Analyses of Soybean Seed Transcriptomics and Metabolomics Data Using R: A Systems Biology Approach to Understanding Seed Composition

Ling Li[1], Wenxu Zhou[2], Wengang Zhou[1], Dan Nettleton[3], Mark Westgate[4], Basil Nikolau[2], and Eve Syrkin Wurtele[1]

[1]Department of Genetics, Development, and Cell Biology, [2]Department of Biochemistry, Biophysics, and Molecular Biology, [3]Department of Statistics, and [4]Department of Agronomy, Iowa State University, Ames, IA 50011 [http://www.metnetdb.org]

ABSTRACT

In addition to being the propagule that ensures the dissemination of plants, seeds also provide one of the major products of agriculture. The biochemical storage reserves that are deposited within the seed during its development fall into three general categories: proteins, oils, and carbohydrates. These seed reserves are biosynthesized by the programmed expression of a metabolic network during seed development. In most commercial lines of soybean grown in the Midwestern states of the US, seeds are composed of 40% protein, 20% oil, 15% soluble carbohydrate, and 15% fiber (http://www.asa-europe.org/SoyInfo/composition_e.htm). There is considerable knowledge concerning the basic biochemical processes by which imported carbon and nitrogen is converted to the final products, protein, oil and carbohydrate. However, there is a great deal to be learned concerning the molecular, biochemical and genetic mechanisms that regulate this complex metabolic network. Recent developments in genomics have started to provide the catalogue of genes that would be required for this process. We have taken advantage of microarray and metabolomics technology to identify the global gene expression profile that regulates the developmental and biochemical network, which determines final seed structure and composition. We have coupled this with bioinformatics analyses to gain insights as to the regulation of the biochemical program that determines soybean seed development.

Using R in the exploRase software, we have analyzed high dimensional transcriptomics and metabolomics data derived from seeds of elite lines of soybean seeds high and low in protein composition. These analyses have enabled the identification of genes and metabolites that may be important for soybean seed composition.

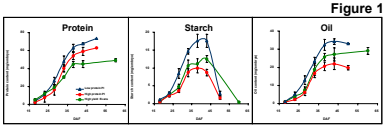
ANALYSES USING R

Affymetrix Signal values were natural log transformed and median centered within each GeneChip, and normalized log signals were analyzed separately for each probe set using a linear model in R. Each linear model included fixed effects for replications and time points. As part of each linear model analysis, an overall F-test was conducted to scan for any change in mean expression across the time points examined. The p-values from these F-tests were converted to q-values using the method of Storey and Tibshirani to control false discovery rate (FDR) at specified levels. In addition, the p-values for all pairwise comparisons between time-specific expression level means were computed in each gene-specific linear model analysis. The set of p-values for each comparison of one time to another were also converted to q-values as described above.

K-medoids cluster analysis was used to organize and visualize the expression patterns of the 2869 genes with q-values less than 0.01 in the overall F-test for change of expression over time (Figure 3). Euclidean distance between standardized expression profiles was used to determine the dissimilarity between genes when clustering. Cluster number was selected using the criterion described by Krzanowski and Lai. A similar clustering method was also applied to the metabolomics data (Figure 2). Different R functions were written for RCBD analysis (randomized complete block design) (produces treatment means and a p-value for each gene if the data correspond to a RCBD with no missing data and one experimental unit per treatment in each block; p-values for differences between all pairs of treatment means are also computed), calculate FDR, get an estimate of K for K-medoids using the gap statistic, plot gene expression and clusters.

RESULTS

Figure 1

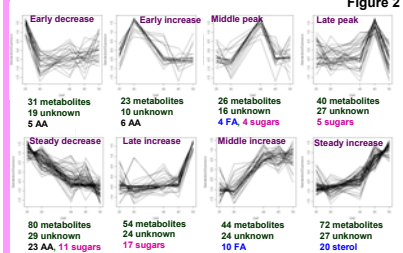


Protein, oil and starch content in high protein, low protein, and Evans soybean lines during seed development.
Values are means of three or more replications. Bars represent standard error. Protein content increases over seed development; starch content increases from the beginning and reaches a peak at 40 days after flowering (DAF), and decreases rapidly after 45 DAF; oil accumulation is similar to that of protein's, but it reaches a peak at about 45 DAF.

