Aim 2: **Trait-to-gene network module discovery pipeline**

**Rationale**: We propose to develop and implement methods to build gene networks from co-expression data on multiple crop species to identify genes that are potentially central to a particular trait of interest. We will then test some of those candidate genes by over-expression, knock-outs, and knock-ins in Arabidopsis. The most promising of those will, in turn, be tested in Maize.

**Novelty**: The trait-to-network module approach follows the spirit of AraNet [Lee 2010] and PlaNet [Mutwill 2011] in that multiple species are used to identify functionality in gene networks. What is novel in our proposed approach is: (i) we use expression data from trait-relevant experiments on multiple crop species to “weight” edges in the network and identify sets of genes associated with a trait, (ii) we identify orthologous genes that are relevant to a trait, some of which may be missing in Arabidopsis, (iii) our experimental strategy is based on a medium throughput validation testing in an inducible expression system in Arabidopsis, and (iv) a low throughput validation of selected genes in Maize.

**Significance: Agronomic traits and phylogenomic context**: Since the dawn of agriculture, farmers and scientists have improved crops by selection and breeding. Among the 21 sequenced crop and non-crop species in the phylogenomic tree of Fig. 2, the crop species lie at phylogenetic distances of a hundred million years or more from a common ancestor [Chaw et.al. 2004]. Nevertheless, specific gene functions are conserved across these species, sometimes at large phylogenetic distances [Irish and Yamamoto 1995]. As such, complex agronomic traits, such as seed development, seed composition, root architecture, flowering time etc., likely result from medium-sized conserved networks of genes rather than single genes [Espinosa-soto et al., The Plant Cell Nov 2004, To et al., The Plant Cell July 2006]. Arabidopsis, through its amenability to mutational and transformational studies, has provided the vast majority of knowledge about these traits, e.g. flowering time [Espinosa-soto et al., The Plant Cell Nov 2004], seed development [To et al., The Plant Cell July 2006] and root architecture [Péret et. Al., Trends in Plant Science July 2009]. Because Arabidopsis has limited tolerance to extreme conditions, however, data about environment-specific responses of genes is easier to obtain in other species [Li et al. 2011, Tuteja et al. 2010]. Examples include drought resistance [Shen Y, Venu RC, Nobuta K, Wu X et al. 2011] and early seed development [http://www.ncbi.nlm.nih.gov/geo/ : GSE29163]. Additionally, involvement of specific genes in agronomic traits is perhaps better investigated in those species. Hence, expression atlases [Severin et al. BMC Plant Biology2010] and numerous individual expression assays exist for each of several crop species.

**Gene Discovery**: The combined computational-experimental approach described below, capitalizes on mining transcriptomic data from crop species, to inform the identification of gene network modules associated with traits of agronomic interest. This will lead to novel gene discoveries that can be tested using Arabidopsis as a model. For example, in some cases, this comparative network analysis will identify Arabidopsis orthologs of crop genes, whose significance in the trait could not have been identified based solely on Arabidopsis data. In other cases, it will identify nodes of networks that are “missing” in Arabidopsis, but present in the crops. These crop genes could be “knocked-in” to Arabidopsis, using a protoplast transient assay system we describe in Aim 2B. Here is the method:

**Step 1**: **Identify** **trait-related expression datasets**: For each trait, collect one set of experiments [NCBI GEO] that are relevant to the trait (e.g. during seed development). Collect a second set of expression data under unrelated conditions. Transcriptomic data from next-gen sequencing or microarrays can be used as available. Genes that do not show variation, as determined by a minimum variance cut-off across all experiments are “housekeeping” genes, and are removed from further analysis. Further filters are applied to remove poor quality data (e.g. low counts). After filtering, raw counts are normalized using a full-quantile normalization method [Bullard et.al. BMC Bioinformatics 2010].

**Step 2**: **Compute gene correlations**: Build gene correlation networks separately in each of species S1…Sn. We will use methods for correlation proven useful in other plant network tools including Pearson correlation coefficient [Usadel et. al., Plant, cell and environment, 2009], as well as cosine coefficient (CoP) [Ogata et. al., Bioinformatics 2010] and Mutual rank [ATTEDII, PlaNet] [Obayashi et.al., NAR 2008; Mutwil et.al., Plant cell, 2011] and test the other correlation methods discussed in Aim 1. A gene correlation network (CNi) for species Si consists of edges {g1, g2} such that the absolute value of the correlation between these two genes is at least 0.7 with p-value <=0.05

**Step 3. Consensus through “weighting” of nodes and edges:** First form networks in each species of interest (e.g. each crop). A gene g in such a network that passes an orthology cutoff (Stringent BLAST e-value cutoff) with respect to some Arabidopsis gene g’, will give g’ a vote which we will call a weight from now on (because we will eventually give different weights to different species depending on phylogenomic distance). Thus g’ may receive “weights” from several genes in each network and from multiple species networks as in Fig X. If g does not pass the orthology cutoff with respect to any gene in Arabidopsis, but is orthologous enough to genes in other crop species, then we consider that gene to be a candidate for a “knock-in” experiment in Arabidopsis.

