in silico Plants Vol. 2, No. 1, pp. 1–15
doi:10.1093/insilicoplants/diaa002
Advance Access publication 28 February 2020.
Original research

Gene regulatory networks associated with lateral root and nodule development in soybean

Shuchi Smita[1,4], Jason Kiehne[2], Sajag Adhikari[1,5], Erliang Zeng[3,6], Qin Ma[1,7*] and Senthil Subramanian[1*]

[1]Department of Agronomy, Horticulture and Plant Science and Department of Biology and Microbiology, South Dakota State University, Brookings, SD 57007, USA
[2]Simpson College, Indianola, IA 50125, USA
[3]Department of Biology and Department of Computer Science, University of South Dakota, Vermillion, SD 57069, USA
[4]Present address: Department of Computational and Systems Biology and Department of Immunology, School of Medicine, University of Pittsburgh, Pittsburgh, PA 15261, USA
[5]Present address: Celerion Inc., Lincoln, NE 68502, USA
[6]Present address: Division of Biostatistics and Computational Biology, College of Dentistry, University of Iowa, Iowa City, IA 52242, USA
[7]Present address: Department of Biomedical Informatics, Ohio State University, Columbus, OH 43210, USA
*Corresponding authors' e-mail addresses: Qin.Ma@osumc.edu; Senthil.Subramanian@sdstate.edu

Citation: Smita S, Kiehne J, Adhikari S, Zeng E, Ma Q, Subramanian S. 2020. Gene regulatory networks associated with lateral root and nodule development in soybean. *In Silico Plants* **2020**: diaa002; doi: 10.1093/insilicoplants/diaa002

ABSTRACT

Legume plants such as soybean produce two major types of root lateral organs, lateral roots and root nodules. A robust computational framework was developed to predict potential gene regulatory networks (GRNs) associated with root lateral organ development in soybean. A genome-scale expression data set was obtained from soybean root nodules and lateral roots and subjected to biclustering using QUBIC (**QU**alitative **BIC**lustering algorithm). Biclusters and transcription factor (TF) genes with enriched expression in lateral root tissues were converged using different network inference algorithms to predict high-confidence regulatory modules that were repeatedly retrieved in different methods. The ranked combination of results from all different network inference algorithms into one ensemble solution identified 21 GRN modules of 182 co-regulated genes networks, potentially involved in root lateral organ development stages in soybean. The workflow correctly predicted previously known nodule- and lateral root-associated TFs including the expected hierarchical relationships. The results revealed distinct high-confidence GRN modules associated with early nodule development involving AP2, GRF5 and C3H family TFs, and those associated with nodule maturation involving GRAS, LBD41 and ARR18 family TFs. Knowledge from this work supported by experimental validation in the future is expected to help determine key gene targets for biotechnological strategies to optimize nodule formation and enhance nitrogen fixation.

KEYWORDS: Biclustering; gene regulatory network; Inferelator; Lemon-Tree; QUBIC; root nodule; soybean.

INTRODUCTION

Gene regulation is a fundamental process that controls the spatial and temporal patterns of gene expression. Transcription factors (TFs) are central to gene regulation as their activities determine the expression patterns and function of multiple genes (Eeckhoute *et al.* 2009). A TF is a functional protein that binds to short sequences (called TF binding site or *cis*-regulatory elements) on the upstream promoter region of genes to regulate their transcription. One TF can regulate multiple genes including other TFs in signalling, developmental or metabolic pathway. Therefore, TFs act as master regulators of pathways. The nested group of all different TF regulators and their downstream target genes form gene regulatory networks (GRNs) (Blais and Dynlacht 2005). Identification of GRNs and key TFs that are part of these networks is an effective first step to answer multiple biological questions on genotype to phenotype relationships (Petricka and Benfey 2011; Kim and Przytycka 2013). Potential TFs, their co-regulators, downstream signalling pathways and target genes associated with specific biological processes can be predicted by constructing GRNs.

Clustering of large-scale data sets such as global gene expression profiles obtained by RNA sequencing to identify co-regulated TFs and the targeting genes is a promising approach to model and infer the GRNs at a systems level (Udvardi *et al.* 2007; Baitaluk *et al.* 2012). For example, grouping genes/TFs with similar expression patterns (i.e. co-expressed genes) across a group of samples might give insight on TF regulated gene networks and related biological processes. In addition, gene expression is regulated at multiple levels through different mechanisms (Kaufmann *et al.* 2010). Recruitment and binding of other proteins such as 'co-factors' in TF complexes to tightly regulate the location or the extent of transcription is one of the major mechanisms (Guan *et al.* 2014). Often, these interactions between different TFs and co-factor partners are studied using protein–protein interaction (PPI) assays, which provide novel insights into their potential biological function (Rivas and Fontanillo 2010; Szklarczyk *et al.* 2017). Indication of PPIs among co-regulated genes can add confidence to GRN predictions, and PPIs can reveal signalling, regulatory and/or biochemical roles of proteins based on their interactomes (Chaturvedi *et al.* 2007).

The combined use of high-throughput data and mathematical models to build gene co-expression and regulatory networks is the core principle of many systems biology approaches (Sun and Zhao 2009). However, these large-scale data sets are likely to be noisy, and GRN predictions using these big data sets may contain a high number of false positives. Additionally, GRN inference is a computationally intensive job; so filtered data sets consisting of well-defined/accurate data sets (such as significantly co-expressed genes set) might dramatically reduce the computational complexity and time. Most importantly, it would reduce the true search space for the prediction of regulator TFs and their potential target genes and minimize false positives. In order to obtain significantly co-expressed genes, 'biclustering' is a desirable method as it allows two-way clustering of genes as well as samples, i.e. a similar expression pattern (co-expressed genes) under a subset of all samples. Subsequently, the use of sorted, biclustering-filtered data for GRN inference might improve the TF regulator and target gene prediction accuracy. We applied this approach to determine GRNs associated with root lateral organ development in soybean.

Plants produce lateral organs such as leaves, flowers and lateral roots (LRs). Pools of stem cells present in the growing tip of the shoot (the shoot apical meristem) contribute to the formation of aerial/shoot lateral organs. Lateral organs in the root are unique in that they are derived via *de novo* differentiation of mature cells in the root. Lateral roots are present in all vascular plants, but a group of *Fabids* clade plants is capable of producing another root lateral organ, called 'root nodules'. These arise from specific and coordinated interactions with a set of nitrogen-fixing bacteria collectively called rhizobia. For example, the interaction of soybeans with *Bradyrhizobium diazoefficiens* results in root nodules. Biological nitrogen fixation in root nodules helps reduce the need for chemical nitrogen fertilizers, which are expensive and cause environmental pollution. Similarly, proper patterns of LR formation (root branching) are crucial for plants to access water and other nutrients in the soil. Therefore, these two root lateral organs play important roles in the development of soybeans, a major crop in the USA as well as in other countries. A number of genetic and systems biology studies especially in the model plant *Arabidopsis thaliana* have identified developmental pathways and regulators involved in LR development (Benková and Bielach 2010; Atkinson *et al.* 2014; Du and Scheres 2018). Many functional genomics studies have identified genes expressed during nodule development in soybean and other legumes (Zhu *et al.* 2013; Li and Jackson 2016). However, only a few regulators associated with nodule development are known and these were identified primarily using genetic approaches that require the presence of a clear developmental phenotype (Roy *et al.* 2020).

Recently, we obtained transcriptomes of emerging nodules (ENs), mature nodules (MNs), emerging LRs (ELRs) and young LRs (YLRs) in soybean (Adhikari *et al.* 2019). This allowed us to identify genes and TFs that are enriched specifically in nodule tissues and not in LRs. We present a robust computational framework, which we applied to predict TFs and their target GRNs associated with soybean root nodule development. This approach consists of the following steps (Fig. 1): (i) preparing a compendium of soybean lateral organ transcriptome data and cataloging TFs enriched in root nodules; (ii) initial biclustering of transcriptome data using the *QU*alitative *BI*Clustering (QUBIC) algorithm (Li et al. 2009; Xie et al. 2019; Zhang et al. 2017) to identify all (nodule development stage-specific) co-expressed gene modules; (iii) GRN construction and inference using reliable network construction programs, Lemon-Tree (Bonnet *et al.* 2015) and Inferelator (Greenfield *et al.* 2010); (iv) augmentation of GRNs with evidence from physical or direct and indirect regulatory interaction information from PPI and *cis*-regulatory element enrichment analysis; and (v) building a consensus from different modes of GRN inference for potential regulators and their predicted targets. We ran two modes of Lemon-Tree: one with default mode, where Lemon-Tree itself produced the co-expressed clusters and the other mode where Lemon-Tree used reinforced bicluster (BC) information from QUBIC. This study provides a template framework for GRN construction and augmentation by exploiting big datasets, which are being generated, deposited and made available (making use of available data) in public domain at a rapid rate.

MATERIALS AND METHODS

RNA-seq data set for root lateral organ development in soybean

We utilized a genome-wide soybean transcriptome data set generated from root lateral organs (Adhikari *et al.* 2019). This data set contains the transcriptomes of two different developmental stages of two root lateral organs collected in three biological replicates: ENs, MNs, ELRs and YLRs. Adjacent root sections above and below these organs devoid of any lateral organs (designated as ABEN, ABMN, ABELR and ABYLR, respectively) were used to construct respective age- and inoculation-status appropriate control tissue libraries. Comparison of gene expression profiles between each lateral organ tissue type and the corresponding control tissue type (e.g. EN vs. ABEN, ELR vs. ABELR and so on) helped to identify organ-specific/enriched genes. In total, 24 RNA-seq libraries (four target tissue types, four control tissue types, three biological replicates each) were prepared, sequenced and analysed. Expression patterns of previously known marker genes, consistency between replicates, and high sequence quality of this data set indicated that it was well-suited for global gene expression analysis (Adhikari *et al.* 2019). A total of 113 210 gene transcripts (FPKM

