Gene Coexpression Network Alignment and Conservation of Gene Modules between Two Grass Species: Maize and Rice[C][W][OA]

Stephen P. Ficklin and F. Alex Feltus*

Plant and Environmental Sciences (S.P.F., F.A.F.) and Department of Genetics and Biochemistry (F.A.F.), Clemson University, Clemson, South Carolina 29634

One major objective for plant biology is the discovery of molecular subsystems underlying complex traits. The use of genetic and genomic resources combined in a systems genetics approach offers a means for approaching this goal. This study describes a maize (*Zea mays*) gene coexpression network built from publicly available expression arrays. The maize network consisted of 2,071 loci that were divided into 34 distinct modules that contained 1,928 enriched functional annotation terms and 35 cofunctional gene clusters. Of note, 391 maize genes of unknown function were found to be coexpressed within modules along with genes of known function. A global network alignment was made between this maize network and a previously described rice (*Oryza sativa*) coexpression network. The IsoRankN tool was used, which incorporates both gene homology and network topology for the alignment. A total of 1,173 aligned loci were detected between the two grass networks, which condensed into 154 conserved subgraphs that preserved 4,758 coexpression edges in rice and 6,105 coexpression edges in maize. This study provides an early view into maize coexpression space and provides an initial network-based framework for the translation of functional genomic and genetic information between these two vital agricultural species.

The combination of genomics, genetics, and systems-level computational methods provides a powerful approach toward insight into complex biological systems. Of particular significance is the discovery of genetic interactions that lead to desirable agricultural and economic traits in the Poaceae family (grasses). The Poaceae includes valuable crops such as rice (*Oryza sativa*), maize (*Zea mays*), wheat (*Triticum* spp.), and sugarcane (*Saccharum officinarum*), which are globally some of the most agriculturally and economically important crops (FAOSTAT, 2007). Understanding complex interactions underlying agronomic traits within these species, therefore, is of great significance, in particular to help with crop improvements to meet the challenges of plant and human health but also for basic understanding of complex biological systems.

In addition to their pivotal role in agriculture, grasses offer a powerful model system in that their genomes are closely conserved and functional genomic knowledge gained in one species can be hypothesized to occur in another syntenic region (translational functional genomics; Paterson et al., 2009). In cases of grass species with poorly resolved, polyploid genomes such as sugarcane, where genomic resources are not as far progressed as in other grasses (e.g. rice, sorghum [*Sorghum bicolor*], maize, etc.), translational functional genomics methods may be the most cost-effective strategy for crop improvement as well as for unraveling the functional consequences of polyploidy. Additionally, crops rich in genetically mapped loci deposited in sites like Gramene (Jaiswal, 2011) provide a rich source of systems genetic hypotheses that could in principle accelerate the translation of interacting gene sets associated with complex traits into grasses with poor genetic resources (Ayroles et al., 2009; Wang et al., 2010).

One method of identifying interacting gene sets is through the construction of a gene coexpression network, which is constructed through the discovery of nonrandom gene-gene expression dependencies measured across multiple transcriptome perturbations, often derived from a collection of microarray data sets. During coexpression network construction, the tendency of m transcripts to exhibit similar (or not) expression patterns across a set of n microarrays is determined. In the case where dependency is determined via a correlation metric (e.g. Pearson's r), a comprehensive $m \times m$ matrix of correlation values is generated, which represents expression similarity. The "similarity matrix" is then thresholded to form an "adjacency matrix," which represents an undirected graph where edges (coexpression) exist between two nodes (transcripts) when a correlation value in the matrix is above the significance threshold. Computational methods are then

* Corresponding author; e-mail ffeltus@clemson.edu.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantphysiol.org) is: F. Alex Feltus (ffeltus@clemson.edu).

[C] Some figures in this article are displayed in color online but in black and white in the print edition.

[W] The online version of this article contains Web-only data.

[OA] Open Access articles can be viewed online without a subscription.

www.plantphysiol.org/cgi/doi/10.1104/pp.111.173047

applied to circumscribe groups of network nodes that are highly connected (coexpressed gene "modules"; Langfelder and Horvath, 2008; Li and Horvath, 2009; Chang et al., 2010; Rivera et al., 2010; Xu et al., 2010). It has been shown that genes in these modules participate in similar biological processes; therefore, guilt-by-association inferences can be applied to module genes with no known function that are connected to module genes of known function (Wolfe et al., 2005; Aoki et al., 2007).

Global coexpression networks are those that incorporate expression data from a variety of tissues, developmental stages, and environmental conditions into a single network, the goal being to capture stable coexpression relationships across a diverse collection of experimental perturbations. Global gene coexpression networks maintain similar properties as other naturally occurring networks, such as human social networks and protein-protein interaction networks. These networks tend to be scale free, small world, modular, and hierarchical (Ravasz et al., 2002; Barabási and Oltvai, 2004). Detailed descriptions of these properties can be found in the report by Barabási and Oltvai (2004).

Plant coexpression networks have previously been constructed for Arabidopsis (*Arabidopsis thaliana*; Persson et al., 2005; Wei et al., 2006; Mentzen et al., 2008; Atias et al., 2009; Mao et al., 2009; Wang et al., 2009; Lee et al., 2010; Mutwil et al., 2010), barley (*Hordeum vulgare*; Faccioli et al., 2005), rice (Lee et al., 2009; Ficklin et al., 2010), poplar (*Populus* spp.; Ogata et al., 2009), and tobacco (*Nicotiana tabacum*; Edwards et al., 2010). Several online plant resources also exist for searching coexpression relationships within and sometimes between these species as well as incorporating functional and other data types. These include the Arabidopsis Coexpression Toolkit (Manfield et al., 2006), STARNET 2 (Jupiter et al., 2009), RiceArrayNet (PlantArrayNet; Lee et al., 2009), ATTED-II (Obayashi et al., 2009), the Coexpressed Biological Processes database (Ogata et al., 2010), AtCOECiS (Vandepoele et al., 2009), The Gene Coexpression Network Browser (Ficklin et al., 2010), AraNet (Lee et al., 2010), and a second AraNet (Mutwil et al., 2010). Clearly, there is a burgeoning interest in using a network approach to discover gene-gene dependencies across the field of plant biology.

Given the recent and rapid increase of available biological networks, an important method is the identification of common patterns of connectivity between two networks. Internetwork comparisons are used for several purposes, including improved identification of functional orthologs between species (Bandyopadhyay et al., 2006) and identification of evolutionarily conserved subgraphs, or sets of highly connected genes that demonstrate conserved function (Stuart et al., 2003). Several different network comparison methods exist that perform either local or global comparisons. Local network alignments attempt to align small subsets of nodes between multiple networks, whereas global network alignments attempt to find the best alignment of all nodes in one network with another (Singh et al., 2008). Various heuristics exist for global alignment of two or more networks, and typically these methods first use homology to prioritize the alignment of nodes and then incorporate a measure of topology to refine alignments (Hu et al., 2005; Flannick et al., 2009; Kalaev et al., 2009; Liao et al., 2009; Zaslavskiy et al., 2009; Chindelevitch et al., 2010). Some methods strictly use topology to guide alignments (Kuchaiev et al., 2010), given that network motifs are often conserved in functionally related systems (Milo et al., 2002; Shen-Orr et al., 2002). The majority of network alignment methods have been used to align protein-protein interaction networks, whereas one method has recently been published for the alignment of gene coexpression networks (Zarrineh et al., 2011).

This study adds to the growing compendium of systems-level knowledge for plants by first describing a maize gene coexpression network, and then through a global network alignment with a rice coexpression network (Ficklin et al., 2010) we identified common subgraphs of coexpressed gene sets between the two grass species. For network alignment, we applied a tool, IsoRankN (Liao et al., 2009), which incorporates both gene homology and network topology in its alignment algorithm. The use of homology contributes conservation of sequence, and topology contributes conservation of coexpression—both of which are associated with functional relatedness. We describe the discovery of multiple sets of modules between rice and maize that are both enriched for similar functional terms and that are potentially evolutionarily conserved between the two grasses. This functional similarity between modules in maize and rice seems to agree with the idea that function may be translated through the aligned nodes of two networks. This may serve as a method for identifying functional modules in other grass species. Phenotypic associations available in the rice network may also provide an initial glimpse at the possibilities of translational systems genetics from rice to maize and other cereals. In practice, this method may assist with the prioritization of genes for future mutational studies.

