Construction and Optimization of a Large Gene Coexpression Network in Maize Using RNA-Seq Data[1][OPEN]

Ji Huang,[a] Stefania Vendramin,[a] Lizhen Shi,[b] and Karen M. McGinnis[a,2]

[a]Department of Biological Science, Florida State University, Tallahassee, Florida 32306
[b]Department of Computer Science, Florida State University, Tallahassee, Florida 32306

ORCID IDs: 0000-0002-6182-800X (J.H.); 0000-0001-6612-3570 (S.V.); 0000-0002-9564-8146 (K.M.M.).

With the emergence of massively parallel sequencing, genomewide expression data production has reached an unprecedented level. This abundance of data has greatly facilitated maize research, but may not be amenable to traditional analysis techniques that were optimized for other data types. Using publicly available data, a gene coexpression network (GCN) can be constructed and used for gene function prediction, candidate gene selection, and improving understanding of regulatory pathways. Several GCN studies have been done in maize (*Zea mays*), mostly using microarray datasets. To build an optimal GCN from plant materials RNA-Seq data, parameters for expression data normalization and network inference were evaluated. A comprehensive evaluation of these two parameters and a ranked aggregation strategy on network performance, using libraries from 1266 maize samples, were conducted. Three normalization methods and 10 inference methods, including six correlation and four mutual information methods, were tested. The three normalization methods had very similar performance. For network inference, correlation methods performed better than mutual information methods at some genes. Increasing sample size also had a positive effect on GCN. Aggregating single networks together resulted in improved performance compared to single networks.

Maize (*Zea mays*) is the most widely produced crop in United States, and U.S. agriculture accounted for 36% of world maize production in 2015 (USDA, 2016). Maize has also been in the center of genetics research for more than 100 years, including McClintock's pioneering work with transposable elements (reviewed by McClintock, 1983; Fedoroff, 2012). Due to recent technological advances in nucleic acid sequencing and the availability of the maize genome sequence (Schnable et al., 2009), maize genomics research has been greatly expedited.

RNA-sequencing (RNA-Seq) has become the favored technique for detecting genomewide expression patterns. RNA-Seq has some advantages over microarray analysis of gene expression, including single base-pair resolution, detection of novel transcripts, and the ability to analyze transcript abundance without existing genome information (reviewed by Wang et al., 2009; Han et al., 2015; Conesa et al., 2016). RNA-Seq data provides information about single nucleotide polymorphisms, which facilitates genomewide association studies (Fu et al., 2013; Li et al., 2013a; Lonsdale et al., 2013; Fadista et al., 2014). Because of its widespread adaptability, greater than 5000 Illumina platform Maize RNA-Seq libraries (Fig. 1A) are available in the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) database (Leinonen et al., 2010), adding to the body of data that can be used to study the maize genome.

The maize genome is large and heterogeneous, and the genome annotation is still far from complete (Cigan et al., 2005; Ficklin and Feltus, 2011). Although recent work has made substantial progress toward describing genomewide expression patterns in many genotypes, environmental conditions, and tissues, relatively little is known about the function and regulation of most maize genes. Because genes with related biological functions or regulatory mechanisms often have similar expression patterns (Aoki et al., 2007), one way to enhance understanding of gene function is by construction of a gene coexpression network (GCN; D'haeseleer et al., 2000; Aoki et al., 2007; Usadel et al., 2009; Li et al., 2015c; Serin et al., 2016). GCNs are constructed using data mining tools and algorithms that describe the relatedness between the expression patterns of multiple genes in a pairwise fashion.

The use of GCNs predates the availability of RNA-Seq expression data (Ficklin and Feltus, 2011; Sato et al., 2011; De Bodt et al., 2012), meaning that these approaches were initiated and optimized predominantly with microarray datasets. Maize RNA-Seq samples are

[1] Funding for this work came from the National Science Foundation (NSF).

[2] Address correspondence to mcginnis@bio.fsu.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: Dr. Karen McGinnis (mcginnis@bio.fsu.edu).

J.H. and K.M.M. designed the experiments; J.H. conducted experiments. J.H. and S.V. analyzed the data; J.H., K.M.M., and S.V. interpreted the data; L.S. and J.H. made the website; J.H., K.M.M., and S.V. wrote the article.

[OPEN] Articles can be viewed without a subscription.
www.plantphysiol.org/cgi/doi/10.1104/pp.17.00825

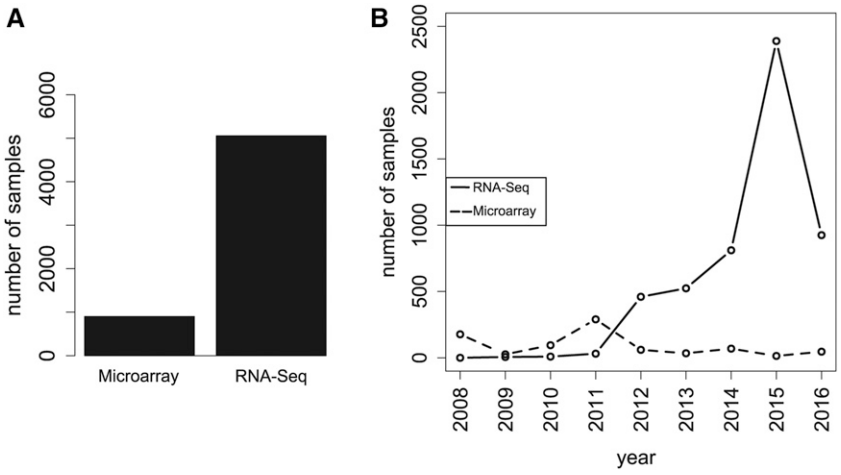

Figure 1. Number of maize microarray and RNA-Seq samples submitted to NCBI (https://www.ncbi.nlm.nih.gov) from 2008 to 2016. A, A text search of the NCBI Gene Expression Omnibus database identified samples generated by microarray platforms GPL4032 and GPL12620, the total values for years 2008 to 2016 were combined to represent the number of microarray studies (Microarray). A text search of the NCBI SRA database was used to identify RNA-seq samples generated between 2008 and 2016 using the Illumina sequencing platform (RNA-Seq). B, Number of samples submitted to NCBI Gene Expression Omnibus database each year generated by microarray platform GPL4032 and GPL12620 were identified by a text search (dashed line) and compared to the number of RNA-Seq Illumina samples (solid line) per year 2008 to 2016.

already five times more abundant than microarray (Fig. 1) and increasing in number, meaning that an RNA-Seq oriented maize GCN protocol would be valuable to the scientific community. Although the initial inputs and results from microarray and RNA-Seq are similar, there are many differences between the data types and analytical approaches. It is therefore anticipated that some adjustments to GCN parameters may improve the efficacy of GCN analysis of RNA-Seq data. GCN construction is typically a multistep process starting with normalization of input datasets, network inference, and network evaluation and interpretation (Supplemental Fig. S1).

Both RNA-Seq and microarrays are affected by systematic variations (Park et al., 2003; Oshlack and Wakefield, 2009; Zheng et al., 2011; Li et al., 2014b). Therefore, genomewide expression results generated by either technique need to be normalized before analysis (Dillies et al., 2013; Li et al., 2015b). Variance stabilizing transformation (VST), counts per million (CPM), and reads per kilobase million (RPKM) are three popular normalization methods for RNA-Seq experiments (Mortazavi et al., 2008; Anders and Huber, 2010; Rau et al., 2013).

Some work has been done to evaluate the efficacy of different normalization methods for expression analysis. Giorgi et al. (2013) showed VST normalization of RNA-Seq data resulted in a GCN with similar characteristics to a microarray-supported network in terms of coefficient and node degree distribution. Normalizations with CPM and using the trimmed mean of *m*-values to adjust the composition bias between RNA-Seq datasets by calculating normalization factors (Robinson et al., 2010) increased the robustness of analysis among diverse library sizes and compositions (Dillies et al., 2013). These studies suggest that optimizing normalization methods might improve GCN performance.

There are several methods for gene network inference, including correlation, mutual information (MI), Bayesian network, and probabilistic graphical models. Typically, correlation and MI methods are used for constructing large-scale GCNs with more than 10,000 genes (Krouk et al., 2013). Correlation methods include Pearson correlation coefficient (PCC), Spearman's correlation coefficient (SCC), Kendall rank correlation coefficient (KCC), Gini correlation coefficient (GCC), and biweight midcorrelation (BIC; Langfelder and Horvath, 2008; Kumari et al., 2012; Ma and Wang, 2012; Ballouz et al., 2015). Cosine similarity coefficient (CSC) has also been used for computing similarities in sparse datasets, such as text (Dhillon and Modha, 2001) and protein-protein interaction data (Luo et al., 2015). MI methods include accurate cellular networks (ARACNE), minimum redundancy network (MRNET), and context likelihood of relatedness (CLR; Margolin et al., 2006; Faith et al., 2007; Meyer et al., 2007). The network inference method might also influence GCN performance.

Several resources are already available for GCN analysis in maize, including COB (Schaefer et al., 2014), CORNET (De Bodt et al., 2012), CoP (Ogata et al., 2010), PLANEX (Yim et al., 2013), and ATTED-II (Obayashi et al., 2009). All of the databases except ATTED-II used PCC to build GCN from 128 to 379 microarray datasets. ATTED-II recently updated their database to provide both GCNs from microarray and RNA-Seq using PCC-based mutual rank (Aoki et al., 2016). Although PCC is widely used, there is very limited evidence to indicate that it is the optimal approach for GCN analyses.

GCNs could also be improved by metaanalysis using ranked aggregation from individual networks (Zhong et al., 2014; Ballouz et al., 2015; Wang et al., 2015a). By aggregating individual experiments, only interactions consistent among networks are preserved, which helps reduce noise and highlights conserved interactions. Furthermore, the ranked aggregation method provides a way to efficiently increase the size of the aggregated network with newly available datasets, and recalculation with all datasets is not required when a new one is added. This provides an efficient way to process and incorporate emerging information.

Herein, an extensive evaluation in constructing maize GCNs is reported. Three methods were tested: the normalization method, the network inference algorithm, and the ranked aggregation method. To our knowledge, this is the first comprehensive attempt at optimizing GCN construction using plant RNA-Seq datasets. The network is publicly accessible at http://www.bio.fsu.edu/mcginnislab/mcn/main_page.php. A tutorial is also provided as Supplemental Dataset 2.

RESULTS

Manually Curated Maize mRNA Expression Profiling from Publicly Available Datasets

Recently, the usage of RNA-Seq in maize (*Zea mays*). has increased dramatically—from generating no data entries in NCBI-SRA in 2008 to greater than 900 in 2016 (Fig. 1B). In contrast, the most widely used Affymetrix expression array for maize had 177 samples in 2008, but only 46 in 2016 (Fig. 1B). GCN construction approaches have not been optimized for RNA-Seq datasets in plants, and doing so could improve the quality and robustness of GCNs. To support a comprehensive evaluation on the effect of RNA-Seq normalization methods and network inference methods on the performance of GCNs, maize RNA-Seq datasets were compiled and processed with a computational pipeline (Supplemental Fig. S1). One-thousand, two-hundred and sixty-six high-quality RNA-Seq maize libraries from 17 different experiments were selected as input to an expression matrix. The corresponding experimental descriptions and publications, where available, of each library were manually checked for sample information (Supplemental Table S1). Also, a filter for reads depth and alignment rate were used to remove unqualified libraries (see "Materials and Methods" for detail). Tissue type and haplotype from those libraries were manually curated and found to include a range of sample types (Supplemental Table S1). Shoot apical meristem, leaf, and root were the top three most abundant tissue types, but a wide range of tissues were represented by multiple libraries in the dataset (Supplemental Fig. S1). The dataset also included multiple haplotypes, although B73 represented approximately 40% of the included libraries. To reduce noise, lowly expressed genes were removed from analysis, leaving 15,116 nonredundant genes across the 1266 libraries. For comparative purposes, the Affymetrix Gene Chip maize array includes 13,339 genes before filtering (GeneChip Maize Genome Array, http://www.affymetrix.com/catalog/131468/AFFY/Maize+Genome+Array#1_1).

