource. This will require basic research into the range and conditions of molecular variation, nature of inheritance, and their relationships to the traditional agronomic traits of interest.

About two decades ago the introduction of inexpensive genetic markers had a profound impact on the methods of plant breeding. Similar improvements in DNA chips, sequencing, and mass spectrometry could represent another fundamental advance. This possibility derives from the understanding that a plant's reproductive productivity depends on a multitude of prior biochemical and physiological responses to environmental conditions and the feasibility of gathering such information. Indeed, the capability to acquire tens to thousands of molecular phenotypes per genetic background, tissue, developmental time point, and environment combination provides a mind boggling number of potential options for collecting data to select on. Obviously, many challenges remain before widespread practical implementation of "omics" based breeding can proceed. Critically, we must determine how omics phenotypes answer the fundamental questions of plant breeding. These questions include: "What are the molecular traits that should be selected for, their variance in my populations, and the accuracy of measurements of them?", "Is there suitable variation to perform selection on and are the genetic effects inherited in a familiar way?", and "How does the trait under selection influence other traits of interest?" Simple answers for all of these straightforward questions are elusive with omics data.

The kinds of relationships that exist between such information and primary traits need to be investigated for a variety of conditions. Conventional analyses appear less suited to the nature of this kind of multivariate data, which has substantial correlation and can change rapidly in response to various stimuli, including the process of sampling and tissue preparation. Moreover, the interrelationship between variables, especially with metabolites, implies that selection on any one of them may result in simultaneous changes to others, influence the steady state or dynamic properties of the system as a whole, and introduce unknown risks of undesirable behavior. In the case of improving metabolic systems, simultaneous selection of many molecular traits towards rationally chosen goals determined from a comprehensive understanding of network relationships might not be just an ideal, but rather a requirement for functional application. Since the relationships between genetic backgrounds and metabolome characteristics are essentially unexplored in maize, we decided to perform a simple experiment to investigating these issues. We compared vigorous hybrids with their less vigorous inbred parents to begin identifying distinguishing features between more productive and less productive metabolic networks and pursue answers to the fundamental questions of plant breeding for profiled molecular traits..

Experimental Design and Methods

A half diallel design plus a set of two inbreds and their reciprocal hybrids were employed to investigate the relationship between genetic background and root metabolic profiles. The inbred parents include flint (3), dent (2), and iodent (1) inbreds, resulting in combinations of inter- and intrapool hybrids (Figure 1). Seeds from the eight hybrids and six parental inbreds were germinated *in vitro* using germination paper rolls. Five seeds per entry were placed on individual pieces of germination paper and rolled up. The fourteen rolls per experimental replication were divided equally between two 2.5 L containers filled with 750 ml distilled water and 20ml of a 2.5g per L Captan solution, which was also used to pre-moisten the rolls. All of the containers were simultaneously incubated at 28 degrees Celsius and 100% relative humidity for 8 days in the absence of light. The experimental design was an incomplete block design with three replications. The germination paper rolls represented plots and the two containers per replication were regarded as blocks. Two pairs of seedling roots were harvested into liquid nitrogen and ground, generating two bulked samples per germination paper roll, for all replications on the same day, with the goal of acquiring a total of six biological replications per entry.

Metabolites were extracted from freeze dried root tissue by incubating 15mg in 1.5ml 80% methanol overnight at 4 degrees Celsius. The methanol solution was removed, filtered, and dried down, and the samples were submitted to the Roy J. Carver Metabolomics Center at the University of Illinois at Urbana-Champaign for derivitization and Gas Chromotograpy - Mass Spectrometry (GC-MS). Relative levels of 159 confidently identified metabolites were provided for each of the 79 samples submitted. Metabolite abundances were reported in relative terms, rather than absolute, due to the unavailability of quantitative standards for all detected molecules. Relative levels are calculated by comparing the detected signal intensity of a chemical standard with the intensities of identified metabolites. Metabolites related to a wide variety of metabolic pathways were identified (Table 1), but metabolism involving sugars or amino acids is particularly well represented.