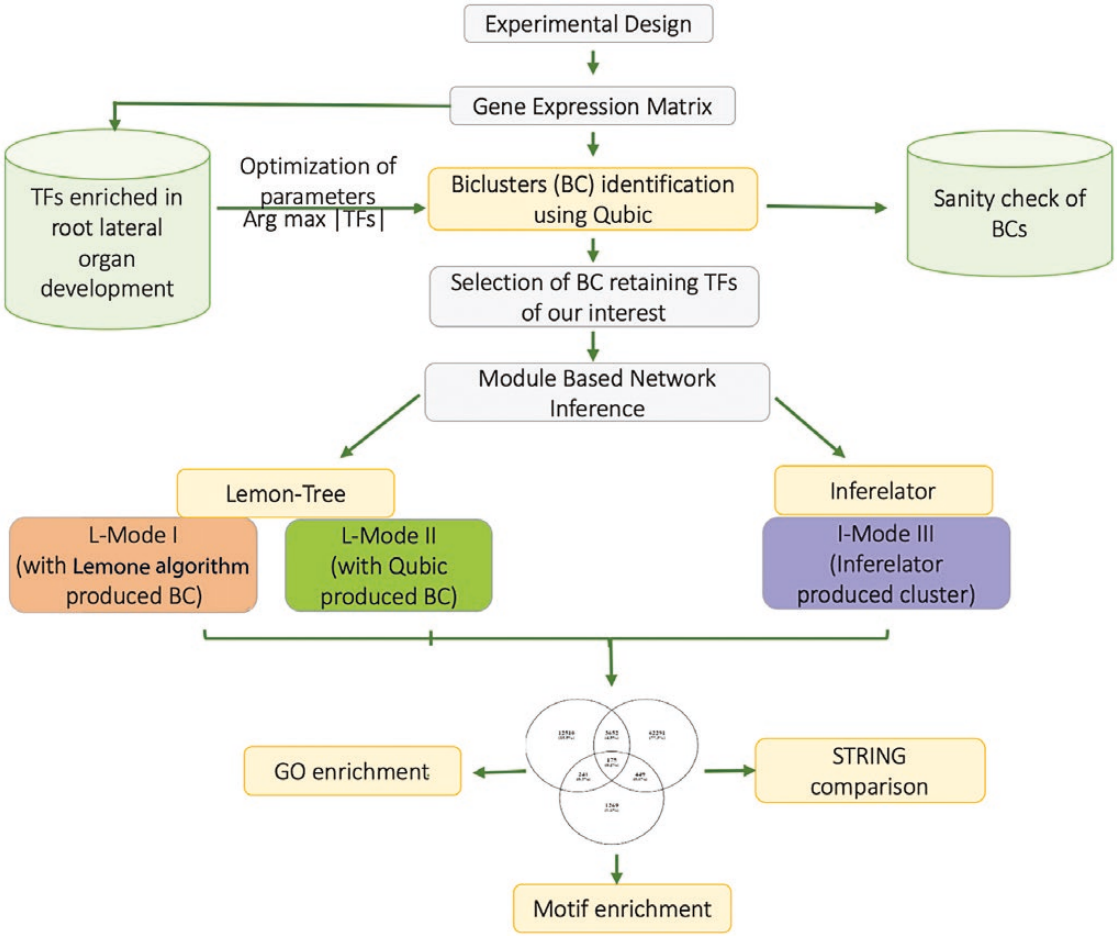

Figure 1. Schematic representation showing our workflows for prediction of regulator TFs and their GRNs associated with root lateral organ development in soybean.

threshold ≥1 in at least one sample) with their normalized expression values in 24 different tissues from the above data set were utilized here.

Furthermore, for expression comparisons at different steps during our analysis, we utilized the following public data sets: Soybean Gene Atlas encompassing RNA-seq data from 14 different soybean tissues (Severin *et al.* 2010) and Soybean eFP Browser, http://bar.utoronto.ca/efpsoybean/cgi-bin/efpWeb.cgi, comprising RNA-seq data from soybean root hair and other tissues (Libault *et al.* 2010*a*,*b*) to evaluate organ-specific enrichment, and Soybean Genome Sequence Assembly version 7.0 (Gmax_109_gene.gff3.gz; ftp://ftp.jgi-psf.org/pub/compgen/phytozome/v9.0/Gmax/annotation/) to obtain gene annotation and Arabidopsis ortholog information.

Cataloging TFs enriched in different stages of root lateral organ development in soybean

To achieve our objective of identifying regulator TFs and prediction of GRNs associated with root nodules, we used soybean transcription factor annotations from the Plant Transcription Factor Database (PlantTFDB v3.0; http://planttfdb.cbi.pku.edu.cn/) (Jin *et al.* 2014) as a starting point. Among the 58 TF families annotated in soybean, 48 TF families had at least one member differentially expressed in at least one of the four organ tissue types in our data. For each TF family, we summed the unique transcripts that were enriched in EN and/or MN to calculate the total number of family members enriched in nodule tissues. Similarly, we calculated the number of TFs enriched in LR tissues. By comparing the number of family members enriched in nodule versus LR tissues, we identified nodule-specific or enriched, LR-specific or enriched and lateral organ non-specific (equal number of transcripts in LR and nodules) TF families (Fig. 2). Statistical analysis (Fisher's exact test, $P < 0.05$) of nodule- versus LR-specific enrichment showed that TFs belonging to TALE, MYB-related, MIKC, C2H2, bZIP, G2-like, WRKY and NFYB families were either nodule-specific or significantly enriched in nodules. Overall, very distinct families of TFs appear to be active either in nodule or LRs despite reported morphological similarities between these organs.

We selected a set of 294 TFs, which were specifically enriched in EN, and MN tissues in our data set as possible regulators (see Results; **Supporting Information—Table S1**). This approach led us to focus

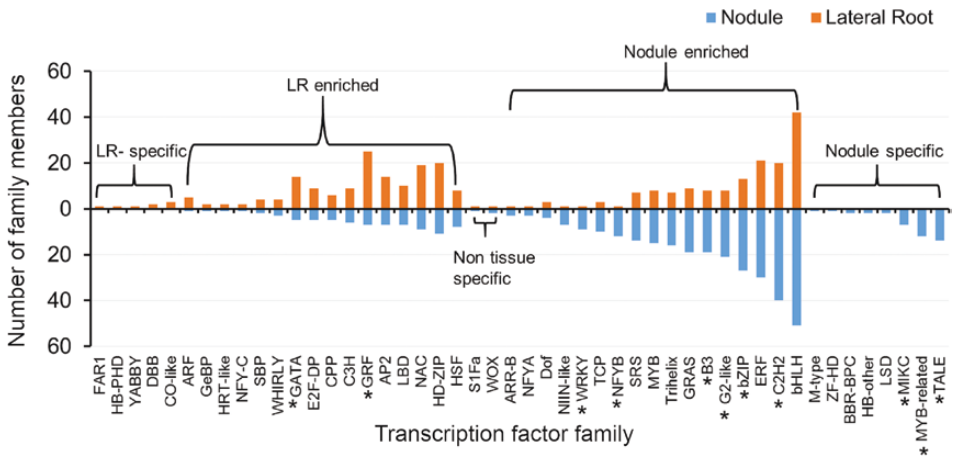

Figure 2. TF families enriched in specific root lateral organs. Bar graphs indicated the number of family members enriched in nodules (blue) or lateral roots (orange). TFs enriched only on nodules and not in LRs are denoted as nodule-specific (and vice versa for LR-specific). TFs with more family members enriched in nodules versus LRs are denoted as nodule-enriched (and vice versa for LR-enriched). TF annotations are based on Plant Transcription Factor Databases (PlantTFDB). Asterisks (*) indicate TF families that were significantly enriched either in nodule or lateral root (Fisher's exact test; $P < 0.05$).

on regulators and their GRNs acting specifically during nodule development. We also included 32 previously characterized TFs/organ-specific marker genes reported elsewhere in literature for their respective role in root lateral organ development in model crop plants as positive controls for validation and relevancy of parameters [**see Supporting Information—Table S2**]. For example, *ENOD40*, *FWL1*, *LBC_A*, *LBC_C1*, *LBC_C2* and *LBC_C3* genes were used as marker genes, and *NIN1* and *NSP1* were used as marker regulators for nodule development. *ARF5*, *CRF2*, *GATA23*, *LRP1* and *TMO7* genes were used as marker regulators for LR development. Together, we used 316 TFs of interest as a starting point for the identification of GRNs.

Initial biclustering of transcriptome data

We utilized normalized expression values of all the 113 210 gene transcripts in 24 libraries for initial biclustering, rather than only significantly differentially expressed gene transcripts. We reasoned that irrespective of enrichment, the TFs and their target gene clusters tend to have similar expression patterns in the root lateral organs, making this an unbiased approach. We chose biclustering (two-way clustering) using QUBIC (Li *et al.* 2009), over traditional clustering to simultaneously identify all the statistically significant BCs of target genes with TFs (if any) as well as the samples where these BCs originated from. Different combinations of QUBIC's parameters were tuned to optimize biclustering to retain the majority of TFs while keeping the total number of co-expressed gene transcripts to the minimum. The program first discretizes the data using the parameters q and r and then a heuristic algorithm is applied to identify BCs, where q is the proportion of affected expression data under all conditions for each gene and r represents the rank of the regulating conditions detected by the parameter q. It is suggested to select a smaller q to focus on a local regulator (Li *et al.* 2009). Parameter f controls the overlap between different BCs and k controls the minimum number of samples in BCs. Another important parameter c, which controls the level of consistency in BCs, was tested to balance the number of TFs and the total number of co-expressed genes covered in BCs. We obtained 219 BCs that contained 240 of the 316 TFs (76 %) and 30 639 out of 113 210 transcripts (~27 %; see Results for details). The output from QUBIC is available in **Supplementary Information—Data Set S1**. This 'filtered' data set was used for regulator and GRN prediction. All programs were tested and implemented on a Linux server with Intel x86-64 processor and 32 cores with 1TB RAM configuration.

Prediction of potential TF regulators and their GRN inference

To improve the confidence of regulator and GRN prediction, we utilized two module-based GRN inference methods: Lemon-Tree (v.3.0) (Bonnet *et al.* 2015) and Inferelator (v.2015.08.05) (Bonneau *et al.* 2006). We compared and scored regulatory predictions from both methods to select high-confidence regulators and their target genes in GRNs.

Lemon-Tree

Lemon-Tree has the option to integrate cluster information; hence, we ran it in two modes: (i) where clusters were generated by Lemon-Tree from the 'filtered' data set (mode I) and (ii) where BC information from QUBIC was fed to Lemon-Tree as co-expressed gene modules for GRN inference (mode II). For mode I, we ran 10 independent Gibbs sampler runs of Lemon-Tree (with default parameters) to identify the most confident regulatory modules and TF regulators. The results were used to extract representative module solutions (tight clusters) from an ensemble of all possible statistical models using the Gibbs sampler method. Lemon-Tree modules are clustered (hierarchical tree) based on samples with similar mean and standard deviation. This tight cluster corresponds to sets of genes, frequently associated across all clustering solutions. For mode II, we prepared this tight cluster data

set using BCs information from QUBIC, but otherwise used the same settings used for mode I.