If there is an edge between g1 and g2 in a species-specific network, and g1 exceeds the orthology cutoff to g1’ in Arabidopsis (as does g2 with respect to g2’), then add a “weight” to the edge between g1’ and g2’. Experimentally validated edges from Arabidopsis provide additional weights to the network edges and nodes. Edges representing predicted Cis-binding sites for Arabidopsis transcription factors, are added to provide confidence and direction to the regulation network. The final network including nodes, edges and weights, forms the Treturn network within Arabidopsis.

**Step 4**: **Identify conserved network modules**: The reciprocal of the weights of the edges form a measure of distance, thus assigning low distance to genes that have often been associated together. Next, we perform k-means clustering or possibly affinity propagation clustering [Frey and Dueck, Science 2007] based on this distance measure. Candidate clusters may be ranked based on the median “weight” count for the nodes, the mean thickness of edges, or edge density.

**Step 5: Mutant studies in Arabidopsis**: Identify genes central to each network module, and prioritize genes that are poorly characterized in (or even absent from) Arabidopsis. Mutagenize those genes by creating knock-out mutants, “knock-ins” (of the missing gene), or over-expression lines (Aim 2B).

**Step 6: Mutant studies in Maize:** Promising candidate genes from Step 5 will be followed up by mutant studies in Maize. In addition to greenhouse conditions, these mutant lines will be tested in field conditions to ascertain the effect of the gene on the trait of interest.

**Dealing with potential Orthology and Paralogy issues**: By collecting weights supporting nodes and edges from multiple species, the “weighted” network includes all paralogs across species. Because the propagation of edges between all pairs of orthologs may inflate the weight assigned to some pairs of paralogs, we will explore methods to prioritize genes for phenotypic assays by ranking the genes within a network module based on the average number of paralogs and/or gene family members across the species.

**Proof-of Principle Preliminary Analysis:** To test this approach, we use seed development as the trait of interest [Baud and Lepiniec, Progress in lipid research, 2010]. The seed development “trait” has been studied in multiple species, and ample mutant phenotype information is available for *in silico* validation of our network predictions [Meinke D et. al., Trends in plant sciences 2008].

**Construction of gene correlation network(CN)**: We have used deep transcriptome data sets from early seed tissue samples of Soybean and Maize to perform this preliminary analysis [NCBI GEO]. Following the specifications of the first two steps above, we found CNsoybean and CNmaize. Then, we assigned orthologs to Arabidopsis, soybean and maize. For this preliminary work, orthology was assigned based on best reciprocal BLAST matches. More sophisticated approaches to orthology assignment will be used in the final work, as discussed above in Aim 1. This unified network contains XX nodes (YY genes) and ZZ edges.

The distance between nodes, computed as the reciprocal of edge weight [REFERENCE], was used to cluster the nodes. This distance matrix was subjected to k-means clustering to identify conserved clusters of co-expression [REFERENCE].

REST OF THE PRELIMINARY RESULTS WILL DEPEND ON THE RESULTS OBTAINED AT THIS STAGE. WE ANTICIPATE THAT LEC1, LEC2, FUS3, ABI3 AND ABI5 GENES WILL BE REDISCOVERED IN ADDITION TO OTHER GENES THAT INTERACT WITH THEM.

IF ANY UNANNOTATED GENES APPEAR IN THE NETWORK WE CAN FOCUS ON THOSE FOR CAREFUL ANNOTATION TO IDENTIFY POSSIBLE ROLES. ANY GENES MISSING INARABIDOPSIS ARE OBVIOUS CANDIDATES FOR KNOCK-IN STUDIES. SUCH “MISSING” GENES WILL BE RANKED BY AVERAGE NUMBER OF PARALOGS ACROSS SPECIES.