Results and Discussion

The shotgun metabolomic approach employed provides a snapshot of observed metabolites at the time of tissue harvest, not experimentally determined flux through pathways. It should also be noted that the estimates are valid only for the entry, time point, tissue, and environmental conditions sampled, which are quite limited in scope in our experiment. However, the goal for this project was not to make future predictions about metabolite levels, but to determine the nature and proportion of variation in metabolite abundances that is attributable to the genetic composition of entries. The metabolite profiles of entries often displayed cases of both consistency and sharp variation (Table 2). In the second case, instances of reproducible variation for a metabolite could generally not be partitioned to effects from the paper rolls (plot), container (block), or replicate of the experiment (environment). This feature of strikingly increased or decreased abundances for certain metabolites in a subset of replicates was observed for all entries. Since we do not have a statistical measure at this time to identify these cases strictly, our working definition of this feature encompasses instances where two or more replicates of an entry with different environments have average differences with the other replicates in the group of approximately 1.5 fold for highly abundant metabolites or 2 or 3 fold for the least abundant ones. Repeated instances of this behavior for a particular metabolite can be correlated with similar effects at other metabolites, though the correlations can be entry dependent. These metabolomic perturbations could result from some post-harvest effect, although it is interesting to note that all samples show some amount of this behavior for some subset of metabolites.

Comparing results of inbreds with hybrids also indicates the role of factors unconsidered at the onset of the experiment. For example, many of the hybrids have levels of core metabolites below the detection threshold, while the inbred lines have abundant amounts. This might be explained by noting the differential growth and development rates between inbreds and hybrids. Although both were harvested at the same time, the hybrid entries had greater lateral root quantity and length, by visual inspection, and may be significantly more advanced developmentally, even at this early stage, resulting in a greater consumption of initial seed resources. This distinction makes it more difficult to compare inbreds and hybrids metabolically without a time series analysis.

The structure of the root metabolite data set is typical of similar current high-throughput approaches to biology, containing information on a large number of correlated variables and a fewer number of observations. These factors preclude most multivariate analysis methods typically used on phenotypic data. Additionally, univariate methods are unattractive since metabolic phenotypes at the molecular level can change rapidly, information on co-variation with other metabolites provides more confident estimates, and results from shotgun style studies lend themselves to interpreting results of biological interest in terms of network pathways. To this end, we endeavored to evaluate some alternative statistical methods to see which ones may be more informative.

One common method for dealing with high dimensional data with few observations, in applications such as Near Infrared Spectroscopy, is called Partial Least Squares or Projection to Latent Structures (PLS). PLS is implemented with an iterative algorithm involving matrix

decomposition that extracts summarizing factors, like the more familiar Principal Components method, which maximize the covariance between the independent and dependant variables. For each iteration, variance is partitioned into a factor and removed, leaving residuals that are passed for the next round of partitioning to use. If the information explained by a factor fails a certain threshold, it is discarded and the process concludes. Each resulting factor is orthogonal, having no overlap of information with other factors, and assigns a value to every Y group that is a weighting of many X variables. The total Y variation that the factors explain combined is reported as R-square.

We can perform a discriminate form of this analysis with entry as a qualitative Y trait and plot several of the factors to see if the information contained in the metabolite data is sufficient to distinguish the different genetic backgrounds represented in the diallel. Initial results obtained using SAS proc PLS showed separate clusters of replicates for most entries when plotted with three factors, so enhanced methods were sought. The metabolite data was next processed using SIMCA-P software and an Orthogonal PLS Discriminate model (OPLS-DA). Orthogonal PLS has the additional feature of compartmentalizing variation in the X data that is unrelated to the Y. With this method, a portion of variation from environmental effects and noise is removed to improve the quality of the factors and facilitate their interpretation (Trygg and Wold, 2002). With this approach, 13 factors and 2 orthogonal components were identified. Investigating the factors revealed that each one primarily discriminated several of the lines, thus plotting only three of the factors in three dimensional space may not fully resolve all fourteen of the entries into independent clusters. Nevertheless, many of the factors provide good separation of most of the entries (Figure 2). For the purposes of plant breeding, such an approach might be useful for identifying a more functional or predictive form of diversity than genetic differences.