Three RNA-Seq Normalization Methods Show Comparable Distribution of Expression

Expression data from distinct sources and experiments can be highly variable because of hybridization artifacts in microarray or variable sequencing depth in RNA-Seq. Many methods have been successfully used for normalizing both microarray and RNA-Seq data to correct for potential biases (Lim et al., 2007; Dillies et al., 2013; Li et al., 2015b). To find an optimal normalization method for building a maize GCN from RNA-Seq data, three widely used normalization methods were compared. This included VST, CPM, and RPKM (Mortazavi et al., 2008; Anders and Huber, 2010; Rau et al., 2013). For all normalization methods, log2 transformation on the normalized expression values reduced the skew of the data distribution (Supplemental Fig. S2). Several network studies from plant RNA-Seq data used log2 transformation (Davidson et al., 2011; Ma and Wang, 2012; Giorgi et al., 2013; Stelpflug et al., 2016; Walley et al., 2016). In our analysis, genes with CPM greater than 2 in more than 1000 samples were included. This filter dramatically reduces zero count values in raw data from 30.949% to 0.367%. Moreover, a prior count of 1 was added at log2 normalization [expression = log2(CPM/RPKM +1)] to avoid a problem with remaining zero values. The log2 transformation reduced skewed distributions and extreme values represented by outliers (Supplemental Fig. S2.). Thus, we think it is important to apply log2 transformation for our data.

The distribution of gene expression across the 1266 libraries formed a bell-shaped curve with a small additional peak of low expression for all three methods (Supplemental Fig. S2). To determine if these low expression values came from a few or multiple libraries, elements within the range of expression that corresponded to the observed peak ($<$ -3.7 CPM; Supplemental Fig. S2B) were extracted from a CPM-normalized expression matrix and matched to the originating libraries. This demonstrated that the low expression elements were not limited exclusively to specific libraries, but eight libraries contributed greater than 25% of lowly expressed elements. A gene ontology (GO) enrichment analysis failed to identify significant GO descriptors within the subset of 43 genes that were defined as being lowly expressed (data not shown). All eight of these libraries were from pollen tissue where the average gene expression at 147 CPM is lower than the average gene expression of the other 79 tissues combined at 183 CPM. Hierarchical clustering and correlation heatmap with the same data (Stelpflug et al., 2016) shows the uniqueness of pollen tissue expression

pattern (Langfelder and Horvath, 2008; Supplemental Fig. S3). When the lowly expressed elements from RPKM- and VST-normalized data were analyzed to determine library origin and GO enrichment (data not shown), we found similarly large numbers of pollen-specific libraries without significant GO categories. In pollen, some highly expressed genes are considered orphan genes (Wu et al., 2014), because they lack detectable homologs in another species. To investigate whether these lowly expressed genes were orphan genes, their gene sequences were blasted against foxtail millet (*Setaria italica*) genome (JGIv2; BLASTX, *e*-value < 1E-03). *S. italica* is a close relative to maize that diverged 23.4 million years ago, as estimated by TimeTree (Kumar et al., 2017). Only 1 out of 43 genes lacked detectable homologs in *S. italica* (data not shown), indicating that the majority of these genes are not likely to be orphan genes.

Because RPKM normalization accounts for gene length, the distribution of gene length versus expression for the RPKM method was compared to data normalized by VST and CPM methods. VST- and CPM-normalized data showed very similar overall patterns with no clear linear relationship between gene length and average expression (Supplemental Fig. S2C). RPKM-normalized data displayed an apparent bias toward elevated expression of a small number of genes that were less than 5000 bp in length and had a lower expression of long genes, suggesting that this normalization method might skew the distribution of expression at some genes. Overall, despite these differences, the three normalization methods resulted in a similar distribution of expression patterns for most of the genes included in the analysis. Additional analysis was completed to determine if the three normalization methods influence network performance.

Network Performance Does Not Differ Based Upon Normalization Method

To compare the efficacy of three normalization and 10 inference methods, a GCN was generated for each combination of normalization and inference methods. Furthermore, all networks were rank-standardized to limit the edge weight ranging from 0 to 1 (see "Materials and Methods"). All networks evaluations used the whole adjacency matrix (15116*15116 in RNA-Seq networks; 11429*11429 or 17862*17862 in protein networks) without a cutoff.

The performance of the different networks was measured by comparing the area under the receiver operator characteristic curves (AUROC). AUROC is a measurement used to evaluate the accuracy of classification models, making it suitable for evaluating GCNs (Gillis and Pavlidis, 2011; Ma and Wang, 2012; Liu et al., 2017). AUROC values range from 0 to 1, with a value closer to 1 indicating that the network is discriminating nonrandom patterns and perfect classification, random networks returning values close to 0.5, and values closer to 0 indicating a high degree of incorrect classification. Whereas an AUROC value close to 1 is optimal, values greater than 0.7 suggest good performance when analyzing large, diverse networks (Gillis and Pavlidis, 2011). To set up the AUROC baseline for the random networks, maize gene IDs were shuffled 10 (for MRNET and CLR) or 1000 times (for PCC) from the normalized expression matrix. The randomized expression matrix was inferenced using designated algorithms and further evaluated. The resulting AUROC values from randomized networks were very close to 0.5 (Supplemental Table S2).

AUROC values were calculated and compared for three different network characteristics. The first characteristic was designed to test if the network identified genes with known or predicted coexpression patterns, based upon prior results and inclusion in two existing datasets that could serve as a positive control for coexpression. The maize metabolic pathway (MaizeCyc) contains 413 pathways with more than two genes and was built based upon collection of evidence from genome annotation, phylogenetic distance, and known genes in maize, rice (*Oryza sativa*), and Arabidopsis (*Arabidopsis thaliana*; Monaco et al., 2013). The maize protein-protein interaction database (PPIM) is based upon both predicted and experimentally detected protein interactions (Zhu et al., 2016) and was the second dataset used in this analysis. Only high-confident interactions from PPIM were used, as defined by ranking the top 5% in their model (Zhu et al., 2016). For comparison with the GCN, genes within the same MaizeCyc or PPIM pathways were considered coexpressed. The MaizeCyc and PPIM datasets were combined and genes with fewer than five interactions were excluded from evaluation, creating a compiled dataset referred to herein as the Protein-Protein and Pathway dataset (PPPTY). PPPTY had 1720 genes and 104,856 interactions that were used in this evaluation. The AUROC value was calculated for each of the 1720 gene terms.

To assess the effect of normalization method on GCNs, AUROC values for all 10 inference methods were averaged for each of the three normalization methods. All three normalization methods scored similarly in comparison with the PPPTY dataset (Fig. 2B), with a mean AUROC value at approximately 0.575 for each, suggesting that the predicted networks were more selective than a random network.

The second characteristic was the presence of similar GO information for maize genes within a detected coexpression set, based upon "guilt by association" that assumes specific subgroups of coexpressed genes have some shared functions (Wolfe et al., 2005). GO annotations were downloaded from AgriGO (Du et al., 2010), which uses signature integration by InterPro to map gene IDs to GO terms rather than coexpression data. InterPro provided greater than 108,000,000 stable GO terms to the functional protein information database UniProtKB at release 2016_01 (Sangrador-Vegas et al., 2016). Thus, the GO annotations provide a reliable evaluation resource independent of coexpression data. To assess this characteristic, GO information was

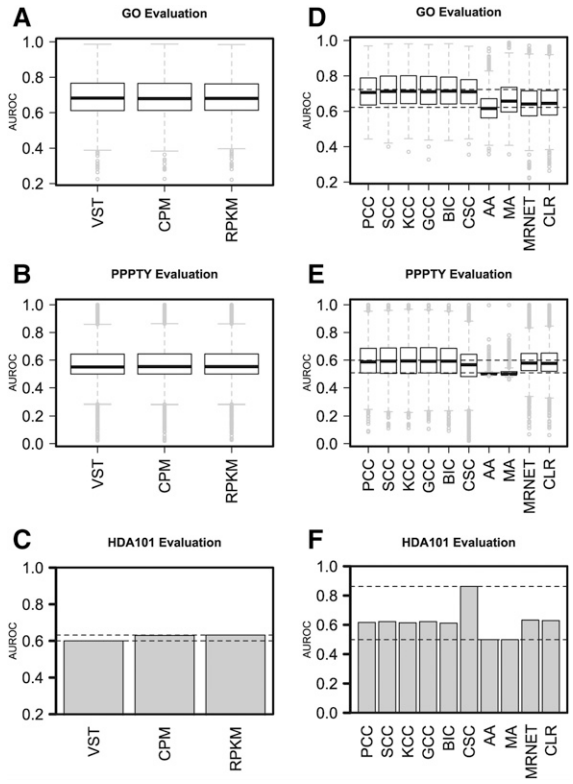

Figure 2. Normalization and network inference methods effect on single network performance. A, Network performance was evaluated by calculating area under the AUROC values from GO datasets for comparisons with samples normalized using the VST, CPM, or RPKM methods. B, Network performance was evaluated by calculating AUROC values from PPPTY dataset comparisons for samples normalized using the VST, CPM, or RPKM methods. C, Network performance was evaluated by calculating AUROC values from comparisons with HDA101 binding targets for samples normalized using the VST, CPM, or RPKM methods. D, Network performance was evaluated by calculating AUROC values from comparisons with the GO dataset for samples constructed using 10 inference methods, including PCC, SCC, KCC, GCC, BIC, CSC, AA, MA, MRNET, and CLR. E, Network performance was evaluated by calculating AUROC values from comparisons with PPPTY for samples constructed using 10 inference methods. F, Network performance was evaluated by calculating AUROC values from comparisons with HDA101 binding targets for samples constructed using 10 inference methods. Outliers were defined as outside of 1.5 times the interquartile range above the 75% quantile or below the 25% quantile. Median values were plotted as bold horizontal lines. For (C) to (F), the horizontal dashed lines indicate the highest and lowest AUROC values.

used in a neighbor voting algorithm (Gillis and Pavlidis, 2011) for sets of coexpression matrices and compared. Coexpression matrices were assessed by 3-fold cross-validation that involved masking GO terms from some genes to test whether the masked GO terms could be predicted based upon gene expression patterns. Two-hundred and seventy-seven GO terms were included for this analysis.

When GO characteristics were used to assess the networks, all three normalization methods performed similarly, but the AUROC values were higher, at approximately 0.689 for each, than those observed for comparisons with PPPTY (Fig. 2A). Because GO addresses gene functions and PPPTY emphasizes protein-protein interactions, this suggests that GCNs are better at predicting functional interactions than physical interactions. The *P* value from one-way ANOVA for testing the normalization method effect on the PPPTY and GO dataset were 0.9535 and 0.4714, respectively, confirming that the normalization method did not create a significant difference in the AUROC scores associated with the GCNs for the characteristics that were tested.