In the next step, the Lemon-Tree algorithm provides a list of weighted TFs with a ranked probability score. The top 1 % regulators with a high score were selected as potential high-confidence regulators for each cluster of co-expressed genes. A global score reflecting the statistical confidence of the regulator assigned to each node in a hierarchal tree manner for each set of co-expressed genes modules was also provided. The regulator score takes into account the number of trees a regulator was assigned to, with what score (posterior probability), and at which level of the tree (Joshi *et al.* 2009). An empirical distribution of scores for randomly assigned regulators-to-module was also provided to assess significance (Bonnet *et al.* 2015). In this data set, the lowest score of a regulator in the top 1 % list was at least three times higher than that of the highest score for a randomly assigned regulator (see Results for details). Therefore, either the top 1 % or at least a 3-fold higher score than randomly assigned regulators appears to be a good threshold to determine high-confidence regulators.

Inferelator

Inferelator (20 bootstraps) was utilized with default settings to build regulatory networks (Bonneau *et al.* 2006). Similar to Lemon-Tree, it also uses the gene expression matrix to predict the regulator TFs and their target genes. However, unlike Lemon-Tree, Inferelator does not take cluster information as input, but generates its own clusters. The program generated a ranked list of target genes for each regulator TF utilizing the gene expression matrix and the TFs of our interest. Unlike Lemon-Tree, there is no 'score-based' selection of TFs in Inferelator, while there are score-based regulatory interactions between TF and their target genes. Inferelator-generated scores (s) for TF (x) regulating gene (g) using input gene expression matrix (RNA-seq) as:

$$s\,(x \rightarrow g \mid \text{RNA-seq}) = \text{Inferelator}\,(x \rightarrow g \mid \text{RNA-seq}) \times \text{sign}\,(\text{cor}\ (x, g))$$

where a regulatory interaction confidence score is multiplied by the sign of the correlation coefficient between the TF and the putative target gene to differentiate activating versus repressing interactions (positive and negative scores, respectively) (Greenfield *et al.* 2010; Ciofani *et al.* 2012).

Combined scoring of regulatory predictions for consensus GRN

By taking advantage of the top regulator prediction feature of Lemon-Tree and top-ranked regulatory target prediction of Inferelator, we compared and combined TF and targeted module genes from all three inference solutions: Lemon-Tree mode I, II and Inferelator (described above). The regulatory TFs and corresponding target genes common among all three inference solutions using Linux '*comm*' command were rated as potential consensus regulators and their targeted GRN interactions. Ranked score function for every predicted regulatory interaction was calculated by normalizing scores produced by each inference solution (score divided by the highest score in each inference solution) and then averaging normalized score calculated from all three inference solutions. These ranked scores were used to select high-confidence candidate TF–target interactions. These scores were shown as edges in the GRN modules, visualized and analysed using Cytoscape (version 3.3.0) (Shannon *et al.* 2003).

$$\text{Average score, As} = \frac{\sum \text{Ns (L-mode I)}, \text{Ns (L-mode II)}, \text{Ns(Inferelator)}}{3}$$

Where $\text{Ns} = \frac{x}{X}$; Ns = normalized score; x = probabilistic score from each mode; X = maximum score in each mode; L-mode I = Lemon-Tree mode I; L-mode II = Lemon-Tree mode II.

Evidence for putative PPIs

Most eukaryotic TFs recruit various co-factors for their DNA-binding specificities and regulatory activities through PPIs. To evaluate potential PPIs that are part of the predicted GRNs, a total of 31 932 066 predicted/experimentally validated soybean protein interactions (NCBI taxon-Id:3847) were obtained from the STRING database (version 10.0) (search tool for the PPI network) (Szklarczyk *et al.* 2015). This database provides information on both experimental and predicted interactions (both physical and functional interactions) from varied sources based on co-expression, experiments and literature mining. We evaluated and compared if the predicted TFs and targets from the different inference solutions (Lemon-Tree mode I, II and Inferelator) were potential PPI partners using all the 31 932 066 STRING PPI interactions in soybean. A non-redundant data set, ignoring the transcript numbers of TFs, targets (from TF–target interactions) predicted by three individual inference solutions and PPI from STRING were compared using the Linux '*comm*' command to identify TF–target pair common in the STRING data set.

Cis-regulatory motif and functional enrichment analysis evidence for direct regulation

Cis-regulatory motif enrichment was carried out using potential promoter sequences (300 bp downstream–600bp upstream of predicted transcription start sites) of target genes for all potential regulator TFs predicted by all three inference solutions (Lemon-Tree mode I, II and Inferelator). Motif enrichment and Gene Ontology (GO) were performed by ShinyGO (http://www.ge-lab.org:3838/go/) using *P*-value cutoff (FDR-corrected; Benjamini–Hochberg enrichment) <0.05 to determine regulation and function.

RESULTS

Optimization of QUBIC parameters for initial biclustering

The primary goal for biclustering in our analysis was to optimize the total number of significant BCs, where the majority of the TFs (out of TFs of interest and marker TFs) are retained while keeping the total number of co-expressed genes to a minimum for GRN prediction. In order to evaluate this condition, we iterated various runs in several steps to empirically optimize key QUBIC parameters. For example, to

focus on a local regulator that typically involve smaller regulatory networks, smaller *q* values were used. To control the overlap by checking the overlapping genes and the number of TFs in between produced BCs, we iterated the run with $f = 0.5–0.65$ (by 0.05). We used $k = 6$ presumably to retain at least three replicates each from either early or late developmental stages or from LR or nodule tissue types in one BC. Importantly, the consistency level of BCs was tested using parameter '*c*' iterated from $c = 0.6–1$ (by 0.1) to balance the number of TFs and a total number of genes covered in BCs (Fig. 3). We noted that the lower the value of consistency level '*c*', the larger was the size of the BC. We evaluated the change in a number of total TFs versus total genes in BCs with increasing consistency levels with the goal of determining the '*c*' value at which we covered the greatest number of TFs in comparison to a total number of genes without losing much consistency (*c*). At $c = 0.98$, 76 % of the TFs of interest were retained with just 27 % of the genes covered in BCs (Fig. 3). The maximum number of marker TFs (18 out of 22) cataloged for root lateral organs were covered at $c = 0.98$, suggesting this to be an ideal value for our analysis. On the other hand, at the highest consistency level ($c = 1$), only three marker TFs were covered in BCs (not shown). Overall, based on results from several iterations and optimizing for the inclusion of greater number of TFs in BCs, we finalized the following parameters: $r = 1$, $q = 0.2$, $c = 0.98$, $o = 500$, $f = 0.25$, $k = 6$, which produced 219 statistically significant BCs. These 219 BCs comprised ~27 % (30 639 out of 113 210) of total gene transcripts. Notably, ~76 % (240 out of 316 TFs of our interest) of the TFs of interest (including 81 % of marker TFs) were retained in 141 of the 219 total BCs produced. The first cluster was the largest cluster with a total of 446 genes. We conclude that the empirical determination of biclustering parameters depending on the biological question and the associated experimental objective is crucial for useful outcomes. The underlying reason is that QUBIC is a heuristic algorithm for two-dimensional clustering without any hidden statistical hypotheses for the minimal number of samples, the number of to-be-identified BCs or the size of a BC.

Evaluation of QUBIC BCs using characterized TFs and co-expressed genes from public LR organ-related data sets

We observed one organ-specific BC each for LR (both ELR and YLR; BC001) and nodule (both EN and MN; BC013) tissues that included all three biological replicate samples in one BC, suggesting that these are likely to be highly consistent and reproducible. Four BCs each were specific to all three replicates of ELR (BC015, 019, 033 and 101) and MN (044, 048, 152 and 155) tissue types. To test the rationality of

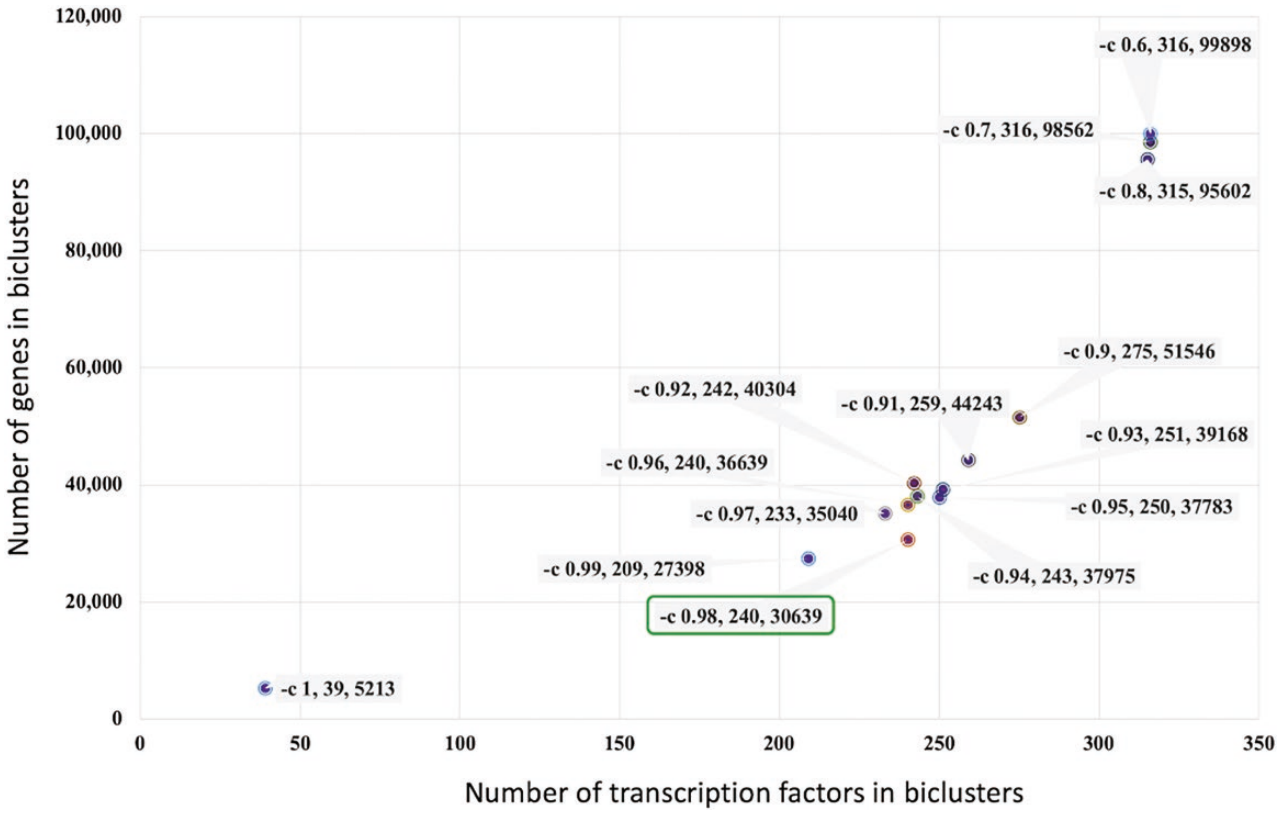

Figure 3. Optimization of QUBIC's consistency parameter. Relationship between QUBIC's consistency parameter '*c*' (tested from 0.6 to 1) and the number of target TFs included in BC versus the size of the BC (total number of genes). Each block denotes –*c* value, TF included in BCs and total number of genes at that '*c*' value. The optimal '*c*' value (0.98) selected for final analysis is highlighted with a green box.