RESULTS

Maize Coexpression Network Construction

The maize coexpression network was constructed using 253 Affymetrix Maize GeneChip Genome Array microarray samples obtained from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) repository. A listing of these array accessions and the experimental conditions under which the transcriptome was measured can be found in Supplemental Table S1. Construction of the maize network was performed using the same method as published previously for rice (Ficklin et al., 2010). Maize microarray data sets were Robust Multichip Average (RMA) normalized (Irizarry et al., 2003), and

40 outlier arrays were removed using the R/arrayQualityMetrics package (Kauffmann et al., 2009). Upon inspection, these outliers seemed to be a result of low-quality hybridizations or nonstandard experimental conditions and did not appear to derive from a common biological system. Next, all pairwise gene expression correlations were determined (Pearson's *r*). The resulting correlation (similarity) matrix was used as input into both the Weighted Correlation Network Analysis (WGCNA) soft-threshold (Langfelder and Horvath, 2008) and Random Matrix Theory (RMT) hard-threshold (Luo et al., 2007) methods for network construction. The WGCNA method identified a power of 6 to power raise the similarity matrix and later divided the network into 34 distinct gene modules, whereas 45 modules were detected for rice (Table I). The relationship between maize modules in terms of similarity of expression is shown in Supplemental Figure S1. The RMT method provided a hard-threshold cutoff value of 0.5781 for the WGCNA power-raised matrix. This is the point where the nearest-neighbor spacing distribution within the network transitions from what would appear as random noise to nonrandom signal (χ^2; $P > 0.001$). The final maize network consisted of 31,983 edges between 2,708 probe sets (2,071 gene models), which corresponds to 15.4% of the original probe sets on the array (Table I). A global view of the maize coexpression network can be seen in Figure 1, where individual modules are distinctly colored. A detailed list of edges for the maize network can be found in Supplemental Table S2. The maize network is available online, along with the previously described rice network, for browsing and searching at http://www.clemson.edu/genenetwork. Network properties, such as node degree and clustering coefficient distributions, can be found in Supplemental Figure S2.

Functional Enrichment and Clustering of Coexpressed Maize Gene Modules

Functional enrichment was performed for each of the 34 modules identified by WGCNA using annotation terms from Gene Ontology (GO; Ashburner et al., 2000), InterPro (Apweiler et al., 2001), and Kyoto Encyclopedia of Genes and Genomes (KEGG; Kanehisa et al., 2008) using an in-house method similar to the tool DAVID (for Database for Annotation, Visualization, and Integrated Discovery; Dennis et al., 2003; Huang et al., 2009). A total of 1,928 unique annotation terms were found to be enriched in the maize modules (Fisher's exact test; $P < 0.1$). Cofunctional clusters, or subsets of nodes within a module that share enriched functional annotation, were identified using the DAVID approach. The identified clusters were sorted first by average connectivity ($<k>$) and second by enrichment score (e-score), the geometric mean of enrichment *P* value. A total of 35 cofunctional gene clusters identified from 596 enriched terms were found within 10 modules. Detailed lists of probe sets and genomic loci within modules and cofunctional clusters, as well as enriched annotation terms, can be found in Supplemental Tables S3 to S6. A total of 383 maize loci are represented in the network with no known functional annotation (Supplemental Table S7). Of these, approximately 50%, or 193 of the 391 genes of unknown function, have 3,092 coexpressed edges with genes in cofunctional clusters (Supplemental Table S8). Therefore, it may be possible to infer function for these loci using the principle of guilt by association.

Interestingly, cofunctional clusters ordered first by $<k>$ in both the maize and rice networks seem quite similar. A list of the top-10 ordered clusters in both networks can be seen in Table II. For example, the highest ordered maize cluster by $<k>$ was enriched

Table I. *Characteristics of the rice and maize networks*

Characteristic	Rice Network	Maize Network
Array	Affymetrix Rice GeneChip	Affymetrix Maize Gene Chip
NCBI GEO accession for array	GPL2025	GPL4032
Probe sets on array	54,168	17,555
Genomic loci mapped to probe sets	46,499	14,792
Microarray samples[a]	508	253
WGCNA selected power threshold	4	6
WGCNA module dendrogram cutoff	0.20	0.20
RMT hard threshold	0.7101	0.5781
Probe sets in network	4,528	2,708
Edges in probe set network	43,144	31,983
Loci in network	2,257	2,071
Edges in loci network	32,820	33,397
Modules	45	34
Enriched terms	2,373	1,928
Functional clusters	76	35
Clustered terms	960	596
Enriched phenotypic terms	17	N/A[b]

[a] Number of samples remaining after outlier detection and removal. [b] N/A, Not applicable.

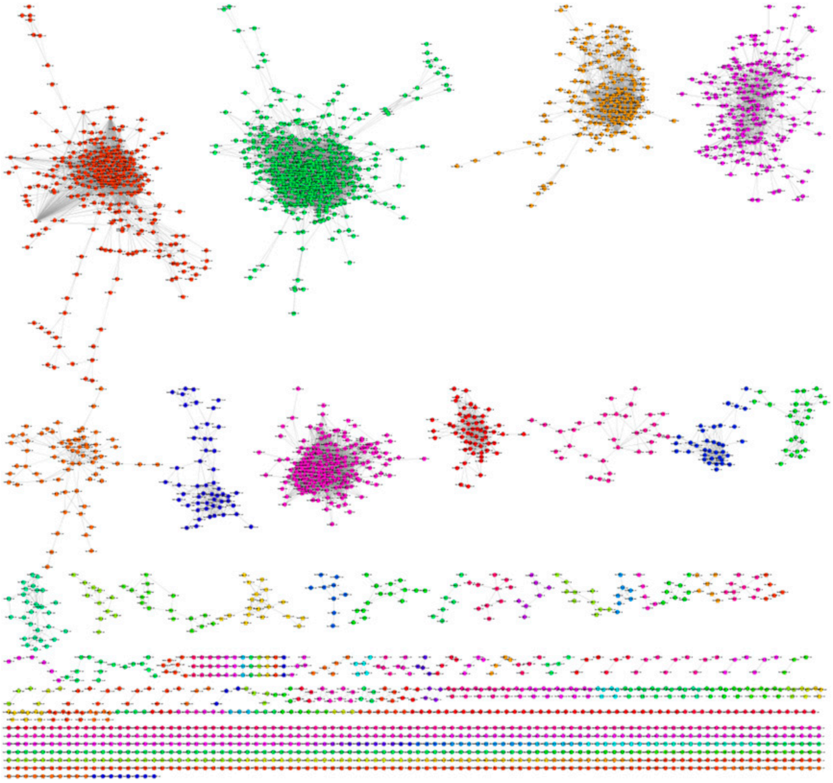

Figure 1. Maize coexpression network. Nodes are probe sets from the Affymetrix GeneChip Maize Genome Array. Edges indicate significant coexpression between probe sets above a hard threshold. The various colors indicate the different modules of the network.

for functional terms related to the ribosome and translation (module ZmM6C25: <k> = 20.2, e-score = 2.62). Similarly, the corresponding rice cluster was also enriched for terms related to the ribosome and translation (module OsM6C25: <k> = 17.0, e-score = 1.98). The second highest maize cluster was enriched for seed storage activity (ZmM5C1: <k> = 10.0, e-score = 3.38), as was the third highest rice cluster (OsM13C1: <k> = 12.9, e-score = 12.22). Additional top-10 ordered clusters with similar annotation terms in both maize and rice, although not in the same order, include clusters enriched for photosynthesis, glycolysis, and microtubule activity.

Rice and Maize Coexpression Network Alignment

Using the constructed maize network, a comparison with the existing rice network was performed with the goal of identifying evolutionarily conserved coexpression patterns. A comparative summary of the statistics for both the rice and maize networks can be seen in Table I. To allow for direct comparison of various network alignment methods, the probe set-based networks were first condensed into a locus-based network that contained 2,071 loci for maize and 2,257 loci for rice (Supplemental Tables S9 and S10). IsoRankN (Liao et al., 2009) was used to perform global alignments between the maize and rice networks. To help clarify the meaning of the various modules, clusters, and subgraphs constructed with this analysis, we provide definitions as well as naming conventions (Table III).