Finally, proteins that regulate gene expression or modify chromatin structure might interact with the DNA of a subset of coexpressed genes. The interactions between such a protein and regulated DNA could be detected by chromatin precipitation of associated DNA followed by DNA sequencing (ChIP-Seq). In maize, there are five ChIP-Seq datasets available (Bolduc et al., 2012; Morohashi et al., 2012; Li et al., 2015a; Pautler et al., 2015; Yang et al., 2016), some of which involved lowly expressed or tissue-specific genes. For example, *Opaque2* is specifically expressed in endosperm (Li et al., 2015a), *Knotted1* is expressed in shoot apical meristem and floral tissues (Bolduc et al., 2012), and *Pericarp Color1* has low expression except in inflorescence and seed (Morohashi et al., 2012). Histone Deacetylase 101 (HDA101) ChIP-Seq data provided the largest dataset for comparison with 26 confirmed binding targets that are relatively highly expressed in most maize tissues (Yang et al., 2016). Histone deacetylation often correlates with decreases in gene expression (Verdin and Ott, 2015). High-confidence HDA101 targets were defined as those discovered by ChIP-Seq and that also showed increased gene expression in *hda101* mutant. Networks associated with the 26 high-confidence HDA101 targets were compared by calculating AUROC. Based upon this analysis, the AUROC values were very similar among networks normalized by VST, CPM, and RPKM (Fig. 2C), which is consistent with GO and PPPTY evaluation.

Correlation Methods Perform Better than MI at Some Genes

After normalization of the expression matrices, they can be processed by different methods for GCN inference. To optimize this step, the AUROC values of six correlation (PCC, SCC, KCC, GCC, BIC, and CSC) and four MI methods [additive ARACNE (AA), multiplicative ARACNE (MA), MRNET, and CLR) were compared for the expression matrices that were generated from each of three normalization methods (VST, CPM, and RPKM) and then averaged. In general, correlation methods are more computationally efficient whereas MI methods are able to reveal nonlinear relationships (Li et al., 2015c). PCC is widely used but may be influenced by outliers (Mukaka, 2012). SCC, KCC, and BIC are less sensitive to outliers, because SCC and KCC only consider the rank information and BIC is

calculated based on dataset median instead of mean (Serin et al., 2016). Recently, GCC has been shown to be a better correlation method for gene expression analysis because of its capacity to detect nonlinear relationships and its insensitivity to outliers (Ma and Wang, 2012). CSC is widely used for text mining and analyzing sparse data with many zeros (Dhillon and Modha, 2001). ARACNE, MRNET, and CLR showed extended gene-dependent relationships under variable biological settings (Margolin et al., 2006; Faith et al., 2007; Meyer et al., 2007; Li et al., 2013b). To estimate the effectiveness of the inference methods, the same testing parameters with AUROC calculations were performed as described for the testing of normalization methods.

Assessed by GO datasets, the 277 AUROC values were averaged to create one average value for each of the 10 inference methods ranging from 0.620 to 0.724 (Fig. 2D). The average AUROC across all normalization methods for six correlation methods was 0.718, whereas the average AUROC for all four MI methods was 0.646. The majority of the 277 GO terms had similar AUROC values in the different correlation-method-generated GCNs, and these patterns are different from those observed in the MI-generated GCNs (Fig. 3A). The similarity among different methods was also detectable by pairwise comparison and comparing Pearson correlations between the different methods (Supplemental Fig. S4A).

To evaluate network inference methods with the PPPTY dataset, the AUROC values for 1720 genes were averaged for each combination of normalization and inference methods (Fig. 2E). This evaluation also showed that the networks constructed using correlation methods resulted in higher AUROC values than MI methods, although the CSC method resulted in lower AUROC values than other correlation methods. As demonstrated for the GO evaluation, results from correlation methods were more similar with each other than the MI methods (Supplemental Fig. S4B). Interestingly, heatmap results indicated that a subset of genes consistently had higher AUROC values when CSC, MRNET/CLR, or AA/MA were used (Fig. 3B), although this includes a small enough number of genes that the average AUROC value over the whole gene set was relatively low for those methods. The gene sets with highest AUROC values in PCC, CSC, or MRNET were extracted. Characteristics of each gene sets were compared in average expression (CPM) and average number of low expressed elements (CPM < 0). The CSC gene set had the smallest number of low expression elements and had higher average expression than both the 1720 gene set and the PCC gene set (Supplemental Fig. S5). This may indicate that the CSC method is better at determining coexpression for highly expressed genes.

The AUROC values from 26 targets of HDA101 ChIP-Seq datasets reveals that CSC GCN had the highest AUROC value and the use of MRNET/CLR GCNs resulted in slightly higher scores than correlation methods (Fig. 2F). This could be explained by the small number of targets creating skewed results, but may also indicate that CSC/MI methods are more suitable for specific types of genes or interactions between genes (Tzfadia et al., 2016). HDA101 is a highly expressed gene in all samples with average expression value equal to 8.64 CPM and minimum expression equal to 2.89 CPM, so it is possible that HDA101 is more suitable for the CSC method. Using two models of ARACNE (AA and MA), the coexpression matrices contain less than 0.5% nonzero values for all comparisons and so these techniques were not included in any additional analyses.

In conclusion, our results indicated the widely used correlation methods resulted in a more predictive maize GCN from a single expression matrix, but coexpression with some individual genes may be better detected using MI methods. Normalization method did not have a substantial influence on GCN's performance, so only CPM normalization was used in conjunction with PCC, SCC, MRNT, and CLR inference for subsequent optimization of other parameters.

Increase Sample Size Had a Positive Effect On GCN

GCN analysis can be accomplished with a variable number of samples and datasets, but sample size can influence the quality of the resulting GCN (Wei et al., 2004; Ballouz et al., 2015). Separate analyses were conducted with different numbers of samples and experiments to empirically determine the effect of sample number on GCN effectiveness. The data in our analysis consisted of 17 experiments, each including between 12 and 404 libraries. For this analysis, the CPM normalization method, followed by each of four inference methods (PCC, SCC, MRNET, and CLR) was applied to the 17 experiments and the 68 resulting networks were evaluated by both GO and PPPTY.

From GO and PPPTY evaluation, all algorithms exhibit a positive linear relationship between sample size with natural logarithm transformed and average AUROC values (Fig. 4). The linear relationships are stronger in the PCC and SCC methods with higher r^2 values, indicating correlation methods benefit more from increasing sample size. Thus, for building correlation-based GCNs, as many samples as possible should be included. We also found that, as seen for the total GCN analysis, PCC and SCC had higher average AUROC values than the MRNET and CLR methods for PPPTY and GO analysis for most of the individual networks (Fig. 5).

Ranked Aggregation of Networks Improved Performance of GCNs

Ranked aggregation for metaanalysis can also be modified to change the outcomes of GCN by buffering the effect of sample heterogeneity (Zhong et al., 2014; Wang et al., 2015a; Asnicar et al., 2016). Aggregated rank standardized correlation/MI matrices were calculated from separate experiments to determine if this approach enhanced GCN performance. Aggregating individual networks together for metaanalysis can help

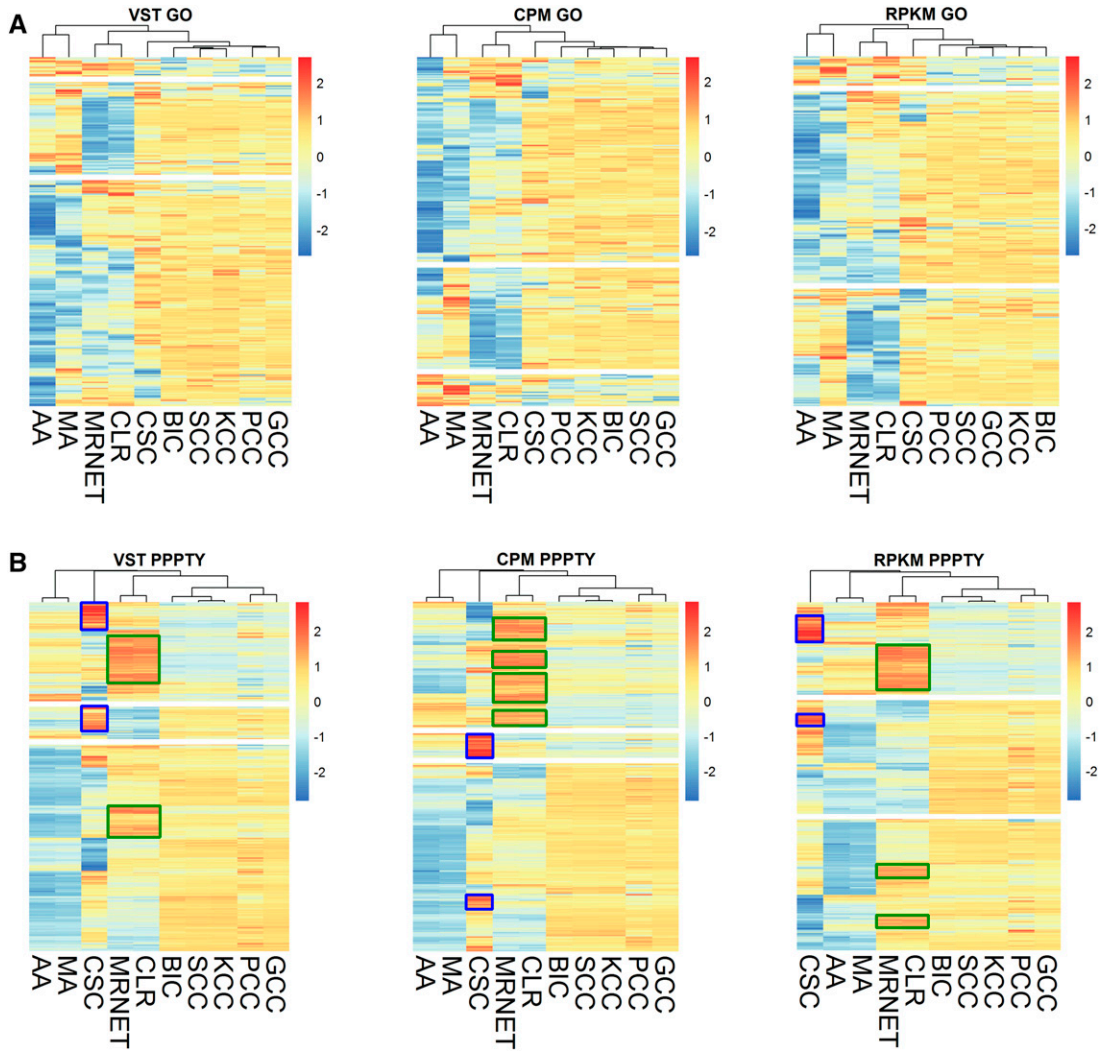

Figure 3. Similarity among 10 inference methods of network performance based upon GO (A) and PPPTY (B) evaluation. Genes with the highest AUROC value in CSC or MRNET/CLR are enclosed in a blue or green box, respectively. AUROC values for each GO term or genes were scaled to standard normal distribution, resulting in scaled AUROC values between −3 (blue) and 3 (red). Samples normalized by VST, CPM, and RPKM were analyzed using 10 inference methods (PCC, SCC, KCC, GCC, BIC, CSC, AA, MA, MRNET, and CLR) and clustered based on Euclidian distance.

to highlight true coexpression interactions and reduce noise (Zhong et al., 2014; Wang et al., 2015a, 2015b). This analysis was conducted with the 17 differently sized experiments using the PCC, SCC, MRNET, and CLR methods for GCN inference as we did previously, resulting in 68 single GCNs. The 17 experiments were aggregated for PCC, SCC, MRNET, and CLR individually and evaluated by GO and PPPTY datasets.