BCs, we compared the expression patterns of co-expressed genes with marker TFs in publicly available transcriptome data (Severin *et al.* 2010). The transcription factor *NSP1* (Glyma16g01020), crucial for nodule development, was present in BC037 and BC045. BC037 was specific to nodule tissues and comprised of 367 co-expressed genes. Among these, 52 % had more than 2-fold up-regulation in EN and MN tissues in our RNA-seq data. A marker gene highly enriched in nodule tissues, *ENOD40* (Glyma02g04180), was found in five BCs (BC013, 22, 40, 45, 53 and 95) with different combinations of nodule samples clustered together in each BC. All genes in BC013 showed specificity for nodule tissue samples with all three replicates in EN and MN in our study. Also, 50 % of the genes from this BC showed greater expression in nodule tissue relative to other tissues types in the soybean Gene Expression Atlas (Severin *et al.* 2010) [**see Supporting Information—Table S3**]. Gene Ontology enrichment analysis for this BC indicated enrichment of the biological process GO term 'nucleic acid metabolic process' (FDR-corrected *P*-value 0.02) and molecular function GO term 'Purine ribonucleoside triphosphate binding' (FDR-corrected *P*-value 0.05), both of which are associated with biological nitrogen fixation that occurs specifically in nodule tissues. For example, soybean nodules export nitrogen in the form of ureides (purines) (Collier and Tegeder 2012). The above observations indicate the appropriate clustering of relevant transcripts and validate the parameters used for clustering. Notably, we observed few novel transcripts and genes with unknown function, co-expressed in the nodule-specific BCs [**see Supporting Information—Table S3**]. This observation suggests a potential role for these genes in nodule development and offers candidate genes for functional characterization.

Furthermore, we took advantage of the time course data for IAA-induced LR development in Arabidopsis (Lewis *et al.* 2013), to select and evaluate marker genes present in LR-related BCs in soybean. For example, the LR marker TF, *GmTMO7* (Glyma04g34080), a potential ortholog of Arabidopsis *TMO7* identified in the above study, was present in BCs 110, 120 and 173. Of the 113 genes present in BC120, 96 showed coordinated up-regulation with *TMO7* in LR tissues, whereas 17 showed negative co-expression. We evaluated the expression patterns of potential Arabidopsis orthologs of these genes in the LR induction time course data set (Lewis *et al.* 2013). Data were available for orthologs of 13 of the 96 positively correlated genes and 2 of 17 negatively correlated genes. Many of these Arabidopsis orthologs (see marked blue and red box in **Supporting Information—Table S4**) were induced in roots with LR primordia at stage 4 and beyond which corresponds to soybean ELR and YLR tissues we used in our study. This suggested that the potential orthologs have a similar expression pattern during LR development in both Arabidopsis and soybean. The other LR marker *LRP1* was in BC019 that comprised of 845 genes. Among these genes, 746 were positively and 99 were negatively co-expressed with *LRP1* in all three replicates of ELR. Arabidopsis orthologs of 30 positively co-expressed genes also showed induction during a similar developmental stage [**see Supporting Information—Table S4**] in the LR induction time course data set (Lewis *et al.* 2013). This suggested that the biclustering parameters used indeed helped to group functionally relevant co-expressed genes together. Therefore, using the identified BCs as input GRN algorithms can identify regulators and regulatory relationships of target genes with higher efficiency and fewer false positives (due to spurious correlations).

Regulatory TFs and their GRNs associated with root lateral organ development in soybean

For the prediction of regulators and inference of corresponding GRNs, we utilized only those 141 BCs that contained our TFs of interest and marker TFs (240 TFs) which comprised 25.8 % (29 270 out of 113 210) of expressed gene transcripts. This approach potentially reduced the computational complexity and time required for modelling GRNs relevant to our study. This sum expression matrix of 29 270 genes and 240 TF genes was used as input for GRN inference by Lemon-Tree mode I, mode II and Inferelator.

Lemon-Tree produced 828 tight clusters in step 1 from the input expression matrix. A higher number of clusters (828 vs. 141 BCs from QUBIC) suggested that Lemon-Tree clusters were relatively more discrete/smaller in comparison to QUBIC BCs. In step 2, two separate options/modes were utilized (see Materials and Methods and Fig. 1). In mode I, we utilized the 828 tight-clustered modules generated by Lemon-Tree (mode I) and in mode II, the 141 BCs produced by QUBIC (mode II). In mode I, 176 TFs were ranked as the top 1 % regulators, whereas in mode II, 92 TFs were ranked as top 1 % regulators [**see Supporting Information—Table S5**]. Score evaluation was performed for the top 1 % and randomly predicted regulators from both modes. In both cases, the minimum score for a top regulator (14.22, mode I and 12.13, mode II) was ~three times higher than the maximal score (4.99, mode I and 4.23, mode II) for a randomly assigned regulator (Fig. 4). This suggested that the scores for top regulators are greater than what could be expected by chance. Inferelator algorithm predicted 132 TFs as potential regulators and five predicted groups [**see Supporting Information—Table S5**]. Comparison of 176, 92 and 132 TFs predicted as regulators respectively, by Lemon-Tree mode I, mode II and Inferelator, revealed that 56 TFs (~27 %) were predicted by all three different modes (Fig. 5A). We ranked these common 56 TFs as high-confidence TF regulators. In addition, ~62 % of the TFs predicted as a regulator by Lemon-Tree mode I were also identified as regulators by Lemon-tree mode II and/or Inferelator [**see Supporting Information—Fig. S1A**].

Furthermore, a total of 113 668 non-redundant TF–target regulatory interactions were predicted by all three modes (Lemon-Tree mode I—26 012, mode II—95 845 and Inferelator—3287) [**see Supporting Information—Table S6**]. A higher number of regulatory interactions in Lemon-Tree mode II is likely due to larger BCs produced by QUBIC. There was relatively smaller overlap among the three modes [**see Supporting Information—Fig. S1B**]. We evaluated whether known LR and nodule marker TFs were predicted as regulators as a measure of successful TF prediction by the three different modes. Soybean orthologs of LR marker TFs, *LRP1* (Glyma14g03900), *ARF5* (Glyma14g40540), *CRF2* (Glyma08g02460) and *TMO7* (Glyma04g34080 and Glyma06g20400), were predicted as regulators by all three inference modes. Additional orthologs of *ARF5* (Glyma17g37580) and *CRF2* (Glyma05g37120) were predicted as regulators by Lemon-Tree mode I and II. However, orthologs of *GATA23* (Glyma03g39220, Glyma19g41780) and *LRP1* (Glyma02g44860, Glyma07g35780) were not identified as regulators

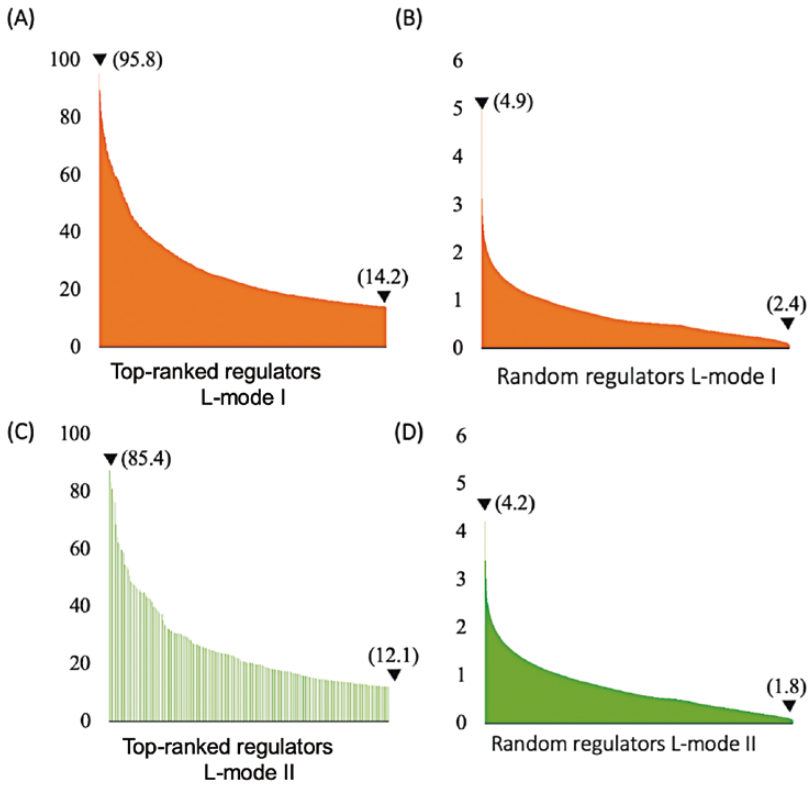

Figure 4. Distribution of Lemon-Tree scores of top-ranked and random regulators. Histogram shows the distribution of scores for top-ranked (A and C) and randomly (B and D) assigned regulators from Lemon-Tree mode I (orange) and mode II (green) workflows. Arrows indicate the minimum and maximum scores from each category with values in parenthesis.