**Aim 2B: Experimental Validation Strategy**. In Aim 2A, we develop and test the method for exploiting data associated with traits in crop species, to inform network studies in Arabidopsis, which ultimately will aid in translational studies back to crop. As proof-of-principle, we tested the seed development trait for which there were ample mutant data with which to validate the genes uncovered in our networks. In the final study, we will apply the same approach to Nitrogen assimilation. An extensive study investigating Maize gene expression changes under various Nitrogen conditions is available from our collaborator [Stephen Moose, Unpublished]. In Aim 2B, we test in Arabidopsis genes important for seed development and Nitrogen assimilation in crops. We will use a medium throughput dexamethasone inducible transient assay system, to assay transcription factors and identify their targets in vivo [Sablowski and Meyerowits Cell 1998]. For TFs that pass initial validation in this transient protoplast system, we will proceed to stable transformants (e.g. T-DNA, overexpression, or “knock in” for cases where the gene is missing in Arabidopsis). For genes that are validated in this approach, mutant lines in Maize will be obtained and field tested for phenotypic effects.

**Experimental method**: We have developed an approach (named “Network Walking”) aimed at analyzing the network targets of a transcription factor in *Arabidopsis* FACS sorted protoplasts. This rapid approach identifies transcription factor targets in less than a week of experimentation, following methods developed by Bargmann and Birnbaum [Bargmann BO, Birnbaum KD (2009) Positive fluorescent selection permits precise, rapid, and in-depth overexpression analysis in plant protoplasts. *Plant Physiol* **149:** 1231-1239.][Bargmann BO, Birnbaum KD (2010) Fluorescence activated cell sorting of plant protoplasts. *J Vis Exp*.] Using Gateway™ technology, we have engineered a vector for which any TF can be fused with a GR (the glucocorticoid receptor) tag. This 35S-TF-GR chimera allows one to i) overproduce the studied TF in the protoplasts, and to ii) control its entrance into the nucleus using dexamethasone (DEX) treatment [Lloyd *et al*, 1994 Lloyd AM, Schena M, Walbot V, Davis RW (1994) Epidermal cell fate determination in Arabidopsis: patterns defined by a steroid-inducible regulator. *Science* **266:** 436-439.][Sablowski and Meyerowitz, 1998 Sablowski RW, Meyerowitz EM (1998) A homolog of NO APICAL MERISTEM is an immediate target of the floral homeotic genes APETALA3/PISTILLATA. *Cell* **92:** 93-103.] [Bargmann BO, Birnbaum KD (2009) Positive fluorescent selection permits precise, rapid, and in-depth overexpression analysis in plant protoplasts. *Plant Physiol* **149:** 1231-1239.][Bargmann BO, Birnbaum KD (2010) Fluorescence activated cell sorting of plant protoplasts. *J Vis Exp*.]. We recently successfully used this “network-walking” approach (FACS assisted protoplast selection + DEX fusion) to study network targets of the well-studied TF, ABI3 [Bargmann et al 2012, In Preparation].

Sentinel genes, predicted to be targets of the TF (based on the “weighted network”), will be assayed by Q-PCR to ascertain any affect this gene might have on their steady state expression levels. A transcription factor that significantly changes the expression level of one or more sentinel genes is assumed to be involved in the regulation (direct or indirect) of that gene. Positive results will be followed up with: (i) transcriptome responses in the protoplast system, and (ii) *in planta* experiments in Arabidopsis (e.g. T-DNA mutants, overexpression) as well as “knock ins”, for genes not present in Arabidopsis. The approach of identifying regulatory hubs, was highly successful in identifying master regulatory genes in N-usage in Arabidopsis [Gutierrez et al 2008]. This “weighted network” approach, which is fueled by gene correlation networks in crops, will enable us to translate candidate genes back to crop species.

Genes validated to affect trait of interest in Arabidopsis will be translated back to Maize using the orthology assignments developed at the earlier stage. Results from the preliminary study in seed development will first be investigated in Maize. Two mutants for each of the ten most promising genes affecting seed development will be identified from the maize transposon library [May BP, Liu H, Vollbrecht E, Senior L, Rabinowicz PD, Roh D, Pan X, Stein L, Freeling M, Alexander D, Martienssen R. Proc Natl Acad Sci U S A. 2003 Sep 30;100(20):11541-6]. Each mutant is assayed for changes in seed development by observing ear phenotypes and/or developmental defects. Where needed, double mutants for candidate genes will be generated by crossing existing mutant lines. In addition, mutant lines for candidate genes from the Nitrogen assimilation network will be identified from all available sources. These mutants will be grown in field conditions under diverse Nitrogen conditions to assay phenotypic changes in ability to assimilate Nitrogen and changes in nitrogen storage.