Of the four aggregated networks that were evaluated, the two correlation methods (PCC and SCC) had higher AUROC values than the single network from 1266 samples (Fig. 6 and Supplemental Fig. S6). However, this aggregation strategy did not result in significantly higher AUROC scores for the MRNET and CLR method networks compared with single networks with 1266 samples (two-tail Wilcoxon rank test for GO evaluation, P values = 0.494 and 0.796). It has been reported that MI estimation accuracy is dependent on sample size (Gao et al., 2015), therefore individual MI networks built with a small number of libraries may not demonstrate improved accuracy from aggregation. In conclusion, the PCC/SCC-built GCN performed best using a ranked aggregation strategy and use of this strategy, in combination with the other optimized parameters, creates a robust GCN.

The Performance of Protein Networks Did Not Exceed Aggregation Networks

In many cases, mRNA levels in a cell are of interest because mRNA level is thought to be related to the level

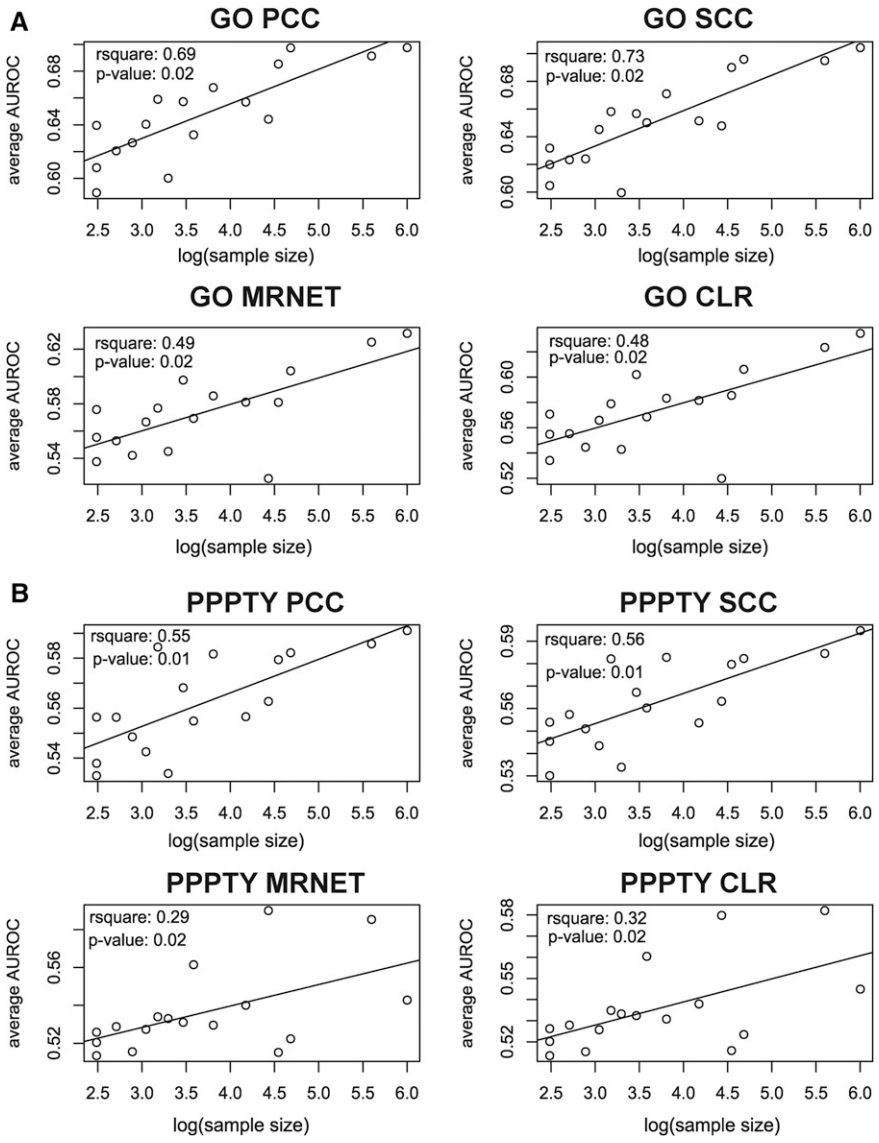

Figure 4. Effect of sample size on network performance. A, Average AUROC values from GO evaluation of 17 different-sized networks plotted against natural logarithm transformed sample size [log(sample size)]. B, Average AUROC values from PPPTY evaluation of 17 different-sized networks plotted against natural logarithm transformed sample size [log(sample size)]. Regression fitting logarithm models were plotted in black lines. R^2 and P values were calculated by the lm() function in R.

and function of a protein of interest. However, many researchers had found inconsistencies between mRNA and protein level (Baerenfaller et al., 2008; Schwanhäusser et al., 2011; Ponnala et al., 2014; Walley et al., 2016). Although relatively fewer protein expression data are available, these data are amenable to GCN construction and could represent a more direct reflection of interacting proteins. Using a nonmodified protein expression atlas from 23 maize tissues based upon mass spectrometry data (Walley et al., 2016), four protein networks were built with PCC, SCC, MRNET, and CLR separately and then evaluated using the same PPPTY and GO dataset as previously mentioned.

GCNs constructed from protein expression did not exhibit superior AUROC values to those observed for RNA-Seq-based GCN using the aggregation strategy (Fig. 6). When evaluated by GO and PPPTY dataset, the performance of the protein network was lower than the aggregated network and the single network from 1266 samples. To confirm this result, a two-way ANOVA was computed with pairwise comparison for the GO evaluation, which showed that the effect of network type was significant (Supplemental Table S3). A subsequent pairwise comparison using Wilcoxon rank sum test indicated that the PCC/SCC method was significantly better than MRNET/CLR (Supplemental Table S3), although MI methods may be superior for some types of interactions.

The raw protein expression data included 17,862 genes, of which 11,429 genes overlapped with our RNA-Seq-based network and were therefore used for the analysis. To demonstrate that the performance of the protein network was not biased due to the gene selection, the PCC method was used for the whole group of 17,862 genes to construct a protein network (Supplemental Fig. S7). No improvement could be detected from the protein network derived from 17,862 genes with P value = 0.635 for GO evaluation and

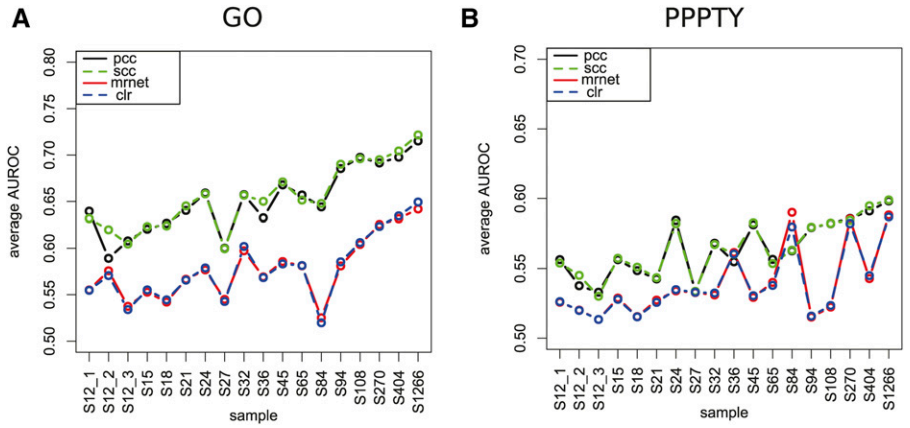

Figure 5. Evaluation of different-sized networks constructed with different numbers of samples using PCC (black), SCC (green), MRNET (red), and CLR (blue) methods. A, Average AUROC values from GO evaluations of networks constructed using 12 (S12) to 1266 (S1266) libraries were plotted against sample size. Seventeen individual networks were labeled as S12_1 to S404; the S1266 included all samples from 17 experiments. B, Average AUROC values from PPPTY evaluations of networks constructed using 12 (S12) to 1266 (S1266) libraries were plotted against sample size. Networks with the same number of samples included are designated as "1", "2", and "3".

P value = 0.995 for PPPTY evaluation from a one-sided Wilcoxon rank sum test.

PCC and SCC-built GCN Exhibits Identical Topological and Functional Properties

In addition to evaluation of network performance based upon biological characteristics, networks can be compared based upon several different network characteristics, including clustering coefficient, number of nodes, network heterogeneity (Dong and Horvath, 2007), network centralization (Dong and Horvath, 2007), number of detected modules, and number of genes in the largest module. Number of nodes is a basic construct in graph theory depicting the scale of a network. Clustering coefficients and number of modules are to model how densely nodes are connected in networks. Heterogeneity measures the variability of node connections. Centralization indicates how likely some nodes are to have significantly more connections than average. In this analysis, each gene corresponds with a node. Based on the extensive evaluation using biological characteristics, like protein-protein interactions (PPPTY) and predicted gene function (GO), three final maize networks were selected for comparison of basic network characteristics based on their overall performance: PCC and SCC-built ranked aggregation network from 17 experiments (PA

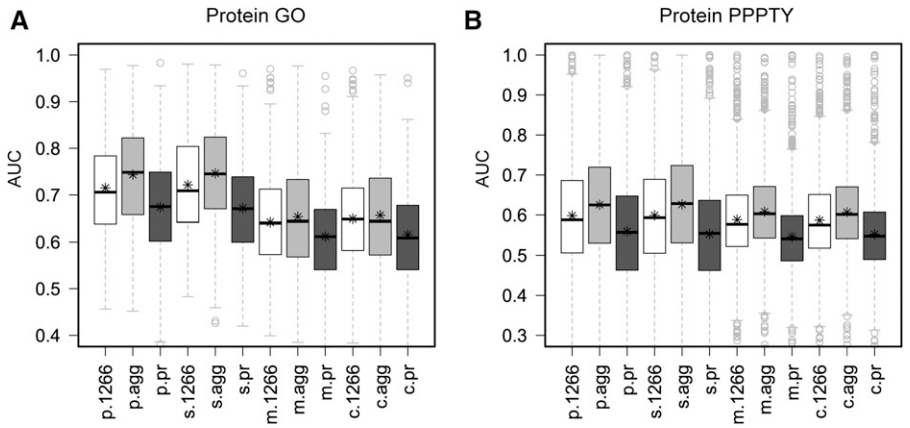

Figure 6. GCN performance comparison of networks constructed with 1266 libraries. A, AUROC values from GO evaluation of single network (white bars), aggregation network (gray bars), and protein network (dark gray bars) were compared for network constructed using PCC(p), SCC(s), MRNET(m), or CLR(c). Bold horizontal lines indicate median. Asterisks indicate mean, and gray dots indicate outliers. B, AUROC values from PPPTY evaluation of single network (white bars), aggregation network (gray bars), and protein network (dark gray bars) were compared for network constructed using PCC(p), SCC(s), MRNET(m), or CLR(c) Bold horizontal lines indicate median. Asterisks indicate mean, and gray dots indicate outliers.

and SA), and MRNET-built single network from 1266 total samples (MS). The three networks were constrained to include the top one-million predicted interactions, or edges.