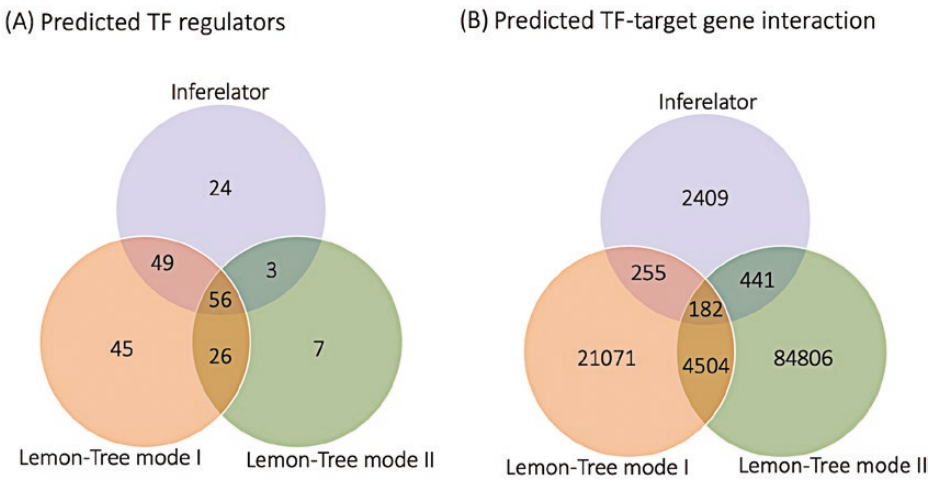

Figure 5. Overlap and differences of predicted (A) regulator TFs and (B) regulatory interactions between TFs and their target genes by three different GRN inference workflows. Numbers in centre indicate the number of potential regulators (in A) and interactions (in B) recovered by all the three different workflows.

by any of the modes. These four genes were not enriched in LR tissues [**see Supporting Information—Table S2**] and this likely why they were not predicted as a regulator in this data set. Prediction of four of the five LR-associated markers correctly as regulators by all three modes suggested that the workflow was reliable and would be useful in predicting previously unknown regulators of nodule development.

A number of TFs were demonstrated to play a crucial role in nodule development through genetic evidence from model legumes (Udvardi *et al.* 2007; Magne *et al.* 2018). These include NODULE INCEPTION (NIN; RWP-RK family (Schauser *et al.* 1999)), NODULATION SIGNALING PATHWAY1 and 2 (NSP1 and NSP2; GRAS domain proteins), Nuclear Factor Y (NF-YA1; (Battaglia *et al.* 2014)), *E*thylene Response Factors *R*equired for *N*odulation (ERN1 and ERN2; AP2/ERF family; (Baudin *et al.* 2015)) and CYCLOPS (coiled-coil domain protein) (Heckmann *et al.* 2006; Heckmann *et al.* 2011; Hayashi *et al.* 2012; Singh *et al.* 2014). In addition, an MYB TF that interacts with NSP2, an ARID domain protein that interacts with SymRK, a bHLH and a set of HD-ZIP IIIs involved in nodule vascular development, and a C2H2 Zn finger TF involved in bacteroid development are also known (Zhu *et al.* 2008). A potential soybean ortholog of *NIN*, Glyma02g48080 (Hayashi *et al.* 2012), belonging to orthogroup OGEF1237 was predicted as a regulator by Lemon-Tree mode I. Only one other *NIN*-like gene in this orthogroup (Glyma04g00210) was included in our list of input TFs based on expression enrichment in nodules, but was not predicted as a regulator by any mode. Two other *NIN*-like genes outside of this orthogroup (Glyma12g05390 and Glyma01g36360) were predicted to be regulators by Lemon-Tree modes I and II. Nodule-enriched *NF-YA*s (Glyma02g35190 and Glyma10g10240) were identified as regulators by Lemon-Tree mode I and Inferelator. In *Lotus japonicus*, two Nuclear Factor-Y (NF-Y) subunit genes, *LjNF-YA1* and *LjNF-YB1*, were identified as transcriptional targets of NIN (Soyano *et al.* 2013). In agreement, our analysis predicted that one of the soybean NIN-like genes, Glyma12g05390, regulates *NF-YA1* (Glyma10g10240; Lemon-Tree mode II) and the other *NIN*-like gene, Glyma01g36360, regulates *NF-YA2* (Glyma02g35190; Lemon-Tree mode I; **see Supporting Information—Table S5**).

Two potential orthologs of *LjERN1* (Glyma02g08020 and Glyma19g29000) were predicted as regulators by Lemon-Tree modes I and II. Among the major nodulation TFs, only *NSP1* was not predicted to be a regulator by our GRN workflow. In summary, the workflow correctly predicted known nodulation and LR TFs including the expected relationships between NIN, NF-YA and ERN1.

Putative PPIs identified in root lateral organ-related GRNs

Co-expressed and co-regulated genes have a higher likelihood of having an indirect functional interaction or direct physical interaction (Xulvi-Brunet and Li 2010). Many TFs form a complex with other proteins for proper molecular and cellular activity. Protein–protein interactions are the physical interactions between two or more proteins which form the crux of a functional protein complex formation (Barabasi and Oltvai 2004). To evaluate if potential regulators identified by us undergo PPIs with other co-regulated proteins, we compared all 113 668 unique TF–target pairs against verified and/or predicted PPIs reported in the STRING database (see Materials and Methods for details). We identified 843 potential interactions among 69 TFs with PPI confidence scores ranging from 150 to 995 [**see Supporting Information—Fig. S2**; **Table S7**]. The high scorer (>800) PPIs were observed from Lemon-Tree mode II run. It was previously suggested that a score of <800 was probably false positives that originated from prediction methods (Isik *et al.* 2015). Also, the maximum number (~64 %) of PPIs was identified by Lemon-Tree mode II, while only four PPIs were predicted by all three modes [**see Supporting Information—Fig. S1C**]. A likely explanation is the comparatively bigger BCs in this mode generated by QUBIC. While overall, in comparison to all predicted interactions by each mode independently, Inferelator had a greater frequency (2 %) of interactions in PPI, i.e. out of total predicted 3288, 61 were observed in PPI, followed by Lemon-Tree mode I (1 %) and then mode II (0.65 %). Two ARF5 LR markers Glyma14g40540 and Glyma17g37580 were predicted to interact with an AUX/IAA protein (Glyma13g43050; PPI score 980) and ATHB14-like homeodomain TF (Glyma15g13640; PPI score 530) present in GRNs predicted by Lemon-Tree mode I and Inferelator, respectively. Glyma13g43050 is an ortholog of Arabidopsis IAA28 that has been demonstrated to interact with AtARF5 (De Rybel *et al.* 2010), and this regulatory module plays a key role in LR development (Rogg *et al.* 2001).

High-confidence TF regulators and their GRNs associated with root lateral organ development in soybean

To determine high-confidence regulatory interactions and build a consensus GRN, we evaluated if TF–target pairs were conserved across all three modes of GRN prediction (Lemon-Tree modes I, II and Inferelator). Results showed that 182 regulatory relationships (that included 21 TFs) were commonly predicted by all three modes (Fig. 5B; **see Supporting Information—Table S8**). Therefore, for 38 % of the TFs predicted as high-confidence regulators (21 of 56), common target genes were also predicted independently by all three modes. These 21 TFs formed independent GRNs with their co-regulated target genes (Fig. 6). We ranked the consensus interactions by computing the average of the normalized score given by all three GRN inference modes (ranged from min = 0.19, max = 0.88; see Materials and Methods for details). Table 1 lists the 21 TFs, their annotation and enrichment in root lateral organs. Supporting Information—**Table S8** lists the score for all high-confidence TF–target pairs predicted by all three modes (Lemon-Tree mode I, Lemon-Tree mode II and Inferelator). Based on the expression of the TF regulator and their predicted target (Fig. 7), we categorized GRN enriched in specific lateral organ tissues.

TF regulators annotated as *AP2*; *ANT* (AINTEGUMENTA), transcriptional factor B3 family protein, *AtGRF5* (Growth-Regulating Factor 5), *C3H*, *AtbZIP52* (*Arabidopsis thaliana* basic leucine zipper 52)-like, *PC-MYB1* and *SHR* (Short Root) appear to co-regulate GRN modules during early nodule (EN) development. Transcription factor regulators annotated as *GRAS*; *SCARECROW-LIKE 6* (*SCL6*), *LOB DOMAIN-CONTAINING PROTEIN 41* (*LBD41*), AP2 domain-containing transcription factor *TINY*, *NUTCRACKER* (*NUC*); nucleic acid binding, *bZIP5*, *FER*-Like Regulator Of Iron Uptake (*FRU*), *RESPONSE REGULATOR* (*RR18*) and two unknown TF proteins appear to co-regulate GRN modules late during nodule (MN) development. Interestingly, four PPIs (out of 843 in PPI network) were also commonly predicted by all three GRN inference networks in our study for LBD41 and FRU in MNs [**see Supporting Information—Table**

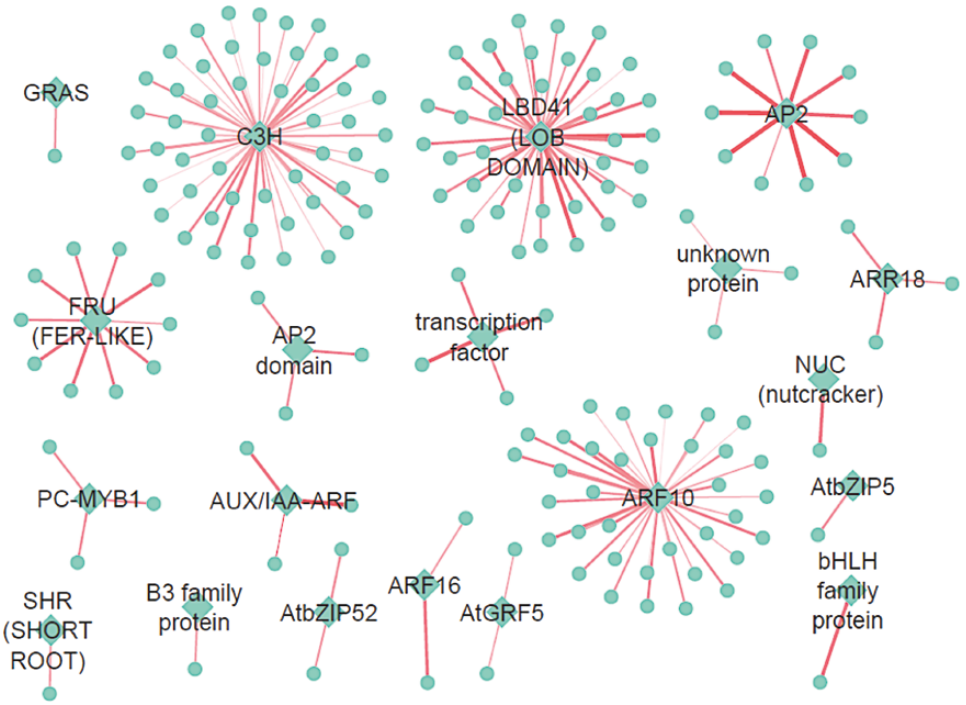

Figure 6. Consensus co-regulatory interactions predicted and recovered by three different GRN inference workflows. Nodes in diamond denote regulator TFs and circles denote predicted target genes. Edges denote the normalized score of interaction calculated by all three workflows. Broader the edges, higher the normalized interaction score.