Another suitable approach for our data involves estimation of the partial correlations between all of the metabolites. When attempting to determine the statistical association between two variables, partial correlations has the advantage, compared with correlation, of removing the influencing effects from other variables. A conventional implementation of partial correlation analysis will not work with data with fewer observations than variables, but alternative means are available. A package for R, Genenet, uses a shrinkage based estimate of partial correlations and a local fdr multiple testing adjustment to determine relationships between variables in datasets with fewer observations than variables (Opgen-Rhein and Strimmer, 2007). The results from this program are then used to construct a Gaussian Graphical Model (GGM), which visualizes the connections between metabolites with the most significant partial correlations. GGMs do not lend themselves to clear interpretation or obvious application, since actual biochemical network is not reconstructed and information on the entirety of the metabolome cannot be acquired from shotgun approaches at this time, but the graphs may be useful for determining roles and relative importances of identified but uncharacterized metabolites. The R code also generates some statistics for partial variances and partial regression coefficients of the metabolites that we might be able to use for better estimates with more traditional methods like comparison of means and contrasts.

Conclusions

The metabolic profiles of our maize genotypes contain sufficient information to distinguish each entry. This implies that allelic diversity has a discernable influence on the metabolic characteristics of primary root tissue, which likely contributes to the final root

phenotype. Possible examples of metabolite traits inherited with simple additive effects were identified by comparison of hybrids with their inbred parents, but require additional data to confirm. Apparent coordinated effects were observed in replicates of each entry that spanned different subsets of metabolites. Partial correlations are a way to construct a relationship network more confidently that includes metabolites without assigned chemical structures, however methods to validate the findings need to be explored. The data showed particular differences between inbreds and hybrids that raise questions about the assumptions involved with a direct comparison that cloud interpretations of heterosis. Additional methods of analysis appropriate for this kind of data are being pursued. Previously, improvement for desired secondary metabolites has taken place using intensive biochemical analyses, but not direct selection on the transcriptome, proteome, or metabolome as a whole towards an omic ideotype. Connections between molecular profiles and traditional agronomic traits need to be established before such an approach can be applied. The evaluation of a sufficient number of contrasts between vigorous and non-vigorous material may provide a framework for future omics based breeding efforts. Currently, the main limitations of omics experiments are the financial resources required for satisfactory replication and the challenge of confidently identifying instances of technical error versus natural variation. “Ome” wide selection is the culmination of the application of functional genomics to plant breeding.

Acknowledgements – We thank Prof. A.E. Melchinger (University of Hohenheim, Germany) for providing the seed of the materials used in this study.

References

Fiehn O., Kopka J., Dörmann Peter, et al. 2000. Metabolite profiling for plant functional genomics. Nature Biotechnology. 18:1157-1161.

Opgen-Rhein R. and Strimmer K. 2007. From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. BMC Systems Biology. 1:37.

Schauer N., Semel Y., Roessner U., et al. 2006. Comprehensive metabolic profiling and phenotyping of interspecific introgression lines for tomato improvement. Nature Biotechnology. 24:4, 447-454.

Trygg J and Wold S. 2002. Orthogonal projections to latent structures, O-PLS. Journal of Chemometrics. 16:119-128.

Young N.D. and Udvardi M. 2008. Translating Medicago truncatula genomics to crop legumes. Current Opinion in Plant Biology. 12:1-9

Table 1 Major pathways represented by profiled metabolites. Some metabolites were assigned to more than one pathway, others were not included.

Aspect of Metabolism	# of related metabolites
Sugar / Starch	40
Amino Acid	38
Phosphotransferase System	13
Phenylpropanoids	12
Fatty Acids	8
Glyoxylate and dicarboxylate	7
Urea	7
Alkaloid synthesis	7
Ascorbate	6
Central Metabolism	6
Pentose Phosphate Pathway	6
Purine	6
Pyrimidine	6
Glycerolipid/phospholipid	5
Steroid	5

Table 2 Relative abundances of selected metabolites for a inbred-hybrid triplet. Quantities of metabolites for replicates of the inbreds are consistent across replications in most cases. Two of the hybrid's replicates show drastic correlated effects across many amino acids and the additional metabolites GABA and sucrose. One of the S028 replicates shows a reduced sucrose level and a replicate of P024 has a greatly increased level, but both lack the correlated response to other metabolites as observed in the hybrid. Note the consistent values of arabinose in replicates and the reduced level in the hybrid compared to the inbreds. The value of zero for glutamine in P024 should be interpreted as being below the sensitivity of the analysis, rather than as an absolute.