IsoRankN provides three input parameters that affect the results of the network alignments. These parameters include an iteration parameter, a threshold parameter, and an α value. Documentation for IsoRankN indicates that the iteration parameter should vary between 10 and 30, the threshold between 1e-3 and 1e-5, and the α value between 0 and 1. The α value controls the contribution of homology and topology; a value of 0 would strictly use homology for alignments, whereas 1 would strictly use topology. A value of 0.5 would weight equally the contributions of both homology and topology. Therefore, to identify an adequate set of parameters for IsoRankN for alignment of the rice and maize networks, these parameters were varied from one extreme to the other within the suggested documented ranges. In total, 189 tests were performed. To measure the change of the biological signal caused by varying these parameters, we used κ statistics to provide a measure of functional similarity between conserved subgraphs. Functional enrichment was performed for each conserved subgraph in both maize and rice. The similarity of terms enriched in corresponding conserved subgraphs of maize and rice is measured by the κ score, where a value greater than 0 indicates that the conserved subgraph in rice is similar, more than could be expected by chance, to the corresponding subgraph in maize. A value of 1 indicates that the two are identical in terms of enriched terms. A plot of average κ scores and subgraph counts across 20 α values, for an iteration value of 30 and threshold value of 1e-4, is shown in Figure 2. Aside from the

Table II. *Side-by-side functional comparison of top-10 maize and rice cofunctional clusters, ordered by average connectivity*

Maize Cluster[a]	<*k*>[b]	E-Score[c]	Summarized Function	Rice Cluster[a]	<*k*>[b]	E-Score[c]	Summarized Function
ZmM2C1	20.2	2.62	Ribosome/translation	OsM6C25	17.0	1.98	Ribosome/translation
ZmM5C1	10.0	3.38	Seed storage	OsM6C4	14.1	4.13	Photosynthesis/light harvesting
ZmM9C1	10.0	7.25	Histone/DNA binding	OsM13C1	12.9	12.22	Seed storage
ZmM1C1	10.0	1.87	Photosynthesis/light harvesting	OsM6C23	10.5	2.04	Carbon fixation/carotenoid biosynthesis
ZmM4C1	8.8	4.07	Ribosome/translation	OsM6C16	9.2	3.14	Photosynthesis
ZmM2C2	5.7	2.43	Translation elongation	OsM2C2	7.4	5.13	Kinesin/microtubule motor activity
ZmM9C2	5.4	4.31	Histone/DNA binding	OsM6C14	5.4	3.19	Glycolysis
ZmM11C1	5.3	3.05	Kinesin/microtubule motor activity	OsM13C5	5.0	11.24	Transcription factor activity
ZmM1C3	4.0	2.89	Glycolysis	OsM13C2	4.6	3.57	Nutrient reservoir activity
ZmM19C1	3.9	5.38	Transcription factor activity	OsM6C11	3.7	4.08	Ribosome binding/protein folding

[a] Modules are numbered sequentially starting from zero and are prefixed with the letter M. Clusters within a module are numbered sequentially and are prefixed with the letter C. Modules and clusters are prefixed with a species abbreviation: Os for rice and Zm for maize. Thus, cluster 1 from module 8 in rice is named OsM8C1. [b] <*k*> is the average connectivity of the nodes in the cluster. [c] E-score is the enrichment score, or geometric mean of the Fisher's test enrichment *P* values of the cluster.

extreme α values near 0 and 1, the functional similarity of conserved subgraphs is relatively consistent across the α values. The graph in Figure 2 was effectively identical for each combination of iteration and threshold we tested. This similarity indicates that convergence of the alignment occurs at low stringency and that selection of almost any parameter blend for those we selected for testing would be effective. The κ score (or functional similarity) of the subgraphs in rice and maize at an α value of 0 is very high; however, the number of conserved subgraphs at that value is very low. The opposite is true for an α value of 1. It seemed that most parameter sets, aside from the extreme α values, would generate an adequate set of subgraphs with a reasonably high average similarity (average κ), so we selected conserved subgraphs derived from alignments from IsoRankN using an α value of 0.8, an iteration value of 30, and a threshold of 1e-4, because this particular combination of parameters seemed to provide the highest average κ. Because average κ score and subgraph count were very similar across all parameter variations, we only present here a single representative result set. Using these parameters, we detected 1,173 aligned loci, which were later connected into 154 conserved subgraphs. These subgraphs preserved 4,758 edges in rice and 6,105 edges in maize (Supplemental Tables S11 and S12). Functional enrichment and clustering, identical to that performed for the network modules, was performed for these subgraphs as well. The cofunctional clusters of these conserved subgraphs can be found in Supplemental Tables S13 and S14.

Common Function in Conserved Maize-Rice Subgraphs

For the IsoRankN alignments, functional enrichment and κ analysis of the conserved subgraphs yielded nine subgraphs with a perfect κ score of 1, indicating an identical set of enriched terms. These include subgraphs enriched for early nodulin 93 proteins, hydrolase activities, DNA binding, peptidase,

Table III. *Synopsis of subgraph definition and naming conventions*

Term	Definition	Naming Schema[a]
Subgraph	Any collection of nodes and edges that form a subset of the global network	
Module	A subgraph within the global network that consists of highly connected groups of nodes; for this study, modules are determined using the WGCNA method that groups nodes by measures of similarity (Supple mental Fig. S1)	SpMx
Functional cluster	A functional cluster is a subgraph within a module where the nodes have a high degree of similarity in functional terms (e.g. GO, InterPro, and KEGG terms)	SpMxCy
Conserved subgraph	A subgraph that is present in one network and has a corresponding subgraph in another network; these subgraphs share nodes that have been locally aligned using a network alignment tool	subgraph_*z*

[a] For naming, Sp indicates a two-letter species abbreviation; the M in Mx indicates the subgraph is a module where x is the module number; C indicates a cluster followed by the cluster number, y; and z is a four-digit number given to uniquely identify conserved subgraphs.

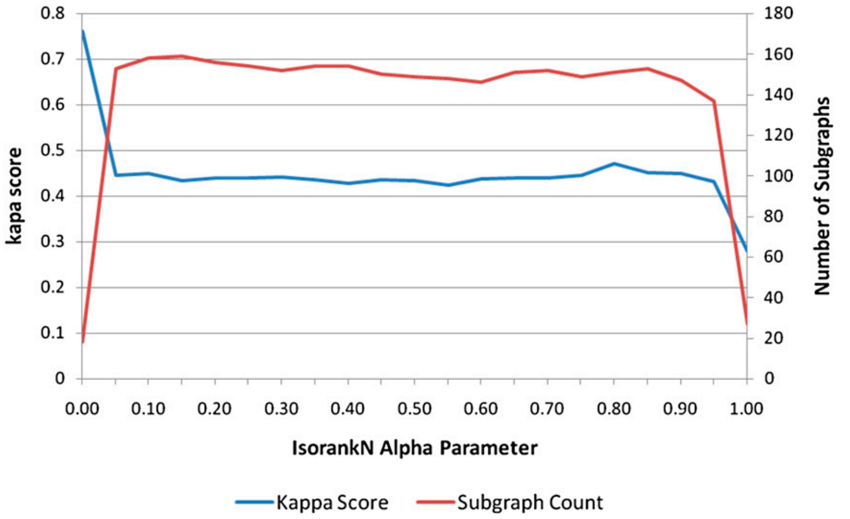

Figure 2. Varying the homology-to-topology ratio has little effect on conserved maize-rice subgraph discovery. This graph shows the distribution of the average κ scores (blue line) and the number of conserved subgraphs (red line) across 20 α values for IsoRankN at an iteration setting of 30 and a threshold value of 1e-4. This graph is a representative plot for 189 trials of IsoRankN where the iteration parameters varied at 10, 20, and 30 and the threshold parameter varied at 1e-3, 1e-4, and 1e-5. Other combinations yielded almost identical graphs. [See online article for color version of this figure.]

nucleosome assembly, transcription factor activity, and others (Supplemental Table S15). However, these subgraphs are relatively small, with two to five edges. The four largest conserved subgraphs are enriched for terms involved in photosynthesis, DNA replication, the ribosome, and starch synthase (Table IV), all of which have a κ score greater than 0.5. Incidentally, these four classes of enriched terms are also present in the top-10 list of enriched clusters as seen in Table II.