Table 1. List of TFs predicted as regulator by all three GRN inference methods used in our study.

21 TFs IDs	TF annotation	Enrichment (log, fold change) in each organ			
		EN	MN	ELR	YLR
Glyma03g27050	AP2 domain-containing protein (TINY)		2.32		
Glyma17g08380	ARR18 (RESPONSE REGULATOR 18)		2.96		
Glyma11g04920	AtbZIP5 (basic leucine-zipper 5)		2.74		
Glyma13g39650	FRU (FER-LIKE REGULATOR OF IRON UPTAKE)		1.8	−1.74	−3.04
Glyma03g03760	GRAS TF; scarecrow-like 6 (SCL6)		2.29	1.22	
Glyma19g06280	LBD41 (LOB DOMAIN-CONTAINING PROTEIN 41)		1.19		
Glyma06g44080	NUC (nutcracker)		1.53		
Glyma03g34730	Putative transcription factor		2.49		
Glyma01g32130	Unknown protein		2.45	−0.87	
Glyma06g05170	AP2; ANT (AINTEGUMENTA)	1.42	−2.23		
Glyma09g07990	AtGRF5 (GROWTH-REGULATING FACTOR 5)	3.34			
Glyma02g40400	Transcriptional factor B3 family protein	2.76			
Glyma14g38460	AtbZIP52 (basic leucine zipper 52)	1.51		1.25	
Glyma16g01296	C3H	2.01		2.17	
Glyma05g22460	SHR (SHORT ROOT)	1.78		1.55	
Glyma06g08660	PC-MYB1	1.4		1.42	
Glyma11g37130	NFYC	3.57		1.49	
Glyma11g20490	ARF10 (AUXIN RESPONSE FACTOR 10)			1.91	2.7
Glyma06g20400	bHLH family protein (TMO7 ortholog)			2.35	
Glyma10g06080	ARF16 (AUXIN RESPONSE FACTOR 16)				−1.62
Glyma19g36571	ARF protein (AUX/IAA-ARF)		−0.77		

S8; **Fig. S2B**], which suggests both transcriptional and post-translational regulatory relationships between the TF–target pairs. It should however be noted that STRING-DB data includes proteins that do not physically interact with the TF. ARF16 and AUX/IAA-ARF proteins were predicted to regulate ELR development, whereas TMO7 and ARF10 (Auxin Response Factor 10) were predicted to co-regulate GRNs during YLR development in soybean.

DISCUSSION

In spite of the economic and environmental importance of biological nitrogen fixation in nodule in soybean, there is still an unanswered question of what key TFs regulate the underlying GRNs in nodules (Udvardi *et al.* 2007). We developed a robust computational framework for GRN construction using genome-scale gene expression data. Specifically, this framework integrates genomic and transcriptomic data to infer the key regulators and GRN associated with nodule development in soybean. The predicted networks consistently included experimentally verified genes, demonstrating the ability of our framework to reveal significant, potentially important GRNs. With a broader impact, the framework can be used as a template for constructing GRNs to address any biological question of interest in any species.

To reduce the computational complexity and make the predicted regulator TFs and GRNs relevant to our biological question, a biclustering method and a regulatory network inference tool were used, where their parameters were optimized via several iterations for data analysis and modelling. Among existing GRN inference algorithms, Lemon-Tree and Inferelator were successfully applied in different biological questions due to their valued feature, i.e. top regulator and top-ranked regulatory target prediction (Michoel *et al.* 2009; Vermeirssen *et al.* 2009; Dolinski and Troyanskaya 2015; Finkle *et al.* 2018). Lemon-Tree detects regulatory modules and regulators from gene

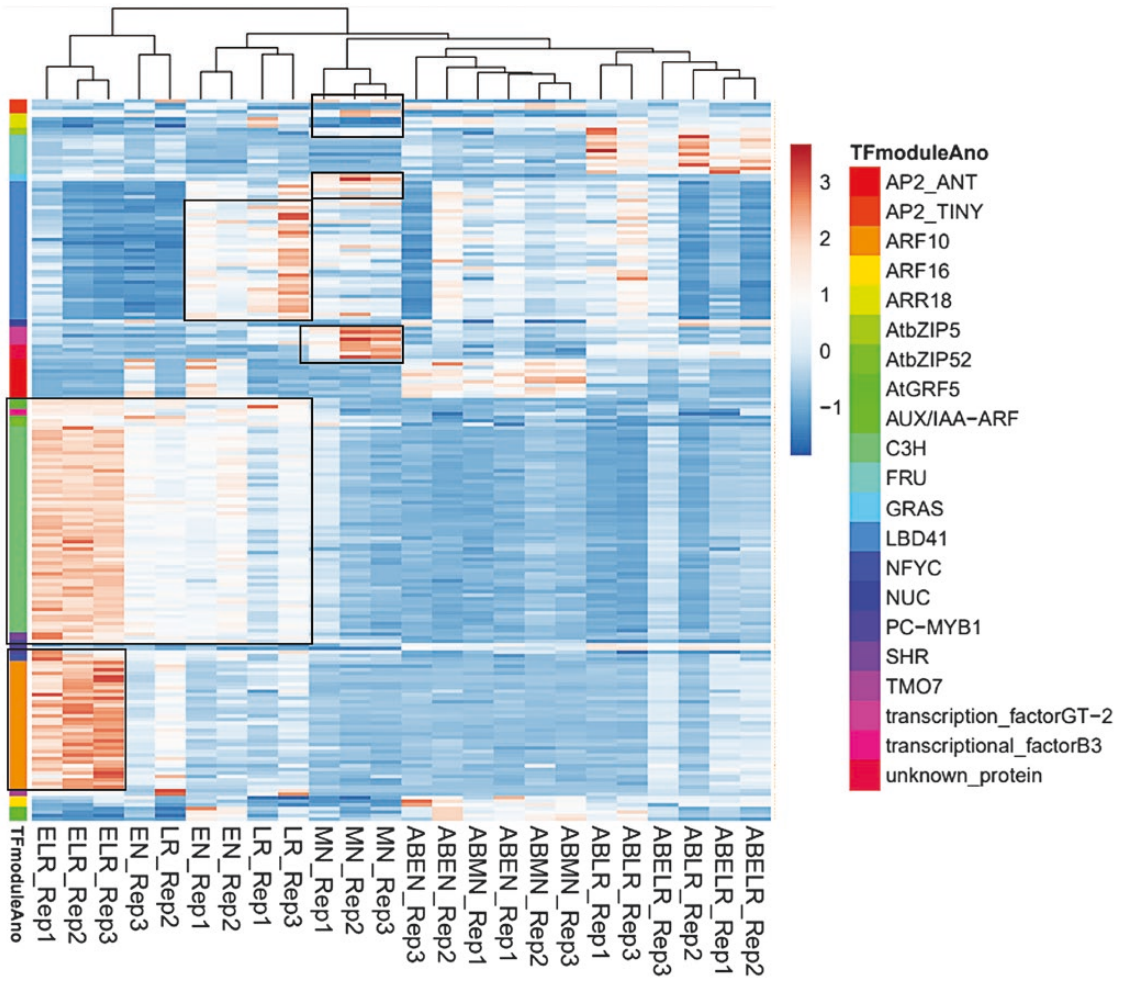

Figure 7. Heat map showing normalized expression from varied samples of root lateral organ development in soybean for regulator TFs and their co-regulatory target genes (TF modules) in consensus networks predicted by all three GRN inference workflows. Row annotation for 21 regulator TFs and their co-regulatory partners are shown in different colours. Co-expressed TFs and their co-regulatory target genes in specific tissues are highlighted in black box.

expression data using probabilistic graphical models (Bonnet *et al.* 2015). Whereas, Inferelator learns a system of ordinary differential equations using the Bayesian Best Subset Regression that describes the rate of change in transcription of each gene or gene cluster, as a function of TFs. It has been shown that predictions made by the Inferelator are highly accurate for top-ranking predictions. Stochastic Lemon-Tree and Inferelator perform better if the transcriptional program can be inferred from a pre-specified list of regulators rather than from a full gene list, because erroneous interactions with non-regulators will be eliminated *a priori* (De Smet and Marchal 2010). So, we took the differentially expressed TFs and predefined marker TFs with a known role in nodule and LRs to infer GRN.

Novel regulators of nodule development

We distinguished organ (LR/nodule) and/or developmental stage-specific (early/mature) consensus GRNs based on organ-specific enrichment of the TFs, their differential expression and expression pattern of their co-regulated genes in our transcriptome data. In addition, we also employed comparative genomics and information from public tissue atlas and transcriptome data. The analysis correctly predicted four of the five LR regulators with high confidence and known nodulation TFs including the expected relationships between them. For example, the phylogenetic analysis suggested that ERN2 may not be present in legumes that form determinate nodules such as soybean, *L. japonicus* or common bean (Kawaharada *et al.* 2017). The expression of *ERN1* and *ERN2* are under the control of NIN and NF-YA in *Medicago*, a legume that forms indeterminate nodules. In fact, NF-YA binds the promoter of *ERN1* directly regulating its expression in *Medicago*. However, *ERN1* expression does not appear to be regulated by NIN or NF-YA in *L. japonicus* as its expression is not altered in *NIN* or *NF-YA* loss of function mutants. Our GRN prediction also did not identify *ERN1* as a target of NF-YA or NIN in soybean. *ERN1* is directly regulated by CYCLOPS in *L. japonicus*. NSP2 and CYCLOPS were not included in the input TF list due to no nodule-specific enrichment and/or incorrect annotation. The inclusion of CYCLOPS in future analyses might reveal regulatory relationships between *ERN1* and CYCLOPS in soybean. It remains to be seen if this is conserved among other determinate nodule forming legumes including soybean. Given the reliability of the workflow in accurately predicting known TFs, we discuss previously unknown regulators of nodule development predicted by the workflow.