In prior studies, most biological networks had scale-free architectures that fit a power-law distribution (Barabási et al., 2004; Doncheva et al., 2012; Schaefer et al., 2014). For the three final maize networks constructed using optimized parameters, both neighborhood connectivity distribution (Supplemental Fig. S8) and node degree distribution (Supplemental Fig. S9) fit power-law models with r^2 values greater than 0.7. The MS network had the highest network centralization value. The network heterogeneity value of MS was over two times that of PA and SA, indicating that MS may contain more highly interacting genes (Supplemental Table S4), consistent with the observed highest centralization values for this network. Centralization and heterogeneity are two variants to model the degree distribution of networks. A scale-free network with more numbers of hubs has larger values of centralization and heterogeneity, whereas a network with larger values of centralization and heterogeneity may contain a larger number of hubs or the number of hubs is not significantly large, but the degree distributions are extremely imbalanced. In biological networks, many observations connected large values of centralization and heterogeneity with more hub genes (Ma and Zeng, 2003; Horvath and Dong, 2008; Iancu et al., 2012; Scott-Boyer et al., 2013), even though theoretically we cannot rule out the possibility that high values were a result from extremely imbalanced degree distribution. For the MS network, most highly connected genes interacted with a large number of lowly connected genes; this pattern is also apparently reflected in the decreasing neighborhood connectivity distribution for the MS network (Supplemental Fig. S8). The genes with the most interactions are expected to act as key components in GCN networks (Langfelder and Horvath, 2008; Allen et al., 2012), and likely represent central regulators of multiprotein biological processes (Ma et al., 2013; Du et al., 2015). The top 1000 interacting genes from all networks were analyzed in more detail, as these were potential hub genes that may regulate other expression patterns and processes. PA and SA shared 95% of the top 1000 interacting genes, whereas MS had 835 unique genes (Fig. 7A). One-hundred and forty-eight genes were shared among all three networks (Supplemental Table S5), making these genes strong candidates for central biological regulators. The annotation of these genes suggests their participation in a range of basic cellular processes (Fig. 7C), including gene expression, DNA replication, translation, and gene silencing (Supplemental Table S5); the top interacting genes were not limited to a subset of cellular biochemistry. Ribosomal proteins were the largest component of top interacting genes (27/148), which was expected because of their cellular abundance and involvement with translation. Interestingly, nine epigenetic regulators were found in the 148 shared genes, including AGO104 (GRMZM2G141818; Singh et al., 2011), CHR106 (GRMZM2G071025; Li et al., 2014a), and LBL1

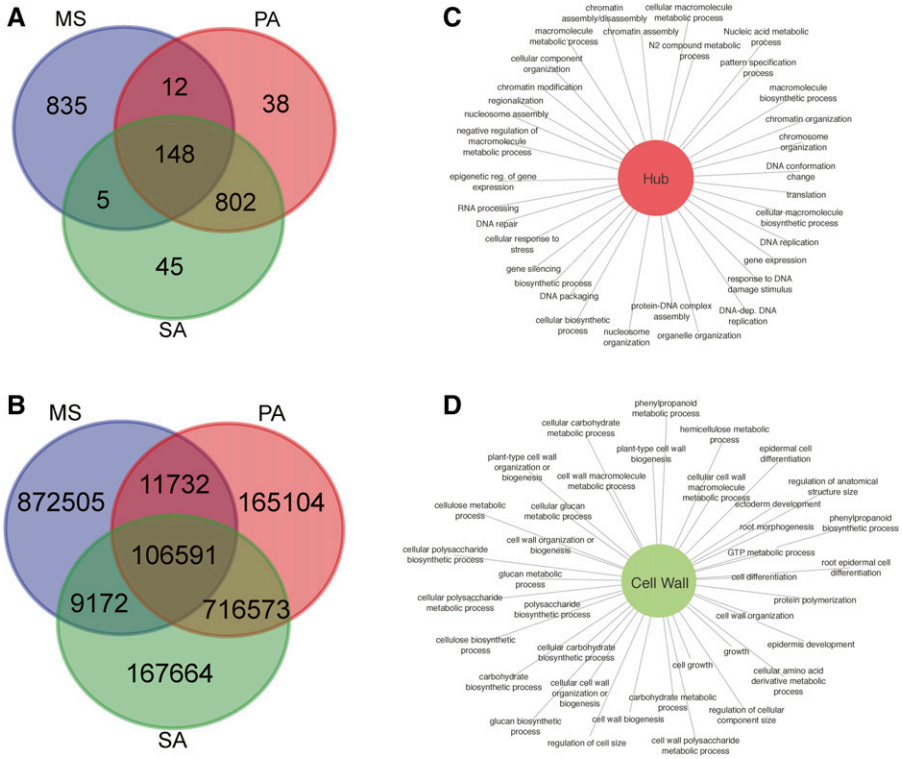

Figure 7. Comparison of top 1000 interacting genes in aggregation and single networks. A, Venn diagram shows the overlap among 1000 highest interacting genes in three networks. One-hundred and forty-eight genes were shared among three networks. B, Venn diagram shows the overlap among top 1×10^6 edges in three networks. A quantity of 10,6591 edges was shared among three networks. C, GO term enrichment analyzed by AgriGO for 148 shared highly interacting genes among PCC ranked aggregation network, SCC ranked aggregation network, and MRNET single network. D, GO term enrichment analyzed by AgriGO for 214 coexpressed genes, queried by 16 cell wall pathway genes. MS, MRNET single network; PA, PCC ranked aggregation network; SA, SCC ranked aggregation network.

(GRMZM2G020187; Dotto et al., 2014), demonstrating the importance of epigenetic regulation for plant development (reviewed by Huang et al., 2017).

To reveal the underlying properties of GCNs, a graph clustering algorithm in the form of a Markov cluster algorithm (MCL) was used to identify network modules (Enright et al., 2002; Morris et al., 2011). The result showed a shared pattern between the PA and SA networks that was distinct from the MS network (Supplemental Table S4). The MS network had fewer but larger modules detected than the PA and SA networks. Consequently, most genes in the MS network clustered into one very large module of 14,054, consistent with the high network centralization value for the MS network. Conversely, PA and SA networks separated into smaller, distinct modules with related GO enrichment (Supplemental Tables S6 and S7). The pattern displayed by the PA and SA networks (Supplemental Fig. S10) seems more likely to represent biologically relevant pathways, and so these methods appear to be better for module detection.

To compile a high-confidence coexpression network, the top-1 million edges from PA, SA, and MS were merged and the intersection of the three produced a 14,277-gene, 106,591-interaction, merged network. PA and SA shared 83.5% of common interactions within the networks whereas MS had 87.3% unique interactions (Fig. 7B). This merged network (Supplemental Dataset S1) was used for a case study analysis of cell wall biosynthesis. The same network can also be accessed at http://www.bio.fsu.edu/mcginnislab/mcn/main_page.php.

Case Study: Cell Wall Biosynthesis and Regulation

To demonstrate the functionality of the network, the predicted cell wall biosynthesis pathway from the merged network was compared to the existing knowledge of this pathway. Sixteen well-characterized components of cell wall biosynthesis were selected as guide genes (Supplemental Table S8), including five cellulose synthase genes, seven cellulose synthase-like genes, three glycosyl hydrolase genes, and one glycosidase gene (Penning et al., 2009; Bosch et al., 2011). Collectively, 214 genes containing 377 edges were extracted from the network with the 16 guide genes (Fig. 8A); two guide genes did not have any coexpressed genes in the network that met the analysis criteria. As expected for these 214 genes, cell-wall-related GO terms were enriched (Fig. 7D; Supplemental Table S9).

The resulting 214 coexpressed genes were queried against the Arabidopsis TAIR 10 protein database to retrieve homologs and their annotations using BLASTP. The literature was manually searched using the maize genes and their Arabidopsis homologs as queries (Supplemental Table S10). The results of the literature survey showed that 31.3% (67/214) of the genes coexpressed with the guide genes had peer-reviewed publications indicating a role in cell wall synthesis or related pathways in plants. A search using 214 randomly selected genes as queries returned only 3.27% genes (7/214) that were involved in cell-wall-related pathways. This suggests that the network discriminated coexpressed genes and identified some known components of the pathway. Lignin biosynthesis genes are expected to function in cell wall biosynthesis to provide rigidity and strength in the secondary cell wall (reviewed by Vanholme et al., 2010). Interestingly, even though no lignin biosynthesis genes were included in our queries, six lignin biosynthesis genes (PAL1, C4H, 4CL2, HCT, CCoAOMT1, and PDR1; reviewed by Zhong and Ye, 2015) were found to be coexpressed with the guide genes. At least nine cellulose biosynthesis and assembly genes were discovered, including CESA1, FLA11, IRX9, IRX14, and IRX10 (reviewed by Zhong and Ye, 2015). Moreover, proteins participating in a well-studied physical interaction, Cellulose Synthase Interactive1

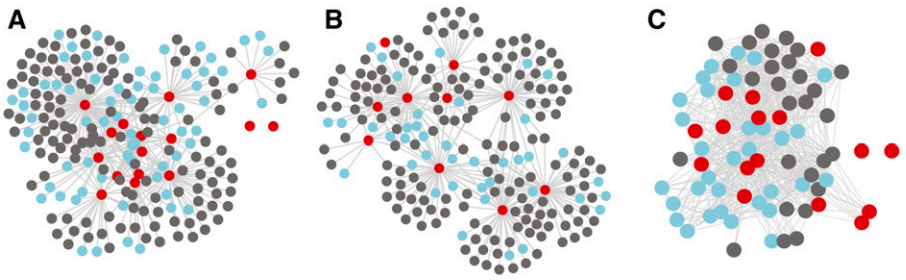

Figure 8. Cell wall pathway subnetworks. A, Intersections of PCC aggregation, SCC aggregation, and MRNET-single networks, queried by 16 cell wall pathway genes (red nodes). Cyan nodes are genes with reported function in cell-wall-related pathways in plant. Dark gray nodes are genes without prior knowledge of involvement in cell-wall-related pathways. Gray lines indicate network-predicted interactions. B, Network retrieved from CORNET database, queried by the 16 cell wall pathway genes (red node). Cyan nodes are genes with reported function in cell-wall-related pathways in plant. Dark gray nodes are genes without prior knowledge of involvement in cell-wall-related pathways. Gray lines indicate network-predicted interactions. C, Network retrieved from STRING database, queried by 16 cell wall pathway genes (red nodes). Cyan nodes are genes with reported function in cell-wall-related pathways in plant. Dark gray nodes are genes without prior knowledge of involvement in cell-wall-related pathways. Gray lines indicate network-predicted interactions.

(CSI1), Cellulose Synthase6 (CESA6), and Cellulose Synthase 3 (CESA3; Desprez et al., 2007; Gu et al., 2010), were also predicted to be expressed in the network. There were 131 genes without reported functions in cell wall pathways, an indication that GCN analysis can be used to predict undiscovered components of biological pathways in maize.