For the rice network, phenotypic terms derived from the *Tos17* retrotransposon insertion mutation studies (Hirochika et al., 1996; Miyao et al., 2003) were mapped to loci and included in the functional enrichment and clustering of network modules (Ficklin et al., 2010). Of the 154 conserved subgraphs, 20 conserved subgraphs from rice are enriched for Tos17 phenotypic terms, which include phenotypes such as "sterile," "pale yellow leaf," "high tillering," "vivipary," and more. A listing of these 20 conserved subgraphs can be seen in Table V. The subgraphs are ranked by descending order of κ score, which indicates the similarity of functional annotations between the rice and maize conserved subgraphs. Several subgraphs have a κ score of 1.0, indicating identical functional similarity, but overall, a high degree of similarity between most of the subgraphs is evident. Figure 3 shows the relationship between the global rice and maize networks (Fig. 3, A and C, respectively) with the conserved subgraphs of each (Fig. 3, B and D, respectively) as constructed using IsoRankN node alignments. Light gray lines between the global rice and maize networks simply map the location of nodes with their conserved counterparts. Light gray lines between the two conserved subgraph networks show node alignments provided by IsoRankN. Light red lines indicate node alignments with phenotypic associations in rice. Figure 4 shows a close-up view of conserved subgraph "subgraph_0107."

DISCUSSION

The purpose of this study was to identify conserved, coexpressed gene sets between two vital agricultural species: rice and maize. To identify these gene sets, we first constructed a maize coexpression network, de novo, and aligned it to a previously described rice coexpression network (Ficklin et al., 2010). Our hypothesis was that the discovery of conserved network nodes (genes) and edges would provide an initial framework

Table IV. *Top 10 largest conserved subgraphs by size*

Subgraph	κ Score	Maize Nodes	Rice Nodes	Top Enriched KEGG/GO Term for Maize Conserved Subgraph	Top Enriched KEGG/GO Term For Rice Conserved Subgraph
subgraph_0107	0.64	323	278	GO:0009765 photosynthesis, light harvesting	GO:0015979 photosynthesis
subgraph_0067	0.59	120	95	GO:0000786 nucleosome	GO:0003777 microtubule motor activity
subgraph_0034	0.73	57	35	GO:0005840 ribosome	GO:0005840 ribosome
subgraph_0282	0.51	49	45	K13679 granule-bound starch synthase	K00703 glgA; starch synthase
subgraph_0624	0.15	11	2	GO:0015934 large ribosomal subunit	GO:0005840 ribosome
subgraph_0341	0.09	11	3	K02634 petA; apocytochrome *f*	K02709 psbH; PSII PsbH protein
subgraph_0046	0.87	9	3	GO:0005840 ribosome	GO:0005840 ribosome
subgraph_0031	0.72	9	7	K10999 CESA; cellulose synthase A	K10999 CESA; cellulose synthase A
subgraph_0033	0.10	9	4	GO:0005773 vacuole	GO:0003676 nucleic acid binding
subgraph_0035	0.91	8	2	GO:0015934 large ribosomal subunit	GO:0015934 large ribosomal subunit

Table V. *Top functional terms for conserved subgraphs derived from IsoRankN in maize and rice with phenotypic associations in rice*

Subgraph	κ	Maize Genes	Rice Genes	Rice Phenotypes	Maize Top Enriched GO/IPR[a] Term	Rice Top Enriched GO/IPR Term
subgraph_0065	1.00	4	5	Pale green leaf, long culm, albino, drooping leaf, yellow, low tillering	IPR005050 early nodulin 93 ENOD93 protein	IPR005050 early nodulin 93 ENOD93 protein
subgraph_0908	1.00	2	2	Pale green leaf	GO:0016787 hydrolase activity	GO:0016787 hydrolase activity
subgraph_0060	1.00	5	4	Late heading	GO:0043565 sequence-specific DNA binding	GO:0043565 sequence-specific DNA binding
subgraph_0005	0.89	2	2	Germination rate	GO:0016788 hydrolase activity, acting on ester bonds	GO:0016788 hydrolase activity, acting on ester bonds
subgraph_0046	0.87	9	3	Spl/lesion mimic	GO:0005840 ribosome	GO:0005840 ribosome
subgraph_0907	0.67	2	2	Others	IPR005516 remorin, C-terminal region	IPR005516 remorin, C-terminal region
subgraph_0107	0.64	323	278	Pale green leaf	GO:0009765 photosynthesis	GO:0015979 photosynthesis
subgraph_0777	0.62	2	2	Long culm	IPR010525 auxin response factor	IPR010525 auxin response factor
subgraph_0067	0.59	120	95	Lamina joint, thick culm, lax panicle, high tillering	GO:0000786 nucleosome	GO:0003777 microtubule motor activity
subgraph_0649	0.57	5	2	Vivipary	GO:0005783 endoplasmic reticulum	GO:0005783 endoplasmic reticulum
subgraph_0727	0.56	2	2	Yellow, narrow leaf	GO:0003899 DNA-directed RNA polymerase activity	GO:0004197 Cys-type endopeptidase activity
subgraph_0092	0.47	2	2	Vivipary, yellow	IPR001944 glycoside hydrolase, family 35	IPR000922 D-galactoside/L-Rha-binding SUEL lectin
subgraph_0029	0.44	8	4	Short panicle, dense panicle	GO:0043687 posttranslational protein modification	GO:0005840 ribosome
subgraph_0218	0.20	2	2	Virescent	GO:0006754 ATP biosynthetic process	GO:0005524 ATP binding
subgraph_0621	0.18	4	3	Abnormal shoot	GO:0045735 nutrient reservoir activity	GO:0005215 transporter activity
subgraph_0893	0.17	2	3	Sterile, stripe	GO:0006464 protein modification process	GO:0004197 Cys-type endopeptidase activity
subgraph_0006	0.17	2	2	Vivipary	GO:0009289 fimbrium	GO:0015079 potassium ion transmembrane transporter activity
subgraph_0624	0.15	11	2	Short panicle, abnormal panicle shape, small grain	GO:0015934 large ribosomal subunit	GO:0005840 ribosome
subgraph_0105	0.12	4	3	Rolled leaf	GO:0016857 racemase and epimerase activity	GO:0016020 membrane
subgraph_0599	0.11	3	3	Rolled leaf, pale green leaf	GO:0004871 signal transducer activity	GO:0007155 cell adhesion

[a] IPR, InterPro records.

for the translation of complex functional genomic and genetic knowledge from one species to another. This strategy is complementary to traditional comparative genomic approaches where known function is translated between taxa via homology and/or synteny. Additionally, the WGCNA and RMT tools were selected to preserve a knowledge-independent approach. The networks were thresholded (using RMT) and modules were constructed (using WGCNA) without prior knowledge of the underlying gene functions.

The Global Maize Gene Coexpression Network

Here, we provide, to our knowledge, the first known maize gene coexpression network. This network facilitates research in maize by providing lists of interacting