LEGENDS

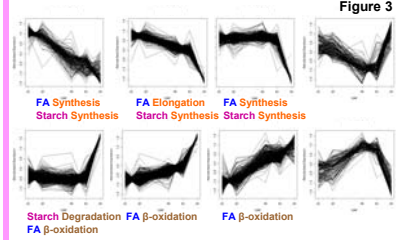
Metabolomics: Study of small-molecule metabolite profiles.
Carbohydrate metabolism (starch, sugar); amino acid (AA); fatty acid (FA) metabolism.
Anabolism (Synthesis/elongation); catabolism (degradation/oxidation).

Figure 2



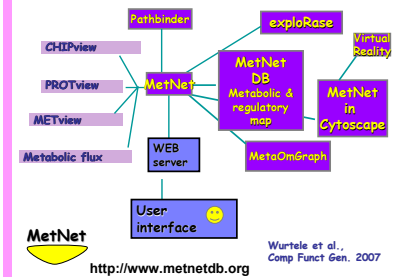
Cluster analysis of 400 metabolites that alter accumulation during Evans seed development.
These metabolites are grouped into eight clusters based on their accumulation patterns.

Figure 3



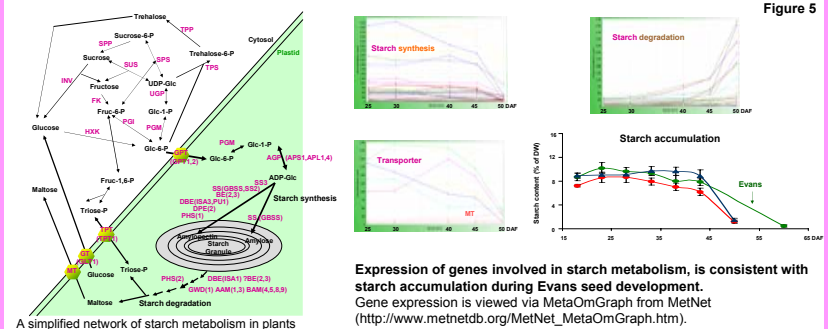
Cluster analysis of the 2869 genes that are differentially expressed over Evans seed development.
When FDR is controlled at the level of 0.01, 2869 genes are declared differentially expressed over time. Different clusters contain genes involved in different metabolic pathways.

Figure 4



MetNet systems biology platform for plant 'omics was also used for data analysis.

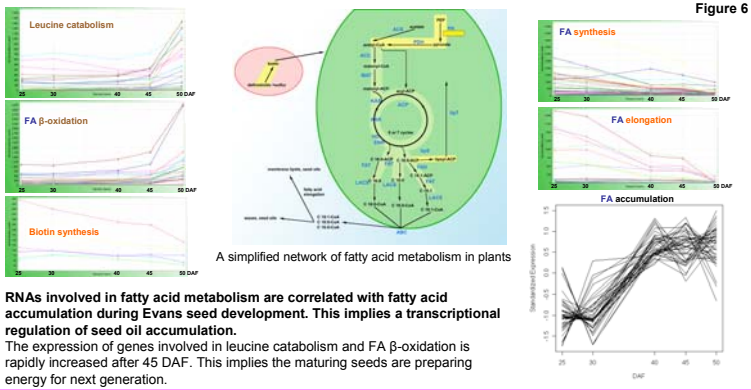
Figure 5



A simplified network of starch metabolism in plants

Expression of genes involved in starch metabolism, is consistent with starch accumulation during Evans seed development.
Gene expression is viewed via MetaOmGraph from MetNet (http://www.metnetdb.org/MetNet_MetaOmGraph.htm).

Figure 6



A simplified network of fatty acid metabolism in plants

RNAs involved in fatty acid metabolism are correlated with fatty acid accumulation during Evans seed development. This implies a transcriptional regulation of seed oil accumulation.
The expression of genes involved in leucine catabolism and FA β-oxidation is rapidly increased after 45 DAF. This implies the maturing seeds are preparing energy for next generation.

Funding: Plant Science Institute, Iowa State University; Consortium for Plant Biotechnology Research, USDA