The cell wall biosynthesis pathway results were also compared with the CORNET coexpression database (De Bodt et al., 2012) and STRING functional protein association network (Szklarczyk et al., 2015) using the same 16 genes and similar parameters (see "Materials and Methods"). From CORNET, 10 out of 16 genes had coexpressed genes (Fig. 8B). In total, 210 genes and 325 interactions were retrieved using CORNET, of which 19% (40/210) had publications supporting their function in cell wall pathways (Supplemental Table S11). STRING performed very well, with 14 out of 16 genes demonstrating predicted protein association (Fig. 8C), resulting in 817 interactions with 76 genes. Forty-eight percent (36/75) of coexpressed genes were experimentally confirmed (Supplemental Table S12), the highest percentage among the three methods. Only one of the lignin biosynthesis genes (PAL1) was found using CORNET and none were found using STRING. Although STRING appears very robust for predicting protein-protein interactions, this suggests that an optimized GCN analysis have more power to find genes that function together without physically interacting. This case study shows that a robust, optimized GCN can discover physical and functional interactions and enhance study of biological relevant interactions. A tutorial is provided on how to use Cytoscape to visualize any coexpressed genes in our network (Supplemental Dataset S2).

DISCUSSION

As the per-read cost of RNA-Seq technology decreases, the use of this technology is quickly increasing. With greater than 5000 libraries available for maize, there is now ample data to support GCN analysis. This comprehensive evaluation of normalization methods and network inference methods using real maize RNA-Seq data will provide a useful set of optimized parameters to support these analyses.

In our analysis, VST, CPM, and RPKM normalization methods had equivalent outcomes for GCN analysis, consistent with prior results using much smaller datasets (Giorgi et al., 2013). Several benchmark studies focusing on differential expression analysis proposed that RPKM performed poorly and should be avoided (Maza et al., 2013; Dillies et al., 2013; Zyprych-Walczak et al., 2015). This was not observed for the maize GCN testing. It is possible that the large number of samples from various labs, created enough heterogeneity within samples that normalization effects were minimized (Paulson et al., 2016). Furthermore, the normalization is on a library basis, which means genes within the same library are normalized by similar factors. So, when the network is constructed by PCC and BIC where expression vectors are centered by mean or median values, the effect of different normalization methods are probably small. Two rank correlations, SCC and KCC, only consider the difference on relative rankings where normalization has a limited effect. It is similar for the GCC method. The estimation of mutual information is based on the k-nearest neighbor method implemented in Parmigene (Sales and Romualdi, 2011). Because the three normalization methods shared similar expression distribution (Supplemental Fig. S2), MI estimations from different normalizations are expected to be similar.

When assessing inference methods, the simple and widely used correlation methods, like PCC and SCC, are less time-consuming than MI methods. This analysis showed PCC/SCC-built GCNs had better overall performance. This is consistent with a study in human GCN analysis (Ballouz et al., 2015) but SCC did not score higher than other correlation methods using GO and PPPTY evaluations. Some genes had higher performance using MI methods, but this effect was limited to evaluation with the PPPTY data. This may indicate that correlation and MI inference methods assert different kinds of interactions (Meyer et al., 2008; Marbach et al., 2012; Song et al., 2012). Marbach et al. (2012) stated that integration of multiple inference methods showed a more robust performance than any single inference methods in in silico and in *Escherichia coli* expression networks, referring to "the wisdom of the crowd". However, for analysis of the available maize data, integration of PCC, SCC, MRNET, and CLR did not result in a network that outperformed PCC and SCC networks (data not shown). This approach was also less effective in more complex *Saccharomyces cerevisiae* datasets than prokaryotic networks (Marbach et al., 2012), suggesting that more work is required to determine whether integrating algorithms can improve GCNs with eukaryotic data.

In conclusion, we extensively evaluated normalization methods and inference methods for building an RNA-Seq-based maize GCN. This optimization may apply to a range of datasets with shared characteristics of maize, including a large and heterogeneous genome, with rich and diverse transposon element composition and limited gene annotation.

MATERIALS AND METHODS

RNA-Seq Data Collection and Process

The maize (*Zea mays*) genome and its annotation were downloaded from Ensembl Plant Release 31 (http://plants.ensembl.org/). The original 1303 RNA-Seq samples based on Illumina HiSeq2000 or Hiseq2500 were downloaded from NCBI SRA (Leinonen et al., 2010). The downloaded files were converted to Fastq format using the Fastq-dump command in SRA Toolkit (version 2.5.2). The adapters for the Fastq files were trimmed by Cutadapt 1.8.1 (Martin, 2011). The adapter-removed files were then quality checked by FastQC v0.11.2 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). HISAT2 v2.0.4 (Kim et al., 2015) was used for genome alignment. Gene-level expression raw read counts were calculated by FeatureCounts 1.5.0 (Liao et al., 2014) from aligned bam files (Supplemental Fig. S1). Twenty-six libraries with fewer than 5,000,000 reads

total and 11 libraries with less than 70% of total alignment rate were excluded, leaving 1266 samples (Supplemental Table S1) for the final expression table. The processing protocol were streamlined by Snakemake v3.7.1 (Köster and Rahmann, 2012).

Gene Count Normalization

The expression data were normalized using three different methods before constructing GCNs. CPM and RPKM were calculated by the edgeR package (Robinson et al., 2010) in the "R" environment and then log2 normalized [expression = log2(CPM/RPKM +1)]. For both methods, scale factors between samples were estimated by trimmed mean of *m*-values in edgeR. VST was calculated by the DESeq2 package (Love et al., 2014). Only genes with expression higher than 2 CPM in greater than 1000 samples were included for additional analysis (15,116 genes).

Network Inference

Six correlation coefficient methods and four mutual information methods were applied to normalized gene expression data to construct GCNs. All computing steps were done in the R 3.3.1 environment. The PCC and SCC was calculated by the cor() function. The KCC was calculated using the cor.fk() function in the pcaPP package (Filzmoser et al., 2009). The GCC was calculated by the adjacencymatrix() function in the RSGCC package (Ma and Wang, 2012). Biweight midcorrelation was computed by bicor() function in the WGCNA package (Langfelder and Horvath, 2008). Cosine similarity coefficient was computed by cosine() function in the COOP package (Schmidt, 2016). Mutual information results were computed using the Parmigene package (Sales and Romualdi, 2011). The adjacency matrix weights derived from 10 inference methods were ranked with smallest value = 1. Then ranks were divided by the number of elements in the matrix and the diagonal was set to 1 to make all networks weights range from 0 to 1.

Network Performance Evaluation

To generate the random networks, gene IDs were shuffled randomly in CPM- or VST-normalized expression matrices. The randomized expression matrices were then inferenced by PCC, MRNET, or CLR methods and evaluated. For PCC methods, 1000 repeats of randomization and evaluation were conducted. For MRNET and CLR, each inference step took 2 h on our server, so 10 repeats were conducted.

Four maize datasets were used for evaluation. First, maize protein-protein interactions were downloaded from PPIM v1.1 (Zhu et al., 2016). Only high-confidence interactions were used for evaluation, as defined by the ranking top 5% in their results. Second, maize pathway information was downloaded from MaizeCyc v2.2 (Monaco et al., 2013). Genes within the same pathways were considered as coexpressed. Third, maize GO data for AGPv3.30 was downloaded from AgriGO (Du et al., 2010). GO terms with 20 to 300 genes were used for evaluation. Fourth, ChIP-Seq confirmed the targets for HDA101 (GRMZM2G172883; Yang et al., 2016) were used as positive coexpressed examples for evaluation.

The widely used AUROC for binary classification problems was used for evaluations. Protein-protein interaction and pathway information was parsed into lists of coexpressed genes. Prediction() and performance() function in the R package ROCR were used to calculate AUROCs (Sing et al., 2005). The 277 AUROC values for GO datasets were calculated by the EGAD package (Ballouz et al., 2016) in R. Basically, it utilizes the "guilt by association" principle that genes with shared GO terms are more likely to connected. Thus, networks normalized and inferred by different methods can be evaluated by hiding a subset of genes GO terms and test whether the hidden GO terms could be predicted from the remaining annotations. The prediction model performance was measured by AUROC values in 3-fold cross-validation. All ANOVA and pairwise Wilcoxon rank tests were analyzed in R using anova() and pairwise.wilcox.test() functions from the Stats package. The *P*-value adjustment method was set to "fdr" (Benjamini and Hochberg, 1995).

Definition of True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). For the evaluation using the PPPTY dataset, TP: a network predicts two genes are coexpressed and they are coexpressed in PPPTY dataset; FP: a network predicts two genes are coexpressed, but they are not; TN: a network predicts two genes are not coexpressed and they are not coexpressed in PPPTY; FN: a network predicts two genes are not coexpressed, but they are coexpressed in PPPTY datasets. For the evaluation using the GO dataset, TP: a network predicts a gene that has a specific GO term and it does have that GO term in our GO dataset; FP: a network predicts a gene has a specific GO term, but it does not have that GO term in our GO dataset; TN: a network predicts a gene does not have a specific GO term and it does not have it in our GO dataset; FN: a network predicts a gene does not have a specific GO terms, but it has that GO term in its GO dataset.

Network Clustering and Characterization

For each network, the top 1,000,000 edges were selected as stringent coexpression networks. The network topological characteristics were computed in Cytoscape (Shannon et al., 2003). The neighborhood connectivity distribution and node degree distributions were plotted by Network Analyzer plugin (Doncheva et al., 2012). Graph clustering was performed using MCL v14.137 with inflation value set to 1.8 (Enright et al., 2002). All networks were visualized in Cytoscape.

GO Enrichment and Visualization

GO enrichment was analyzed in AgriGO's Singular Enrichment Analysis tool (Du et al., 2010). Fifteen-thousand, one-hundred and sixteen genes involved in our networks were used as background references. Hypergeometric testing was used to calculate *P* value, for which a value below 0.05 was considered as significant. The Yekutieli method was used for multiple test correction, and terms with a false discovery rate greater than 0.05 were discarded. The results were then imported into Cytoscape for visualization.

Databases Comparison on Cell Wall Pathway

Sixteen well-characterized (Penning et al., 2009; Bosch et al., 2011) components of cell wall biosynthesis (Supplemental Table S8) were chosen as query genes to search against CORNET Maize (https://bioinformatics.psb.ugent.be/cornet/versions/cornet_maize1.0/) on its website and the STRING database using the Cytoscape stringApp (http://apps.cytoscape.org/apps/stringapp). The parameters for searching CORNET database were: Method = Pearson, Correlation coefficient = 0.75, *P* value ≤ 0.05, and Top genes = 50. This resulted in 210 coexpressed genes and 325 interactions. To search the STRING database, the confidence cutoff was set to 0.4 with the maximum number of interactors set to 100. Seventy-six genes with 817 interactions were retrieved. Maize proteins were blasted against TAIR 10 protein sequences using standalone BLASTP version 2.2.28+ (Camacho et al., 2009).

Supplemental Data

The following supplemental materials are available.

Supplemental Figure S1. Pipeline and datasets used for analysis.

Supplemental Figure S2. Distribution of gene expression values.

Supplemental Figure S3. Maize CPM-normalized with log2 transformed gene expression from all tissues and developmental stages.

Supplemental Figure S4. Pairwise comparison among results of inference methods.

Supplemental Figure S5. Characteristics of the all-1720-PPPTY gene set (ALL_1720), genes with highest AUROC values in the CSC method (CSC), the PCC method (PCC), and the MRNET method (MRNET).

Supplemental Figure S6. Evaluation of network performance based on sample size and inference.

Supplemental Figure S7. GCN performance comparison between protein networks.

Supplemental Figure S8. Average neighborhood connectivity for three selected networks, PCC-aggregated (PA), SCC-aggregated (SA), and MRNET-single (MS).