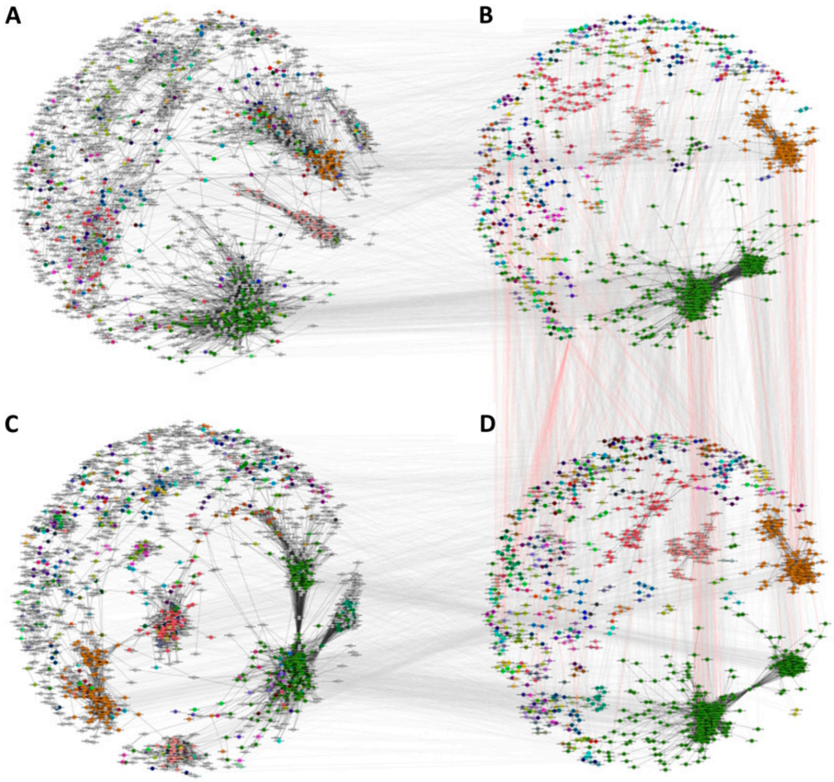

Figure 3. Conserved subgraphs between rice and maize. A, The global locus-based network for rice. B, The conserved network for rice with colored subgraphs. C, The global locus-based network for maize. D, The conserved network for maize with colored subgraphs. Nodes in B and D are color coded according to the conserved subgraphs to which they belong. The same colored nodes in B belong to the same conserved subgraph in D. These same nodes are colored identically in the global networks to show global placement. Nodes colored gray in the global networks are not assigned to a conserved subgraph. Dark-colored edges in the global and conserved subgraphs represent coexpression edges. Lightly colored lines between the global networks in A and C and the conserved subgraphs in B and D simply indicate the positions of the same nodes in both types of networks. Lightly colored gray lines between the conserved subgraphs of rice and maize in B and D show the locations of aligned nodes as indicated by IsoRankN. Lightly colored red lines between B and D originate from the rice conserved subgraph in B and indicate known phenotypic associations in rice with possible translation to maize.

genes annotated for specific biological processes that provide clues to candidate gene (known and novel) involved in those processes. Additionally, 391 genes with unknown function (Supplemental Tables S7 and S8) are coexpressed within modules, and 194 of those unannotated genes are interconnected within 32 different cofunctional modules. For example, cluster ZmM5C1 is the fourth highest ordered cofunctional cluster by <*k*> and contains nine loci. However, there are 24 directly connected neighboring genes that have no ascribed GO, KEGG, or InterPro function. The enriched functional terms for this cluster include seed storage activities. Guilt-by-association inferences would suggest that the 24 genes of unknown function in ZmM5C1 may be involved in seed storage or related processes. Therefore, these genes make interesting, perhaps novel, candidates for understanding the biological process associated with seed storage. In total, 194 genes of unknown function, through 3,092 edges, now suggest inferences for the biological processes summarized by 33 different cofunctional clusters (Supplemental Table S8).

The Small Size of Global Coexpression Networks

The maize network is small in comparison with the number of loci mapped to the probe sets present on the microarray. Using 32,540 gene models from the ZmB73 4a.53 release of the maize genome (Schnable et al., 2009), 14,792 (45%) of the known maize loci were measured on the microarray platform. Of those, only 2,071 loci (14%) were present in the global maize network. We observed a similar phenomenon in rice, where almost 86% of known rice transcripts mapped to the probe sets on the microarray platform; similarly, a low fraction of the measured loci (10%) were present in the final network. Therefore, with regard to the number of potential coexpression relationships, both networks are relatively smaller than what we would expect across the organism's life cycle. The RMT method (Luo et al., 2007) was specifically used to define the threshold to reduce random noise from the final network to ensure that the detected coexpression relationships were strong. Therefore, the small size of the network is most likely caused by relationships lost within the "noise" of the data set, combined with the fact that not all coexpression relationships from all conditions are represented by the data set.

Is it possible to boost the biological signal and increase the gene space fraction captured in coexpression networks? Lowering the significance threshold, even using reasonable methods designed to limit the number of false positives, would increase the number of loci in the network but could reduce the overall quality of the biological signal and possibly confound the interpretation of modules (Perkins and Langston, 2009). Usadel et al. (2009) discuss several reasons that significant coexpression correlations can be lost. These include sample selection, complex interaction types, and selection of normalization and correlation methods. Our data suggest that the rice and maize networks consist primarily of coexpression relationships derived from basal biological processes whose expression is most common across the samples used to build the network.

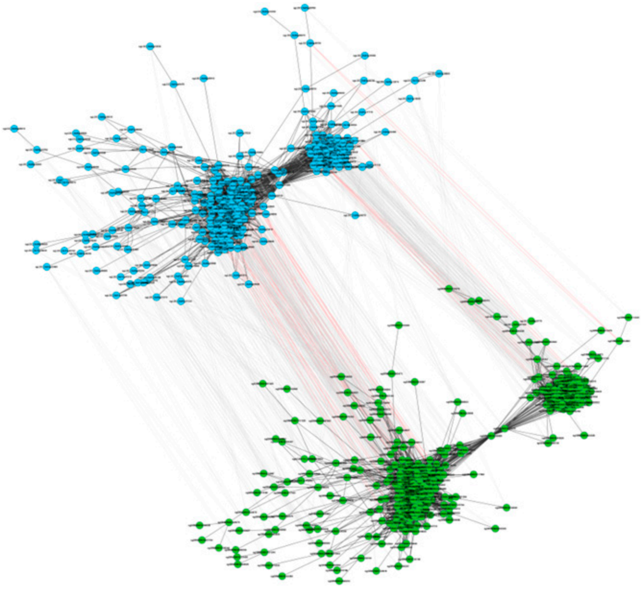

Figure 4. Largest conserved subgraph with implied phenotypic associations. Shown is the largest conserved subgraph, subgraph_0107, between rice (blue nodes) and maize (green nodes). Dark edges in the subgraph are coexpression relationships. Light edges indicate alignments between the two subgraphs determined using IsoRankN. Light red edges indicate phenotypic associations with nodes in rice that are aligned to nodes in maize.

It would seem that the global coexpression networks currently available for plants, including the rice and maize networks we have generated, are immediately useful for these common processes but lack representation of less frequent processes and other subtle interactions. For maize and rice, it may be that more significant coexpression relationships would be detected if (1) additional transcriptome measurements are made from tissue systems not present in the current network, which would increase the sampling frequency and the probability of detecting rarer coexpression relationships; (2) overlap from multiple tissue-specific transcriptomes on a single sample are reduced by segregating data sets to be tissue/condition specific; and (3) additional statistical methods are employed to identify coexpression relationships specific to unique tissues, conditions, or developmental stages, essentially dissecting the input data into subsystems. It should be noted that the detection of coexpression relationships between highly homologous transcripts including gene variants may require extensive transcriptome measurements from a non-hybridization-based platform (e.g. RNAseq) before the full potential of global coexpression networks, measured in the observed number of coexpression relationships, can be realized.

Conservation between Rice and Maize Coexpression Networks

From a qualitative perspective, the apparent collective role of genes in cofunctional clusters from both rice and maize networks, when ordered by average connectivity, were quite similar (Table II). Cofunctional clusters were ordered by connectivity under the premise that highly coexpressed genes are more likely involved in similar biological processes. As mentioned previously, functional terms from processes such as translation, seed storage, glycolysis, photosynthesis, and the cell cycle are all enriched in the top-10 functional clusters of both networks and provide good indication that the two coexpression networks, derived from independent microarray samples for two different species, demonstrate conservation in terms of the connectivity of coexpressed genes for common biological processes.

The apparent conservation of coexpression patterns between rice and maize is further bolstered through a formal global alignment of the two networks via IsoRankN and identification of conserved subgraphs. Many of the conserved subgraphs between rice and maize show a high degree of similarity of enriched functional terms, indicating a high level of conservation, which we quantified using κ statistics (Table IV; Supplemental Table S15). For example, the function of the top-10 conserved subgraphs by size is shown in Table IV. Many of these share similar function, especially when κ scores are closest to 1. This is notable because the likelihood that nodes in the conserved subgraph would be significantly coexpressed, aligned together based on topology and homology, and have nonrandom chance of similarity between their respective enriched function is low. Moreover, modules and cofunctional clusters in maize also align to modules and cofunctional clusters in rice that have similar functional annotations. For example, Figure 5 shows the fourth largest conserved subgraph, “subgraph_0282,” with 49 nodes from maize (bottom left) and 45 from rice (top right). Both the maize and rice loci from this subgraph are enriched for terms involved in seed storage, nutrient reservoir activity, and starch synthase, with a κ score of 0.51. Not only are the functional enrichments similar between aligned nodes, but module coexpression relationships are also maintained. The majority of maize genes in subgraph_0282 belong to module ZmM5 (with only three from ZmM19), and all of the genes from rice belong to module OsM13. Also, cofunctional clusters show evidence of alignment as well. Within this same conserved subgraph, the orange nodes from rice in Figure 5 and the purple nodes from maize are from cofunctional clusters OsM13C1 and ZmM5C1, respectively. Both of these clusters are enriched for seed storage activities, and the nodes from these two clusters have direct alignments with the other.