Env.	Entry	Alanine	Asparagine	Lysine	Met.	Phen.	Proline	Tyrosine	Cysteine	Glut.	Arabinose	GABA	sucrose
1-1	S028	19.3	181.8	5.2	21.5	4.8	1.5	26.1	1	21.7	19.6	23.9	4.3
2-2	S028	13.6	210.1	5.3	23.4	6	1.7	27.2	1.3	21	25.6	26.2	817.3
2-2	S028	17.2	186.2	5	20.8	5.7	1.9	22.1	1.7	17.1	25.2	30.5	975.7
3-1	S028	22.6	193.3	6.2	27.1	8	1	22.7	1.7	24.2	24.3	20	813.6
3-1	S028	15.9	199	5.2	27.2	5.2	1.2	25.7	1.6	15.9	24.4	26.9	899.1
1-2	S028xP024	**96.6**	**96.6**	**88.3**	**47.9**	**47.9**	**28.7**	**98.6**	2.6	8.6	3.1	**313.4**	**10.8**
1-2	S028xP024	35	84.1	9	80.1	13.5	1.7	31.2	1.2	10.2	4.9	34.4	838.6
2-1	S028xP024	24.4	129.3	6.1	69.2	10.8	1.4	22.7	0.8	9.7	3.8	35.5	887.3
2-1	S028xP024	**114.1**	**16.7**	**55.4**	**41.4**	**40.9**	**19.4**	**91.4**	1.1	9.1	2.7	**178.2**	**13.8**
3-1	S028xP024	23	110.8	10.3	61.8	14.2	1.8	29.9	1	8.3	3.2	25.7	1148.6
3-1	S028xP024	25.3	125.5	9.2	63.6	12.9	1.6	35.4	1.2	11.6	3.4	34.2	1122.4
1-1	P024	13	9.6	2.7	17.3	1.9	5.6	69.4	1.2	0	17.6	114.9	10.1
2-2	P024	13.3	12	2.6	16	1.5	6.9	51.1	1.1	0	12.1	8.6	9.3
2-2	P024	14.2	17.5	2.3	15.6	2.6	5.8	55.8	3.5	0	13.1	6.5	9.4
3-2	P024	16.7	13.5	2.6	11.4	1.7	6.2	57.5	1.5	0	10.5	8.6	126
3-2	P024	15	19.1	2.8	11.9	1.3	5.5	46.2	1.5	0	10.7	8.1	9.6

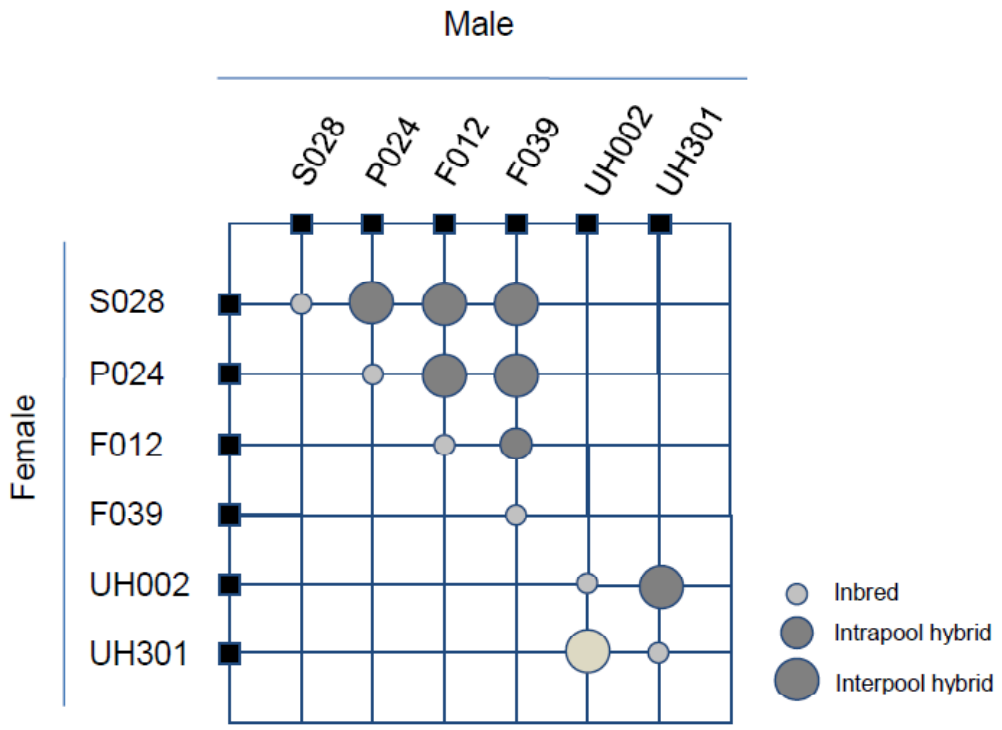

Fig 1 Maize Genotypes and crossing design used in this study.