Supplemental Figure S9. Node distribution for the four selected networks, PCC-aggregated (PA), SCC-aggregated (SA), and MRNET-single (MS), and the intersection among three networks (Merged network).

Supplemental Figure S10. Network representation of PCC ranked aggregation network (PA).

Supplemental Table S1. RNA-Seq libraries used in this analysis.

Supplemental Table S2. Random network AUROC value baseline.

Supplemental Table S3. ANOVA tables and pairwise comparisons.

Supplemental Table S4. Topological characteristics of four maize networks.

Supplemental Table S5. GO annotation for 148 hub genes.

Supplemental Table S6. Enriched GO terms for PCC ranked aggregation networks from module 1 to module 8.

Supplemental Table S7. Enriched GO terms for SCC ranked aggregation networks from module 1 to module 8.

Supplemental Table S8. Sixteen query genes in maize cell wall pathway.

Supplemental Table S9. GO enrichment analysis for 214 coexpressed genes of cell wall query genes in the merged network.

Supplemental Table S10. Annotation for coexpressed genes queried by 16 cell wall pathway genes from the merged network.

Supplemental Table S11. Annotation for coexpressed genes queried by 16 cell wall pathway genes from the CORNET database.

Supplemental Table S12. Annotation for coexpressed genes queried by 16 cell wall pathway genes from the STRING database.

Supplemental Dataset S1. The merged network in Cytoscape-ready format.

Supplemental Dataset S2. Tutorial: Visualizing Coexpression Data in Cytoscape.

ACKNOWLEDGMENTS

We give special thanks to Dr. Peixiang Zhao (FSU Department of Computer Science) for advice and discussion on topological analysis of maize networks. We also thank Dr. Alan Lemmon (FSU Department of Scientific Computing) and Dr. Jonathan Dennis (FSU Department of Biological Science) for the helpful discussion on data analysis.

Received June 19, 2017; accepted July 31, 2017; published August 2, 2017.

LITERATURE CITED

Allen JD, Xie Y, Chen M, Girard L, Xiao G (2012) Comparing statistical methods for constructing large scale gene networks. PLoS One **7:** e29348

Anders S, Huber W (2010) Differential expression analysis for sequence count data. Genome Biol **11:** R106

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

Aoki Y, Okamura Y, Tadaka S, Kinoshita K, Obayashi T (2016) ATTED-II in 2016: a plant coexpression database towards lineage-specific coexpression. Plant Cell Physiol **57:** e5

Asnicar F, Masera L, Coller E, Gallo C, Sella N, Tolio T, Morettin P, Erculiani L, Galante F, Semeniuta S (2016) NES2RA: network expansion by stratified variable subsetting and ranking aggregation. Int J High Perform Comput Appl doi: 10.1177/1094342016662508

Baerenfaller K, Grossmann J, Grobei MA, Hull R, Hirsch-Hoffmann M, Yalovsky S, Zimmermann P, Grossniklaus U, Gruissem W, Baginsky S (2008) Genome-scale proteomics reveals *Arabidopsis thaliana* gene models and proteome dynamics. Science **320:** 938–941

Ballouz S, Verleyen W, Gillis J (2015) Guidance for RNA-seq co-expression network construction and analysis: safety in numbers. Bioinformatics **31:** 2123–2130

Ballouz S, Weber M, Pavlidis P, Gillis J (2016) EGAD: ultra-fast functional analysis of gene networks. bioRxiv 53868

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

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Statist Soc B **57:** 289–300

Bolduc N, Yilmaz A, Mejia-Guerra MK, Morohashi K, O'Connor D, Grotewold E, Hake S (2012) Unraveling the KNOTTED1 regulatory network in maize meristems. Genes Dev **26:** 1685–1690

Bosch M, Mayer C-D, Cookson A, Donnison IS (2011) Identification of genes involved in cell wall biogenesis in grasses by differential gene expression profiling of elongating and non-elongating maize internodes. J Exp Bot **62:** 3545–3561

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. BMC Bioinformatics **10:** 421

Cigan AM, Unger-Wallace E, Haug-Collet K (2005) Transcriptional gene silencing as a tool for uncovering gene function in maize. Plant J **43:** 929–940

Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szcześniak MW, Gaffney DJ, Elo LL, Zhang X, Mortazavi A (2016) A survey of best practices for RNA-seq data analysis. Genome Biol **17:** 13 doi: 10.1186/s13059-016-0881-8

Davidson RM, Hansey CN, Gowda M, Childs KL, Lin H, Vaillancourt B, Sekhon RS, de Leon N, Kaeppler SM, Jiang N, et al (2011) Utility of RNA sequencing for analysis of maize reproductive transcriptomes. Plant Genome J **4:** 191

De Bodt S, Hollunder J, Nelissen H, Meulemeester N, Inzé D (2012) CORNET 2.0: integrating plant coexpression, protein-protein interactions, regulatory interactions, gene associations and functional annotations. New Phytol **195:** 707–720

Desprez T, Juraniec M, Crowell EF, Jouy H, Pochylova Z, Parcy F, Höfte H, Gonneau M, Vernhettes S (2007) Organization of cellulose synthase complexes involved in primary cell wall synthesis in *Arabidopsis thaliana*. Proc Natl Acad Sci USA **104:** 15572–15577

D'haeseleer P, Liang S, Somogyi R (2000) Genetic network inference: from co-expression clustering to reverse engineering. Bioinformatics **16:** 707–726

Dhillon IS, Modha DS (2001) Concept decompositions for large sparse text data using clustering. Mach Learn **42:** 143–175

Dillies M-A, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J, Guernec G, Jagla B, et al; French StatOmique Consortium (2013) A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief Bioinform **14:** 671–683

Doncheva NT, Assenov Y, Domingues FS, Albrecht M (2012) Topological analysis and interactive visualization of biological networks and protein structures. Nat Protoc **7:** 670–685

Dong J, Horvath S (2007) Understanding network concepts in modules. BMC Syst Biol **1:** 24

Dotto MC, Petsch KA, Aukerman MJ, Beatty M, Hammell M, Timmermans MCP (2014) Genome-wide analysis of leafbladeless1-regulated and phased small RNAs underscores the importance of the TAS3 ta-siRNA pathway to maize development. PLoS Genet **10:** e1004826

Du D, Rawat N, Deng Z, Gmitter FG, Jr. (2015) Construction of citrus gene coexpression networks from microarray data using random matrix theory. Hortic Res **2:** 15026

Du Z, Zhou X, Ling Y, Zhang Z, Su Z (2010) agriGO: a GO analysis toolkit for the agricultural community. Nucleic Acids Res **38:** W64–W70

Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res **30:** 1575–1584

Fadista J, Vikman P, Laakso EO, Mollet IG, Esguerra JL, Taneera J, Storm P, Osmark P, Ladenvall C, Prasad RB, Hansson KB, Finotello F, et al (2014) Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism. Proc Natl Acad Sci USA **111:** 13924–13929

Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS (2007) Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. PLoS Biol **5:** 0054–0066

Fedoroff NV (2012) McClintock's challenge in the 21st century. Proc Natl Acad Sci USA **109:** 20200–20203

Ficklin SP, Feltus FA (2011) Gene coexpression network alignment and conservation of gene modules between two grass species: maize and rice. Plant Physiol **156:** 1244–1256

Filzmoser P, Fritz H, Kalcher K (2009) pcaPP: Robust PCA by Projection Pursuit. R Package, version 1. 9–49. http://CRAN.R-project.org/package=pcaPP

Fu J, Cheng Y, Linghu J, Yang X, Kang L, Zhang Z, Zhang J, He C, Du X, Peng Z, Wang B, Zhai L, et al (2013) RNA sequencing reveals the complex regulatory network in the maize kernel. Nat Commun **4:** 2832

Gao S, ver Steeg G, Galstyan A (2015) Efficient estimation of mutual information for strongly dependent variables. Artificial Intel Statist **38:** 277–286

Gillis J, Pavlidis P (2011) The role of indirect connections in gene networks in predicting function. Bioinformatics **27:** 1860–1866

Giorgi FM, Del Fabbro C, Licausi F (2013) Comparative study of RNA-seq- and microarray-derived coexpression networks in *Arabidopsis thaliana*. Bioinformatics **29:** 717–724

Gu Y, Kaplinsky N, Bringmann M, Cobb A, Carroll A, Sampathkumar A, Baskin TI, Persson S, Somerville CR (2010) Identification of a cellulose synthase-associated protein required for cellulose biosynthesis. Proc Natl Acad Sci USA **107:** 12866–12871

Han Y, Gao S, Muegge K, Zhang W, Zhou B (2015) Advanced applications of RNA sequencing and challenges. Bioinform Biol Insights **9**(**Suppl 1**): 29–46

Horvath S, Dong J (2008) Geometric interpretation of gene coexpression network analysis. PLOS Comput Biol **4:** e1000117

Huang J, Lynn JS, Schulte L, Vendramin S, McGinnis K (2017) Epigenetic control of gene expression in maize. Int Rev Cell Mol Biol **328:** 25–48

Iancu OD, Kawane S, Bottomly D, Searles R, Hitzemann R, McWeeney S (2012) Utilizing RNA-Seq data for de novo coexpression network inference. Bioinformatics **28:** 1592–1597

Kim D, Langmead B, Salzberg SL (2015) HISAT: a fast spliced aligner with low memory requirements. Nat Methods **12:** 357–360

Köster J, Rahmann S (2012) Snakemake—a scalable bioinformatics workflow engine. Bioinformatics **28:** 2520–2522

Krouk G, Lingeman J, Colon AM, Coruzzi G, Shasha D (2013) Gene regulatory networks in plants: learning causality from time and perturbation. Genome Biol **14:** 123

Kumar S, Stecher G, Suleski M, Hedges SB (2017) TimeTree: a resource for timelines, timetrees, and divergence times. Mol Biol Evol **34:** 1812–1819

Kumari S, Nie J, Chen H-S, Ma H, Stewart R, Li X, Lu M-Z, Taylor WM, Wei H (2012) Evaluation of gene association methods for coexpression network construction and biological knowledge discovery. PLoS One **7:** e50411

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

Leinonen R, Sugawara H, Shumway M; International Nucleotide Sequence Database Collaboration (2011) The sequence read archive. Nucleic Acids Res **39:** D19–D21

Li C, Qiao Z, Qi W, Wang Q, Yuan Y, Yang X, Tang Y, Mei B, Lv Y, Zhao H, Xiao H, Song R (2015a) Genome-wide characterization of cis-acting DNA targets reveals the transcriptional regulatory framework of opaque2 in maize. Plant Cell **27:** 532–545

Li H, Peng Z, Yang X, Wang W, Fu J, Wang J, Han Y, Chai Y, Guo T, Yang N, Liu J, Warburton ML, et al (2013a) Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. Nat Genet **45:** 43–50

Li J, Wei H, Zhao PX (2013b) DeGNServer: deciphering genome-scale gene networks through high performance reverse engineering analysis. BioMed Res Int **2013:** 856325

Li P, Piao Y, Shon HS, Ryu KH (2015b) Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data. BMC Bioinformatics **16:** 347

Li Q, Eichten SR, Hermanson PJ, Zaunbrecher VM, Song J, Wendt J, Rosenbaum H, Madzima TF, Sloan AE, Huang J, Burgess DL, Richmond TA, et al (2014a) Genetic perturbation of the maize methylome. Plant Cell **26:** 4602–4616