It should be noted that low κ scores between enriched terms of the rice and maize networks do not indicate that the node alignments are weak. κ scores are based on functional similarity and are dependent on the underlying functional annotation of the loci. For instance, conserved orthologous loci in two genomes may not have been annotated identically yet are aligned due to sequence homology and network to-

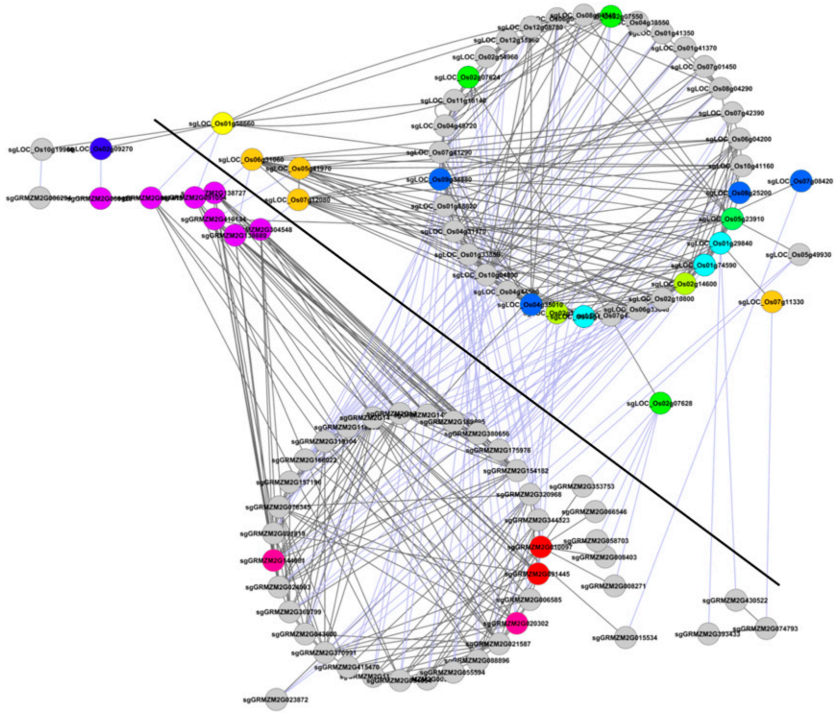

Figure 5. Subgraph_0282 from rice and maize. This subnetwork shows the coexpression edges and conserved alignments for all nodes of subgraph_0282 between maize and rice. Coexpression edges are gray lines, and network alignments are light blue lines. Nodes below the heavy diagonal line are from the maize network, and nodes above it are from rice. All of the rice nodes belong to module OsM13, and the majority of the maize nodes are from module ZmM5, with the exception of the three rightmost nodes in the bottom half, which belong to module Zm19. Yellow nodes in maize are for loci of unknown function. Purple nodes in maize are from cluster ZmM5C1, annotated for nutrient reservoir/seed storage activity. Orange nodes in rice are from cluster OsM13C1, also annotated for nutrient reservoir/seed storage activity. Nodes of other colors belong to the other functional clusters within the module. Gray nodes belong within the subgraph but are not part of a cofunctional cluster.

pology. Also, similar functions may be annotated in somewhat equivalent yet different functional terms. Therefore, a high functional similarity between conserved subgraphs was used to help validate the network alignments, but a lack of functional similarity does not indicate a poor alignment.

Translation of Function and Phenotype

As mentioned previously, conservation between coexpression networks is a powerful tool for validating the correctness of each aligned pair of networks. In essence, conservation reduces the noise within the network because it provides another layer of evidence for coexpression (Obayashi and Kinoshita, 2011). Moreover, the alignment between species strengthens the guilt-by-association inferences made for genes of unknown function. For example, cluster ZmM5C1 was described previously as containing coexpression with 24 genes of unknown function. Seven of the loci from ZmM5C1 appear in conserved subgraph_0282 (purple nodes in Fig. 5). Guilt-by-association inferences may be applied to these genes of unknown function; however, the inference is made stronger because the coexpression relationships are conserved.

A powerful application of alignments between rice and maize networks is the potential to translate gene sets with enriched phenotypes from rice to maize through conserved subgraphs. For instance, Table V provides a list of conserved subgraphs that have enriched phenotypes in rice. These terms are not only present in annotations of genes in the subgraph but enriched. Of particular note is subgraph_0065, with five genes in rice, four genes in maize, and six phenotypic terms: pale green leaf, long culm, albino, drooping leaf, yellow, and low tillering. This subgraph has a κ score of 1, indicating perfect similarity between annotated terms, and is annotated as early nodulin 93 protein. It may be that this high level of similarity is due in part to the fact that sequence homology was employed in network alignment, and sequence homology is often used to transfer functional annotation from one species to another. However, network topology based on coexpression edges was weighted more strongly in the IsoRankN alignment, indicating that coexpression relationships are also maintained between rice and maize alignments. Therefore, it seems appropriate that these phenotypic associations from rice can be inferred to the four maize genes as well as connected neighbors in the subgraph.

CONCLUSION

Gene coexpression network alignments coupled with genetic and functional genomic data provide a method for translation of gene function and genotype-phenotype associations between species, and they are especially useful for species with limited genetic resources. Experimental evidence will be needed to determine the true predictive power of coexpression relationships (intranetwork and internetwork), but the functional similarity we observed in conserved subgraphs seems quite promising. Still, better quantitative measures of biological signal are needed to validate the coexpression relationships. If an in silico metric for biological signal can be identified, it would provide a means to calculate type I and type II errors under alternate network construction protocols. However, given that gene coexpression networks have already been used to successfully identify candidate genes for specific traits (Lee et al., 2010; Mutwil et al., 2010), it is natural to conclude that function and phenotype can also be transferred across species to help identify genes involved in complex traits. The power of this translational systems genetics approach will be increasingly more useful as more genetic data are made available for grasses, especially in the form of genome-wide association studies. In particular, the translation of function and phenotype into large polyploid species, such as sugarcane, would be especially powerful because the capture of genetic associations can be difficult and expensive and genome resources tend to lag behind those of less complex species.

MATERIALS AND METHODS

Maize Network Construction

The method used for construction of the maize (*Zea mays*) gene coexpression network was identical to that previously described for the rice (*Oryza sativa*) gene coexpression network (Ficklin et al., 2010). A total of 293 samples from the Affymetrix Maize GeneChip Genome Array microarray were obtained from NCBI's GEO repository. RMA normalization (Irizarry et al., 2003) using the software package RMAExpress (Bolstad, 2010) and outlier detection using the arrayQualityMetrics (Kauffmann et al., 2009) tool for Bioconductor (Gentleman et al., 2004) were used to remove outlier samples. Arrays that failed all three outlier tests were excluded from further analysis. Then, a similarity matrix was constructed by performing pairwise Pearson correlations for every probe set across all samples. We selected Pearson correlation because it was commonly supported by both the WGCNA and RMT tools. Next, the WGCNA package (Langfelder and Horvath, 2008) was used to convert the similarity matrix into an adjacency matrix by raising the similarity matrix to a power of 6. The power chosen is one that best approximates scale-free behavior in the resulting network and is selected by the software. Finally, the RMT algorithm (Luo et al., 2007) was used to select a hard threshold that limits the noise in the resulting network.

Functional Enrichment and Clustering

The gene models used for this study were from the maize B73 genome (Schnable et al., 2009) version 4.53a obtained from the maizesequence.org Web site. GO (Ashburner et al., 2000), InterPro (Apweiler et al., 2001), and KEGG (Kanehisa et al., 2008) terms were used for functional annotation of these gene models. In the case of GO and InterPro terms, these were obtained directly from the maizesequence.org Web site. KEGG terms were obtained by uploading maize coding sequences to the KEGG/KAAS server, which maps KEGG terms using a homology-based method (Moriya et al., 2007). An in-house tool similar to the online DAVID tool (Dennis et al., 2003; Huang et al., 2009) was used to perform functional enrichment using a Fisher's exact test against each network module and the genome background. Modules were further subdivided into functional clusters using pairwise κ statistics between all genes. Functional clusters were ordered by the geometric mean of the Fisher's *P* values, the e-score, or the $<k>$, which provides a measure of interconnectedness of the nodes in the functional cluster.