With the goal of identifying high-confidence TF–target pairs operating during nodule development, we pre-selected a set of 294 TFs, which were specifically enriched in EN, and MN tissues in our data set as possible regulators. We identified 17 high-confidence TFs among these as predicted by all three modes. Three TFs were predicted to drive GRNs specifically associated with ENs, which are soybean orthologues of Arabidopsis *AINTEGUMENTA* (*ANT*; At4g37750), *AP2/B3* domain transcriptional factor (At5g58280) and *GROWTH-REGULATING FACTOR 5* (*GRF5*). All three genes are associated with sites of cell proliferation in Arabidopsis. While GRF5 plays a role in cell proliferation during leaf primordia formation and leaf development, ANT is crucial for flower development. At5g58280 shows the highest expression level in the shoot apex, particularly in the central zone. Indeed, it is likely that the soybean TFs associated with EN GRNs direct cell proliferation during early nodule development. Seven other TFs belonging to C3H, bZIP, MYB1, NF-YC and SHR families predicted to co-regulate GRN modules in ENs which also happened to be enriched in ELRs (Table 1). These GRNs might act during the initiation of both these lateral organs. Soybean ANT ortholog was the regulator with the highest score in our analysis (0.8) and was predicted to co-regulate 10 target genes specifically in ENs. Its targets included *ATCSLA09*, *ALDH2C4*, *GCL1* (*GCR2*-LIKE 1), *AAP6* and auxin-responsive protein. A maximum of 51 co-regulated target genes were predicted for a C3H TF regulator (enriched in both EN and ELR) by all three modes. Most of the target genes such as glycosyl hydrolase family protein, *CYCA1;1* (Cyclin A1;1), zinc finger (C3HC4-type RING finger), *CDKB1*, *CMT3* (chromomethylase 3); DNA (cytosine-5-)-methyltransferase, calmodulin-binding protein-related, *CYC1BAT*; cyclin-dependent protein kinase regulator, mitotic spindle checkpoint protein, putative (*MAD2*), *ATARP7* (Actin-Related Protein 7); structural constituent of cytoskeleton, kinesin motor protein-related and *CDC20.1*; signal transducer were high scoring target genes.

Gene Ontology enrichment analysis of genes involved in EN and EN-ELR GRNs showed significant enrichment of regulation of a cell cycle, movement of a cell or subcellular component, microtubule-based movement, cell division and cell cycle biological process [**see Supporting Information—Fig. S4**]. This is consistent with biological processes known to occur early during lateral organ development. *Cis*-regulatory motif GACCGTTA was enriched in the EN-related GRN regulated by a Myb/SANT TF.

Similarly, MN-GRN involved in MN development was enriched with meristem initiation and growth. Nine TF regulators annotated as being similar to GRAS (SCL6), LBD41, TINY, NUC bZIP5, FRU, RR18, Myb/SANT-like DNA binding protein and a SCREAM-like protein appear to co-regulate GRN modules late during nodule (MN) development. Among these TFs, LBD41 had the highest score (0.77). LBD41 was predicted to co-regulate 38 target genes, among which *PDC2* (pyruvate decarboxylase-2) had the highest normalized score (0.7). Other targets included *PSAT*, *SIMILAR TO RCD ONE 2* (*SRO2*), *MATERNAL EFFECT EMBRYO ARREST 14* (*MEE14*), AN1-like Zinc finger, *SNF2*, trehalose-6-phosphate phosphatase, hypoxia-responsive family protein, *bHLH*, wound-responsive family protein and an *ASPARTATE AMINOTRANSFERASE 1* (*ASP1*) with normalized score >0.5 (Fig. 7). Arabidopsis LBD41 is associated with hypoxia response and multiple targets predicted for the soybean ortholog of LBD41 in MN were also associated with hypoxia (Gasch *et al.* 2016). Nodule oxygen concentrations are highly regulated to enable the proper functioning of the oxygen-sensitive nitrogenase enzyme complex. It is tempting to suggest that soybean LBD41 might play a role in regulating the response to hypoxia in MN. The Arabidopsis orthologs of SCL-6, a key regulator in MN, play a role in shoot branching by regulating axillary bud development (Wang *et al.* 2010). We had previously suggested that nodules and shoot axillary meristems require a similar hormone balance during development. It is possible that some developmental pathways such as those regulated by SCL6 are shared between these organs. Similarly, the role of Arabidopsis NUTCRACKER protein required in periclinal cell divisions (Long *et al.* 2015), that of FRU in uptake of iron (Jakoby *et al.* 2004) and RR18 in positive regulating cytokinin activity (Veerabagu *et al.* 2012)

are all consistent with biological processes observed in MN tissues (Breakspear *et al.* 2014; Reid *et al.* 2017). Gene Ontology enrichment analysis for MN-GRN genes showed enrichment of specification of axis polarity, adaxial/abaxial axis specification, meristem initiation, meristem growth and regulation of meristem growth [**see Supporting Information—Fig. S4**]. While these processes are known to occur in MNs, TFs associated with these processes had not been identified previously. Genes involved in MN-GRN had significant enrichment (*P*-value ≤0.05 FDR) for *cis*-regulatory motifs GGGCCCAC, ACCG and TGTCGG in their upstream regulatory regions. These are likely to be regulated by TCP, AP2 and B3 TFs, respectively. The study has revealed potential TFs associated with different biological processes in nodule and LR development [**see Supporting Information—Fig. S4**]. Knowledge from this work supported by experimental validation in the future is expected to help determine key gene/TF targets for biotechnological strategies to optimize nodule formation and enhance nitrogen fixation.

DATA AVAILABILITY

Gene expression data used to construct gene regulatory networks are available in NCBI Gene Expression Omnibus (GEO), accession number GSE129509. Raw data files are available in NCBI's Sequence Read Archive (SRA) and can be accessed via links available at the GEO record URL: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE129509.

SUPPORTING INFORMATION

The following additional information is available in the online version of this article—

Figure S1. Venn diagrams outlining overlaps and differences in the outputs among the three different gene regulatory network inference workflows: (A) regulator transcription factor (TF) prediction, (B) identification of targets for predicted TFs and (C) protein–protein interactions.

Figure S2. A total of 843 STRING protein–protein interaction (PPI) predictions matched our co-regulatory expression prediction. The large network is shown in A and smaller discrete networks in B.

Figure S3. Gene Ontology biological process enrichment of target genes in consensus 182 co-regulatory gene network interactions predicted for root lateral organ development in soybean.

Figure S4. Summary of lateral organ gene regulatory networks (GRNs) involving high-confidence transcription factor regulators predicted in this study and the biological processes enriched in those GRNs.

Table S1. List of transcription factor genes enriched in emerging and mature nodule tissues.

Table S2. Nodule (emerging nodules (ENs), mature nodules (MNs)) and lateral root (emerging lateral roots (ELRs) and young lateral roots (YLRs)) marker genes used to determine clustering parameters.

Table S3. Normalized expression (read counts) from Soybean Gene Atlas (Sevrin *et al.* 2010) for genes in ENOD40- and NSP1-containing biclusters. Cell(s) with the highest value in each row (gene) are highlighted.

Table S4. Expression patterns of potential Arabidopsis orthologs of soybean genes in TMO7 and LRP1 containing biclusters during lateral root initiation (data from Lewis *et al.* 2013).

Table S5. High-confidence regulators predicted by each workflow.

Table S6. Unique regulator-target prediction by all three workflows.

Table S7. Protein–protein interactions between regulator–target pairs as predicted by STRING database.

Table S8. List of 21 high-scoring regulators and their target genes forming lateral organ-specific gene regulatory networks. Transcription factor–target pairs highlighted in orange were predicted to be involved in protein–protein interactions by STRING-DB.

Data Set S1. Bicluster output from QUBIC.

ACKNOWLEDGEMENTS

We gratefully acknowledge the South Dakota State University for providing their high-performance computing clusters for data analysis and and for the technical support from their research information technology team (Dr. Brian Moore).

FUNDING

This work was supported by grant awards from the National Science Foundation/EPSCoR Cooperative Agreements (IIA-1355423 and 1849206); National Science Foundation's Plant Genome Research Program (IOS-1350189 to S.S.); United States Department of Agriculture National Institute of Food and Agriculture (2016-67014-24589 to S.S.); and SD Agricultural Experiment Station (SD00H543-15). J.K. was an NSF-REU fellow supported by award no. OAC-1559978.

CONFLICT OF INTEREST

None declared.

LITERATURE CITED

Adhikari S, Damodaran S, Subramanian S. 2019. Lateral root and nodule transcriptomes of soybean. *Data* **4**:64.

Atkinson JA, Rasmussen A, Traini R, Voß U, Sturrock C, Mooney SJ, Wells DM, Bennett MJ. 2014. Branching out in roots: uncovering form, function, and regulation. *Plant Physiology* **166**:538–550.

Baitaluk M, Kozhenkov S, Ponomarenko J. 2012. An integrative approach to inferring gene regulatory module networks. *PLoS One* **7**:e52836.

Barabási AL, Oltvai ZN. 2004. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics* **5**:101–113.

Battaglia M, Rípodas C, Clúa J, Baudin M, Aguilar OM, Niebel A, Zanetti ME, Blanco FA. 2014. A nuclear factor Y interacting protein of the GRAS family is required for nodule organogenesis, infection thread progression, and lateral root growth. *Plant Physiology* **164**:1430–1442.

Baudin M, Laloum T, Lepage A, Rípodas C, Ariel F, Frances L., Crespi M, Gamas P, Blanco FA, Zanetti ME, Carvalho-Niebel FD, Niebel A. 2015. A phylogenetically conserved group of nuclear factor-Y transcription factors interact to control nodulation in legumes. *Plant Physiology* **169**:2761–2773.

Benková E, Bielach A. 2010. Lateral root organogenesis—from cell to organ. *Current Opinion in Plant Biology* **13**:677–683.

Blais A, Dynlacht BD. 2005. Constructing transcriptional regulatory networks. *Genes & Development* **19**:1499–1511.

Bonneau R, Reiss DJ, Shannon P, Facciotti M, Hood L, Baliga NS, Thorsson V. 2006. The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets *de novo*. *Genome Biology* **7**:R36.

Bonnet E, Calzone L, Michoel T. 2015. Integrative multi-omics module network inference with Lemon-Tree. *PLoS Computational Biology* **11**:e1003983.