Li S, Łabaj PP, Zumbo P, Sykacek P, Shi W, Shi L, Phan J, Wu P-Y, Wang M, Wang C, Thierry-Mieg D, Thierry-Mieg J, et al (2014b) Detecting and correcting systematic variation in large-scale RNA sequencing data. Nat Biotechnol **32:** 888–895

Li Y, Pearl SA, Jackson SA (2015c) Gene networks in plant biology: approaches in reconstruction and analysis. Trends Plant Sci **20:** 664–675

Liao Y, Smyth GK, Shi W (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics **30:** 923–930

Lim WK, Wang K, Lefebvre C, Califano A (2007) Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks. Bioinformatics **23:** i282–i288

Liu S, Liu Y, Zhao J, Cai S, Qian H, Zuo K, Zhao L, Zhang L (2017) A computational interactome for prioritizing genes associated with complex agronomic traits in rice (*Oryza sativa*). Plant J **90:** 177–188

Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, Walters G, Garcia F, Young N; GTEx Consortium (2013) The genotype-tissue expression (GTEx) project. Nat Genet **45:** 580–585

Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. Genome Biol **15:** 550

Luo X, You Z, Zhou M, Li S, Leung H, Xia Y, Zhu Q (2015) A highly efficient approach to protein interactome mapping based on collaborative filtering framework. Sci Rep **5:** 7702

Ma C, Wang X (2012) Application of the Gini correlation coefficient to infer regulatory relationships in transcriptome analysis. Plant Physiol **160:** 192–203

Ma H-W, Zeng A-P (2003) The connectivity structure, giant strong component and centrality of metabolic networks. Bioinformatics **19:** 1423–1430

Ma S, Shah S, Bohnert HJ, Snyder M, Dinesh-Kumar SP (2013) Incorporating motif analysis into gene co-expression networks reveals novel modular expression pattern and new signaling pathways. PLoS Genet **9:** e1003840

Marbach D, Costello JC, Küffner R, Vega NM, Prill RJ, Camacho DM, Allison KR, Kellis M, Collins JJ, Stolovitzky G; DREAM5 Consortium (2012) Wisdom of crowds for robust gene network inference. Nat Methods **9:** 796–804

Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, Califano A (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics **7**(**Suppl 1**): S7

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J doi: http://dx.doi.org/10.14806/ej.17.1.200

Maza E, Frasse P, Senin P, Bouzayen M, Zouine M (2013) Comparison of normalization methods for differential gene expression analysis in RNA-Seq experiments: a matter of relative size of studied transcriptomes. Commun Integr Biol **6:** e25849

McClintock B (1983) The significance of responses of the genome to challenge. Science **80:** 792–801

Meyer PE, Kontos K, Lafitte F, Bontempi G (2007) Information-theoretic inference of large transcriptional regulatory networks. Eurasip J Bioinform Syst Biol doi: 10.1155/2007/79879

Meyer PE, Lafitte F, Bontempi G (2008) minet: a R/Bioconductor package for inferring large transcriptional networks using mutual information. BMC Bioinformatics **9:** 461

Monaco MK, Sen TZ, Dharmawardhana PD, Ren L, Schaeffer M, Naithani S, Amarasinghe V, Thomason J, Harper L, Gardiner J, et al (2013) Maize metabolic network construction and transcriptome analysis. Plant Genome **6:** 12

Morohashi K, Casas MI, Falcone Ferreyra ML, Falcone Ferreyra L, Mejía-Guerra MK, Pourcel L, Yilmaz A, Feller A, Carvalho B, Emiliani J, Rodriguez E, Pellegrinet S, et al (2012) A genome-wide regulatory framework identifies maize pericarp color1 controlled genes. Plant Cell **24:** 2745–2764

Morris JH, Apeltsin L, Newman AM, Baumbach J, Wittkop T, Su G, Bader GD, Ferrin TE (2011) clusterMaker: a multi-algorithm clustering plugin for Cytoscape. BMC Bioinformatics **12:** 436

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods **5:** 621–628

Mukaka MM (2012) Statistics corner: a guide to appropriate use of correlation coefficient in medical research. Malawi Med J **24:** 69–71

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

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

Oshlack A, Wakefield MJ (2009) Transcript length bias in RNA-seq data confounds systems biology. Biol Direct **4:** 14

Park T, Yi S-G, Kang S-H, Lee S, Lee Y-S, Simon R (2003) Evaluation of normalization methods for microarray data. BMC Bioinformatics **4:** 33

Paulson J, Chen C-Y, Lopes-Ramos CM, Kuijjer ML, Platig J, Sonawane AR, Fagny M, Glass K, Quackenbush J (2016) Tissue-aware RNA-Seq

processing and normalization for heterogeneous and sparse data. bioRxiv 81802

Pautler M, Eveland AL, LaRue T, Yang F, Weeks R, Lunde C, Il, JB, Meeley R, Komatsu M, Vollbrecht E, et al (2015) FASCIATED EAR4 encodes a bZIP transcription factor that regulates shoot meristem size in maize. Plant Cell Online **2:** tpc.114.132506

Penning BW, Hunter III CT, Tayengwa R, Eveland AL, Dugard CK, Olek AT, Vermerris W, Koch KE, McCarty DR, Davis MF, Thomas SR, McCann MC, et al (2009) Genetic resources for maize cell wall biology. Plant Physiol **151:** 1703–1728

Ponnala L, Wang Y, Sun Q, van Wijk KJ (2014) Correlation of mRNA and protein abundance in the developing maize leaf. Plant J **78:** 424–440

Rau A, Gallopin M, Celeux G, Jaffrézic F (2013) Data-based filtering for replicated high-throughput transcriptome sequencing experiments. Bioinformatics **29:** 2146–2152

Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics **26:** 139–140

Sales G, Romualdi C (2011) Parmigene—a parallel R package for mutual information estimation and gene network reconstruction. Bioinformatics **27:** 1876–1877

Sangrador-Vegas A, Mitchell AL, Chang H-Y, Yong S-Y, Finn RD (2016) GO annotation in InterPro: why stability does not indicate accuracy in a sea of changing annotations. Database (Oxford) Published online 2016 Mar 19. doi: 10.1093/database/baw027

Sato Y, Antonio BA, Namiki N, Takehisa H, Minami H, Kamatsuki K, Sugimoto K, Shimizu Y, Hirochika H, Nagamura Y (2011) RiceXPro: a platform for monitoring gene expression in japonica rice grown under natural field conditions. Nucleic Acids Res **39:** D1141–D1148

Schaefer RJ, Briskine R, Springer NM, Myers CL (2014) Discovering functional modules across diverse maize transcriptomes using COB, the co-expression browser. PLoS One **9:** e99193 10.1371/journal.pone.0099193

Schmidt D (2016) Co-Operation: Fast Correlation, Covariance, and Cosine Similarity. R Package, Version 0.6-0. https://cran.r-project.org/package=coop

Schnable P, Ware D, Fulton R, Stein J (2009) The B73 maize genome: complexity, diversity, and dynamics. Science **326:** 1112–1115

Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M (2011) Global quantification of mammalian gene expression control. Nature **473:** 337–342

Scott-Boyer M-P, Haibe-Kains B, Deschepper CF (2013) Network statistics of genetically-driven gene co-expression modules in mouse crosses. Front Genet **4:** 291

Serin EAR, Nijveen H, Hilhorst HWM, Ligterink W (2016) Learning from co-expression networks: possibilities and challenges. Front Plant Sci **7:** 444

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res **13:** 2498–2504

Sing T, Sander O, Beerenwinkel N, Lengauer T (2005) ROCR: visualizing classifier performance in R. Bioinformatics **21:** 3940–3941

Singh M, Goel S, Meeley RB, Dantec C, Parrinello H, Michaud C, Leblanc O, Grimanelli D (2011) Production of viable gametes without meiosis in maize deficient for an ARGONAUTE protein. Plant Cell **23:** 443–458

Song L, Langfelder P, Horvath S (2012) Comparison of co-expression measures: mutual information, correlation, and model based indices. BMC Bioinformatics **13:** 328

Stelpflug SC, Sekhon RS, Vaillancourt B, Hirsch CN, Buell CR, de Leon N, Kaeppler SM (2016) An expanded maize gene expression atlas based on RNA-sequencing and its use to explore root development. Plant Genome **9:** 314–362

Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, Kuhn M, Bork P, et al (2015) STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res **43:** D447–D452

Tzfadia O, Diels T, De Meyer S, Vandepoele K, Aharoni A, Van de Peer Y (2016) CoExpNetViz: comparative co-expression networks construction and visualization tool. Front Plant Sci **6:** 1194

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

USDA (2016) Grain: World Markets and Trade https://www.fas.usda.gov/data/grain-world-markets-and-trade

Vanholme R, Demedts B, Morreel K, Ralph J, Boerjan W (2010) Lignin biosynthesis and structure. Plant Physiol **153:** 895–905

Verdin E, Ott M (2015) 50 years of protein acetylation: from gene regulation to epigenetics, metabolism and beyond. Nat Rev Mol Cell Biol **16:** 258–264

Walley JW, Sartor RC, Shen Z, Schmitz RJ, Wu KJ, Urich MA, Nery JR, Smith LG, Schnable JC, Ecker JR, Briggs SP (2016) Integration of omic networks in a developmental atlas of maize. Science **353:** 814–818

Wang W, Zhou X, Liu Z, Sun F (2015a) Network tuned multiple rank aggregation and applications to gene ranking. BMC Bioinformatics **16(Suppl 1**): S6

Wang YXR, Jiang K, Feldman LJ, Bickel PJ, Huang H (2015b) Inferring gene–gene interactions and functional modules using sparse canonical correlation analysis. Ann Appl Stat **9:** 300–323

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet **10:** 57–63

Wei C, Li J, Bumgarner RE (2004) Sample size for detecting differentially expressed genes in microarray experiments. BMC Genomics **5:** 87

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

Wu D-D, Wang X, Li Y, Zeng L, Irwin DM, Zhang Y-P (2014) "Out of pollen" hypothesis for origin of new genes in flowering plants: study from *Arabidopsis thaliana*. Genome Biol Evol **6:** 2822–2829

Yang H, Liu X, Xin M, Du J, Hu Z, Peng H, Rossi V, Sun Q, Ni Z, Yao Y (2016) Genome-wide mapping of targets of maize histone deacetylase HDA101 reveals its function and regulatory mechanism during seed development. Plant Cell **28:** 629–645

Yim WC, Yu Y, Song K, Jang CS, Lee B-M (2013) PLANEX: the plant co-expression database. BMC Plant Biol **13:** 83

Zheng W, Chung LM, Zhao H (2011) Bias detection and correction in RNA-sequencing data. BMC Bioinformatics **12:** 290

Zhong R, Allen JD, Xiao G, Xie Y (2014) Ensemble-based network aggregation improves the accuracy of gene network reconstruction. PLoS One **9:** e106319

Zhong R, Ye ZH (2015) Secondary cell walls: biosynthesis, patterned deposition and transcriptional regulation. Plant Cell Physiol **56:** 195–214

Zhu G, Wu A, Xu X-J, Xiao P, Lu L, Liu J, Cao Y, Chen L, Wu J, Zhao X-M (2016) PPIM: a protein-protein interaction database for Maize. Plant Physiol **170:** pp.15.01821

Zyprych-Walczak J, Szabelska A, Handschuh L, Górczak K, Klamecka K, Figlerowicz M, Siatkowski I (2015) The impact of normalization methods on RNA-Seq data analysis. BioMed Res Int **2015:** 621690