Maize-Rice Network Comparison

The maize network was compared with the previously described rice network (Ficklin et al., 2010). The maize and rice networks, as well as functional enrichment and cluster discovery, were constructed with an identical protocol. However, as a result of improvements to the in-house scripts that perform functional enrichment, the functional enrichment and clustering were performed again for the rice network before comparison. The maize and rice networks, including both the original and updated functional enrichment results for rice, are available online at http://www.clemson.edu/genenetwork.

Before network comparisons were performed, nodes in both the rice and maize networks were converted from microarray probe sets to genomic loci. In some cases, these were one-to-one mappings between probe sets and genes. However, some microarray probe sets map to more than one genomic locus and vice versa. These mappings are ambiguous but were retained with the assumption that a significant edge to these nodes could be informative because one or more mapped genes would be producing the correlated transcript. During conversion from a probe set to a loci-based network, edges were placed between two loci whenever they mapped to connected probe sets. Edges were also preserved in cases where a single locus mapped to more than one probe set in a different module.

Network comparisons between the rice and maize gene coexpression networks were performed using IsoRankN (Liao et al., 2009), which provided a mixed topology and homology-based global alignment methodology. First, the maize and rice protein sequence data sets were obtained from the Michigan State University version 6.0 assembly for rice (Ouyang et al., 2007) and the maize B73 genome (Schnable et al., 2009) version 4.53a and were aligned against one another using BLASTP (with expectation value [-e]: 1e-6; filtering options [-F]: 'm S'; local Smith-Waterman alignments [-s]: T; single result per query [-b]: 1; and in tabular output [-m]: 8) following the recommendations given by Moreno-Hagelsieb and Latimer (2008) for selecting BLAST parameters for reciprocal best hits. The homolog scores derived from BLAST and the network edges list were used as input to IsoRankN. Several iterations were performed by varying the parameters for the software. IsoRankN's own iteration parameter was adjusted at values of 10, 20, and 30. The threshold parameter was adjusted at values of 1e-3, 1e-4, and 1e-5, and the α value, which controls the contribution of topology versus homology in aligning the networks, was varied from 0.0 to 1.0 in 0.1 increments. In total, 189 iterations were performed for IsoRankN. IsoRankN generates sets of one-to-many mappings, where in some cases multiple aligned loci are in a single set. Each pair or group was referred to as an alignment set. Conserved subgraphs were generated using these output files with an in-house Perl script. Conserved subgraphs were constructed in a two-step method. The first step selected edges that were conserved between the two networks, and the second step identified subgraphs of interconnected loci.

The process for selecting preserved edges was performed by comparing two loci from two different alignment sets in one network with two loci from the same alignment sets from the other network. If an edge existed in both networks using the four selected loci, then both edges were marked as conserved. The following pseudocode describes the process:

```
S = the array of aligned sets of loci for each set in S as $s_i$
  for each set in S as $s_j$ where $s_i$ is not $s_j$
    for each locus $l_i$ in $s_i$
      for each locus $l_j$ in $s_j$
        for each locus $k_i$ in $s_i$ where $k_i$ is not $l_i$
          for each locus $k_j$ in $s_i$ where $k_i$ is not $k_j$
            if $l_i$ and $l_j$ are in the same network and connected and
              if $k_i$ and $k_j$ are in the same network and connected
                then mark both edges as conserved.
```

Edges and nodes that were not marked as conserved were discarded, and the remaining networks, one for rice and the other for maize, became the "conserved" subnetworks.

Finally, conserved subgraphs within the conserved networks were identified by first selecting an edge from one conserved network to serve as a seed for the subgraph. The aligned loci in the other conserved network were also used as a seed. Thus, the process of defining subgraphs was performed in parallel in both networks. Next, the edges of all of the connected neighbors of the seed were added to the subgraph. The process was continued by iterating recursively through the neighbors and adding their edges until all possible edges were exhausted. Then, a new edge, which had not yet been added to a subgraph, was selected to act as the next seed until all edges in the conserved networks were placed in subgraphs. These subgraphs were labeled numerically, and a label for a subgraph in rice was the same for the corresponding conserved subgraph in maize and vice versa.

Functional enrichment was then performed for each subgraph using the same method described previously for modules in the global network. Subgraphs were then compared using κ statistics. As described previously, κ statistics were used to provide a measure of similarity between the functionally enriched terms of genes in a network module. Here, κ statistics are used to provide a measure of similarity between the two conserved subgraphs of maize and rice that have the same label. Subgraphs are then ranked by κ score from greatest to smallest. Conserved subgraphs are given a four-digit unique number prefixed with the word "subgraph."

Supplemental Data

The following materials are available in the online version of this article.

Supplemental Figure S1. Maize network module clustering.

Supplemental Figure S2. Network characteristics.

Supplemental Table S1. Microarray samples used in network construction.

Supplemental Table S2. Edges from the probe set-based maize network.

Supplemental Table S3. Maize locus identifiers in functionally enriched clusters.

Supplemental Table S4. Maize Affymetrix probe set identifiers in functionally enriched clusters.

Supplemental Table S5. Enriched functional terms in maize clusters.

Supplemental Table S6. Enriched functional terms in maize modules.

Supplemental Table S7. Maize genes in modules with no annotated function.

Supplemental Table S8. First neighbors of maize genes in modules with no annotated function.

Supplemental Table S9. Edges from the locus-based rice network.

Supplemental Table S10. Edges from the locus-based maize network.

Supplemental Table S11. Rice conserved subgraph edges from IsoRankN alignments.

Supplemental Table S12. Maize conserved subgraph edges from IsoRankN alignments.

Supplemental Table S13. Rice conserved subgraph functional clusters.

Supplemental Table S14. Maize conserved subgraph functional clusters.

Supplemental Table S15. κ rankings of IsoRankN-based conserved subgraphs.

Received January 23, 2011; accepted May 20, 2011; published May 23, 2011.

LITERATURE CITED

Aoki K, Ogata Y, Shibata D (2007) Approaches for extracting practical information from gene co-expression networks in plant biology. Plant Cell Physiol **48:** 381–390

Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti L, Corpet F, Croning MD, et al (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Res **29:** 37–40

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al (2000) Gene Ontology: tool for the unification of biology. Nat Genet **25:** 25–29

Atias O, Chor B, Chamovitz DA (2009) Large-scale analysis of Arabidopsis transcription reveals a basal co-regulation network. BMC Syst Biol **3:** 86

Ayroles JF, Carbone MA, Stone EA, Jordan KW, Lyman RF, Magwire MM, Rollmann SM, Duncan LH, Lawrence F, Anholt RR, et al (2009) Systems genetics of complex traits in Drosophila melanogaster. Nat Genet **41:** 299–307

Bandyopadhyay S, Sharan R, Ideker T (2006) Systematic identification of functional orthologs based on protein network comparison. Genome Res **16:** 428–435

Barabási AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet **5:** 101–113

Bolstad BM (2010) RMAExpress. http://rmaexpress.bmbolstad.com/ (January 1, 2011)

Chang RL, Luo F, Johnson S, Scheuermann RH (2010) Deterministic graph-theoretic algorithm for detecting modules in biological interaction networks. Int J Bioinform Res Appl **6:** 101–119

Chindelevitch L, Liao CS, Berger B (2010) Local optimization for global alignment of protein interaction networks. Pac Symp Biocomput **15:** 123–132

Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol **4:** 3

Edwards KD, Bombarely A, Story GW, Allen F, Mueller LA, Coates SA, Jones L (2010) TobEA: an atlas of tobacco gene expression from seed to senescence. BMC Genomics **11:** 142

Faccioli P, Provero P, Herrmann C, Stanca AM, Morcia C, Terzi V (2005) From single genes to co-expression networks: extracting knowledge from barley functional genomics. Plant Mol Biol **58:** 739–750

FAOSTAT (2007) Food and Agricultural Organization of the United Nations, Commodities Production Statistics. http://faostat.fao.org/site/339/default.aspx (January 1, 2011)

Ficklin SP, Luo F, Feltus FA (2010) The association of multiple interacting genes with specific phenotypes in rice using gene coexpression networks. Plant Physiol **154:** 13–24