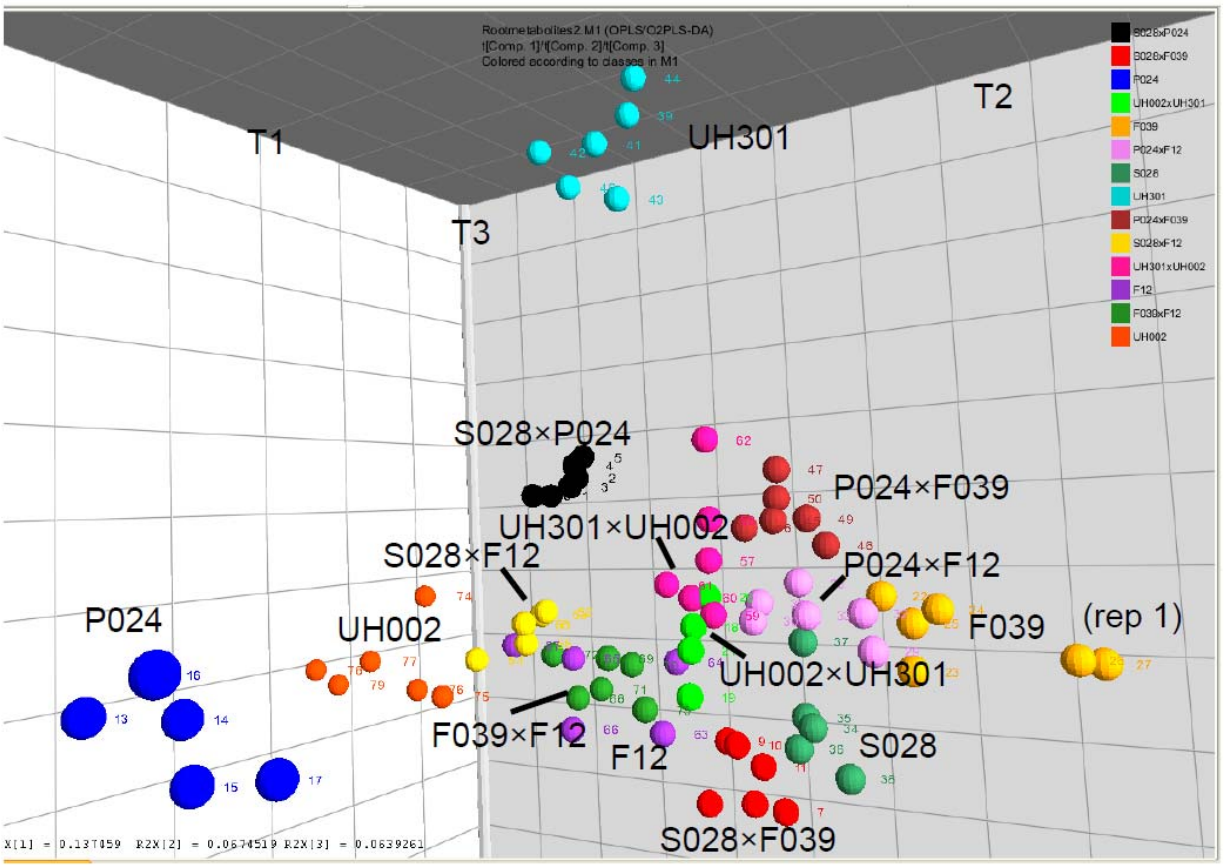

Fig 2 SIMCA visualization of entry clusters discriminated using the three most significant OPLS factors. Each sphere is a replicate of an entry with SIMCA assigned observation numbers beside them. For these factors, replicates of the reciprocal hybrids derived from crosses between UH002 and UH301 were not well segregated and entry F12 is similar to two of the hybrids it is a parent of. The isolation of two replicates of F039 from the same germination paper roll from the others along the T2 factor could be explained by environmental effects.