Flannick J, Novak A, Do CB, Srinivasan BS, Batzoglou S (2009) Automatic parameter learning for multiple local network alignment. J Comput Biol **16:** 1001–1022

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, et al (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol **5:** R80

Hirochika H, Sugimoto K, Otsuki Y, Tsugawa H, Kanda M (1996) Retrotransposons of rice involved in mutations induced by tissue culture. Proc Natl Acad Sci USA **93:** 7783–7788

Hu H, Yan X, Huang Y, Han J, Zhou XJ (2005) Mining coherent dense subgraphs across massive biological networks for functional discovery. Bioinformatics (Suppl 1) **21:** i213–i221

Huang W, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc **4:** 44–57

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics **4:** 249–264

Jaiswal P (2011) Gramene database: a hub for comparative plant genomics. Methods Mol Biol **678:** 247–275

Jupiter D, Chen H, VanBuren V (2009) STARNET 2: a Web-based tool for accelerating discovery of gene regulatory networks using microarray co-expression data. BMC Bioinformatics **10:** 332

Kalaev M, Bafna V, Sharan R (2009) Fast and accurate alignment of multiple protein networks. J Comput Biol **16:** 989–999

Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, et al (2008) KEGG for linking genomes to life and the environment. Nucleic Acids Res (Database issue) **36:** D480–D484

Kauffmann A, Gentleman R, Huber W (2009) arrayQualityMetrics: a Bioconductor package for quality assessment of microarray data. Bioinformatics **25:** 415–416

Kuchaiev O, Milenkovic T, Memisevic V, Hayes W, Przulj N (2010) Topological network alignment uncovers biological function and phylogeny. J R Soc Interface **7:** 1341–1354

Langfelder P, Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics **9:** 559

Lee I, Ambaru B, Thakkar P, Marcotte EM, Rhee SY (2010) Rational association of genes with traits using a genome-scale gene network for Arabidopsis thaliana. Nat Biotechnol **28:** 149–156

Lee TH, Kim YK, Pham TT, Song SI, Kim JK, Kang KY, An G, Jung KH, Galbraith DW, Kim M, et al (2009) RiceArrayNet: a database for correlating gene expression from transcriptome profiling, and its application to the analysis of coexpressed genes in rice. Plant Physiol **151:** 16–33

Li A, Horvath S (2009) Network module detection: affinity search technique with the multi-node topological overlap measure. BMC Res Notes **2:** 142

Liao CS, Lu KH, Baym M, Singh R, Berger B (2009) IsoRankN: spectral methods for global alignment of multiple protein networks. Bioinformatics **25:** i253–i258

Luo F, Yang Y, Zhong J, Gao H, Khan L, Thompson DK, Zhou J (2007) Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory. BMC Bioinformatics **8:** 299

Manfield IW, Jen CH, Pinney JW, Michalopoulos I, Bradford JR, Gilmartin PM, Westhead DR (2006) Arabidopsis Co-expression Tool (ACT): Web server tools for microarray-based gene expression analysis. Nucleic Acids Res (Web Server issue) **34:** W504–W509

Mao L, Van Hemert JL, Dash S, Dickerson JA (2009) Arabidopsis gene co-expression network and its functional modules. BMC Bioinformatics **10:** 346

Mentzen WI, Peng J, Ransom N, Nikolau BJ, Wurtele ES (2008) Articulation of three core metabolic processes in Arabidopsis: fatty acid biosynthesis, leucine catabolism and starch metabolism. BMC Plant Biol **8:** 76

Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U (2002) Network motifs: simple building blocks of complex networks. Science **298:** 824–827

Miyao A, Tanaka K, Murata K, Sawaki H, Takeda S, Abe K, Shinozuka Y, Onosato K, Hirochika H (2003) Target site specificity of the Tos17 retrotransposon shows a preference for insertion within genes and against insertion in retrotransposon-rich regions of the genome. Plant Cell **15:** 1771–1780

Moreno-Hagelsieb G, Latimer K (2008) Choosing BLAST options for better detection of orthologs as reciprocal best hits. Bioinformatics **24:** 319–324

Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M (2007) KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res (Web Server issue) **35:** W182–W185

Mutwil M, Usadel B, Schütte M, Loraine A, Ebenhöh O, Persson S (2010) Assembly of an interactive correlation network for the Arabidopsis genome using a novel heuristic clustering algorithm. Plant Physiol **152:** 29–43

Obayashi T, Hayashi S, Saeki M, Ohta H, Kinoshita K (2009) ATTED-II provides coexpressed gene networks for Arabidopsis. Nucleic Acids Res (Database issue) **37:** D987–D991

Obayashi T, Kinoshita K (2011) COXPRESdb: a database to compare gene coexpression in seven model animals. Nucleic Acids Res (Database issue) **39:** D1016–D1022

Ogata Y, Suzuki H, Sakurai N, Shibata D (2010) CoP: a database for characterizing co-expressed gene modules with biological information in plants. Bioinformatics **26:** 1267–1268

Ogata Y, Suzuki H, Shibata D (2009) A database for poplar gene co-expression analysis for systematic understanding of biological processes, including stress responses. J Wood Sci **55:** 395–400

Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, Childs K, Thibaud-Nissen F, Malek RL, Lee Y, Zheng L, et al (2007) The TIGR Rice Genome Annotation Resource: improvements and new features. Nucleic Acids Res (Database issue) **35:** D883–D887

Paterson AH, Bowers JE, Feltus FA, Tang HB, Lin LF, Wang XY (2009) Comparative genomics of grasses promises a bountiful harvest. Plant Physiol **149:** 125–131

Perkins AD, Langston MA (2009) Threshold selection in gene co-expression networks using spectral graph theory techniques. BMC Bioinformatics (Suppl 11) **10:** S4

Persson S, Wei H, Milne J, Page GP, Somerville CR (2005) Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets. Proc Natl Acad Sci USA **102:** 8633–8638

Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási AL (2002) Hierarchical organization of modularity in metabolic networks. Science **297:** 1551–1555

Rivera CG, Vakil R, Bader JS (2010) NeMo: network module identification in Cytoscape. BMC Bioinformatics (Suppl 1) **11:** S61

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, et al (2009) The B73 maize genome: complexity, diversity, and dynamics. Science **326:** 1112–1115

Shen-Orr SS, Milo R, Mangan S, Alon U (2002) Network motifs in the transcriptional regulation network of Escherichia coli. Nat Genet **31:** 64–68

Singh R, Xu J, Berger B (2008) Global alignment of multiple protein interaction networks. Pac Symp Biocomput **13:** 303–314

Stuart JM, Segal E, Koller D, Kim SK (2003) A gene-coexpression network for global discovery of conserved genetic modules. Science **302:** 249–255

Usadel B, Obayashi T, Mutwil M, Giorgi FM, Bassel GW, Tanimoto M, Chow A, Steinhauser D, Persson S, Provart NJ (2009) Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. Plant Cell Environ **32:** 1633–1651

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

Wang K, Li M, Hakonarson H (2010) Analysing biological pathways in genome-wide association studies. Nat Rev Genet **11:** 843–854

Wang Y, Hu Z, Yang Y, Chen X, Chen G (2009) Function annotation of an SBP-box gene in Arabidopsis based on analysis of co-expression networks and promoters. Int J Mol Sci **10:** 116–132

Wei H, Persson S, Mehta T, Srinivasasainagendra V, Chen L, Page GP, Somerville C, Loraine A (2006) Transcriptional coordination of the metabolic network in Arabidopsis. Plant Physiol **142:** 762–774

Wolfe CJ, Kohane IS, Butte AJ (2005) Systematic survey reveals general applicability of “guilt-by-association” within gene coexpression networks. BMC Bioinformatics **6:** 227

Xu G, Bennett L, Papageorgiou LG, Tsoka S (2010) Module detection in complex networks using integer optimisation. Algorithms Mol Biol **5:** 36

Zarrineh P, Fierro A, Sánchez-Rodríguez A, De Moor B, Engelen K, Marchal K (2011) COMODO: an adaptive coclustering strategy to identify conserved coexpression modules between organisms. Nucleic Acids Res **39:** e41

Zaslavskiy M, Bach F, Vert JP (2009) Global alignment of protein-protein interaction networks by graph matching methods. Bioinformatics **25:** i259–i267