Breakspear A, Liu C, Roy S, Stacey N, Rogers C, Trick M, Morieri G, Mysore KS, Wen J, Oldroyd GED, Downie JA, Murray JD. 2014. The root hair 'infectome' of *Medicago truncatula* uncovers changes in cell cycle genes and reveals a requirement for auxin signaling in rhizobial infection. *The Plant Cell* **26**:4680–4701.

Chaturvedi I, Sakharkar MK, Rajapakse JC. 2007. Validation of gene regulatory networks from protein–protein interaction data: application to cell-cycle regulation. In: Rajapakse JC, Schmidt B, Volkert G, eds. *Pattern recognition in bioinformatics*. Berlin, Heidelberg: Springer, 300–310.

Ciofani M, Madar A, Galan C, Sellars M, Mace K, Pauli F, Agarwal A, Huang W, Parkhurst CN, Muratet M, Newberry KM, Meadows S, Greenfield A, Yang Y, Jain P, Kirigin FK, Birchmeier C, Wagner EF, Murphy KM, Myers RM, Bonneau R, Littman DR. 2012. A validated regulatory network for th17 cell specification. *Cell* **151**:289–303.

Collier R, Tegeder M. 2012. Soybean ureide transporters play a critical role in nodule development, function and nitrogen export. *The Plant Journal* **72** (3):355–67.

De R, Bert V, Vassileva BP, Demeulenaere M, Grunewald W, Audenaert D, Van Campenhout J, Overvoorde P, Jansen L, Vanneste S, Möller B, Wilson M, Holman T, Van Isterdael G, Brunoud G, Vuylsteke M, Vernoux T, De Veylder L, Inzé D, Weijers D, Bennett MJ, Beeckman T. 2010. A novel Aux/IAA28 signaling cascade activates GATA23-dependent specification of lateral root founder cell identity. *Current Biology* **20**:1697–1706.

De Smet R, Marchal K. 2010. Advantages and limitations of current network inference methods. *Nature Reviews Microbiology* **8**:717–729.

Dolinski K, Troyanskaya OG. 2015. Implications of big data for cell biology. *Molecular Biology of the Cell* **26**:2575–2578.

Du Y, Scheres B. 2018. Lateral root formation and the multiple roles of auxin. *Journal of Experimental Botany* **69**:155–167.

Eeckhoute J, Métivier R, Salbert G. 2009. Defining specificity of transcription factor regulatory activities. *Journal of Cell Science* **122**:4027–4034.

Finkle JD, Wu JJ, Bagheri N. 2018. Windowed granger causal inference strategy improves discovery of gene regulatory networks. *Proceedings of the National Academy of Sciences of the United States of America* **115**:2252–2257.

Gasch P, Fundinger M, Müller JT, Lee T, Bailey-Serres J, Mustroph A. 2016. Redundant ERF-VII transcription factors bind to an evolutionarily conserved *cis*-motif to regulate hypoxia-responsive gene expression in arabidopsis. *The Plant Cell* **28**:160–180.

Greenfield A, Madar A, Ostrer H, Bonneau R. 2010. DREAM4: combining genetic and dynamic information to identify biological networks and dynamical models. *PLoS One* **5**:e13397.

Guan D, Shao J, Zhao Z, Wang P, Qin J, Deng Y, Boheler KR, Wang J, Yan B. 2014. PTHGRN: unraveling post-translational hierarchical gene regulatory networks using PPI, ChIP-Seq and gene expression data. *Nucleic Acids Research* **42**:W130–W136.

Hayashi S, Reid DE, Lorenc MT, Stiller J, Edwards D, Gresshoff PM, Ferguson BJ. 2012. Transient nod factor-dependent gene expression in the nodulation-competent zone of soybean (*Glycine max* [L.] Merr.) roots. *Plant Biotechnology Journal* **10**:995–1010.

Heckmann AB, Lombardo F, Miwa H, Perry JA, Bunnewell S, Parniske M, Wang TL, Downie JA. 2006. *Lotus japonicus* nodulation requires two GRAS domain regulators, one of which is functionally conserved in a non-legume. *Plant Physiology* **142**:1739–1750.

Heckmann AB, Sandal N, Bek AS, Madsen LH, Jurkiewicz A, Nielsen MW, Tirichine L, Stougaard J. 2011. Cytokinin induction of root nodule primordia in *Lotus japonicus* is regulated by a mechanism operating in the root cortex. *Molecular Plant–Microbe Interactions* **24**:1385–1395.

Isik Z, Baldow C, Cannistraci CV, Schroeder M. 2015. Drug target prioritization by perturbed gene expression and network information. *Scientific Reports* **5**:17417.

Jakoby M, Wang HY, Reidt W, Weisshaar B, Bauer P. 2004. FRU (BHLH029) is required for induction of iron mobilization genes in *Arabidopsis thaliana*. *FEBS Letters* **577**:528–534.

Jin, J, Zhang H, Kong L, Gao G, Luo J. 2014. PlantTFDB 3.0: a portal for the functional and evolutionary study of plant transcription factors. *Nucleic Acids Research* **42**:D1182–D1187.

Joshi, A, De Smet R, Marchal K, Van de Peer Y, Michoel T. 2009. Module networks revisited: computational assessment and prioritization of model predictions. *Bioinformatics* **25**:490–496.

Kaufmann K, Pajoro A, Angenent GC. 2010. Regulation of transcription in plants: mechanisms controlling developmental switches. *Nature Reviews Genetics* **11**:830–842.

Kawaharada Y, James EK, Kelly S, Sandal N, Stougaard J. 2017. The ethylene responsive factor required for nodulation 1 (ERN1) transcription factor is required for infection-thread formation in *Lotus japonicus*. *Molecular Plant–Microbe Interactions* **30**:194–204.

Kim Y-A, Przytycka TM. 2013. Bridging the gap between genotype and phenotype via network approaches. *Frontiers in Genetics* **3**:227.

Lewis DR, Olex AL, Lundy SR, Turkett WH, Fetrow JS, Muday GK. 2013. A kinetic analysis of the auxin transcriptome reveals cell wall remodeling proteins that modulate lateral root development in *Arabidopsis*. *The Plant Cell* **25**:3329–3346.

Li, Y, Jackson SA. 2016. Crowdsourcing the nodulation gene network discovery environment. *BMC Bioinformatics* **17**:223.

Li G, Ma Q, Tang H, Paterson AH, Xu Y. 2009. QUBIC: a qualitative biclustering algorithm for analyses of gene expression data. *Nucleic Acids Research* **37**:e101.

Libault, M, Farmer A, Brechenmacher L, Drnevich J, Langley RJ, Bilgin DD, Radwan O, Neece DJ, Clough SJ, May GD, Stacey G. 2010*a*. Complete transcriptome of the soybean root hair cell, a single-cell model, and its alteration in response to *Bradyrhizobium japonicum* infection. *Plant Physiology* **152**:541–552.

Libault M, Farmer A, Joshi T, Takahashi K, Langley RJ, Franklin LD, He J, Xu D, May G, Stacey G. 2010*b*. An integrated transcriptome atlas of the crop model *Glycine max*, and its use in comparative analyses in plants. *The Plant Journal: for Cell and Molecular Biology* **63**:86–99.

Long, Y, Smet W, Cruz-Ramírez A, Castelijns B, de Jonge W, Mähönen AP, Bouchet BP, Perez GS, Akhmanova A, Scheres B, Blilou I. 2015. Arabidopsis BIRD zinc finger proteins jointly stabilize tissue boundaries by confining the cell fate regulator SHORT-ROOT and contributing to fate specification. *The Plant Cell* **27**:1185–1199.

Magne K, Couzigou JM, Schiessl K, Liu S, George J, Zhukov V, Sahl L, Boyer F, Iantcheva A, Mysore KS, Wen J, Citerne S, Oldroyd GED, Ratet P. 2018. MtNODULE ROOT1 and MtNODULE ROOT2 are essential for indeterminate nodule identity. *Plant Physiology* **178**:295–316.

Michoel T, De Smet R, Joshi A, Van de Peer Y, Marchal K. 2009. Comparative analysis of module-based versus direct methods for reverse-engineering transcriptional regulatory networks. *BMC Systems Biology* **3**:49.

Petricka JJ, Benfey PN. 2011. Reconstructing regulatory network transitions. *Trends in Cell Biology* **21**:442–451.

Reid D, Nadzieja M, Novák O, Heckmann AB, Sandal N, Stougaard J. 2017. Cytokinin biosynthesis promotes cortical cell responses during nodule development. *Plant Physiology* **175**:361–375.

Rivas JDL, Fontanillo C. 2010. Protein–protein interactions essentials: key concepts to building and analyzing interactome networks. *PLoS Computational Biology* **6**:e1000807.

Rogg LE, Lasswell J, Bartel B. 2001. A gain-of-function mutation in IAA28 suppresses lateral root development. *The Plant Cell* **13**:465–480.

Roy S, Liu W, Nandety RS, Crook A, Mysore KS, Pislariu CI, Frugoli J, Dickstein R, Udvardi MK. 2020. Celebrating 20 years of genetic discoveries in legume nodulation and symbiotic nitrogen fixation. *The Plant Cell* **32**:15–41.

Schauser L, Roussis A, Stiller J, Stougaard J. 1999. A plant regulator controlling development of symbiotic root nodules. *Nature* **402**:191.

Severin AJ, Woody JL, Bolon Y-T, Joseph B, Diers BW, Farmer AD, Muehlbauer GJ, Nelson RT, Grant D, Specht JE, Graham MA, Cannon SB, May GD, Vance CP & Shoemaker RC. 2010. RNA-Seq Atlas of *Glycine max*: a guide to the soybean transcriptome. *BMC Plant Biology* **10**:160.

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 Research* **13**:2498–2504.

Singh S, Katja K, Jayne L, Marion C, Martin P. 2014. CYCLOPS, a DNA-binding transcriptional activator, orchestrates symbiotic root nodule development. *Cell Host & Microbe* **15**:139–152.

Soyano T, Kouchi H, Hirota A, Hayashi M. 2013. Nodule inception directly targets NF-Y subunit genes to regulate essential processes of root nodule development in *Lotus japonicus*. *PLoS Genetics* **9**:e1003352.

Sun N, Zhao H. 2009. Reconstructing transcriptional regulatory networks through genomics data. *Statistical Methods in Medical Research* **18**:595–617.

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, Jensen LJ, von Mering C. 2015. STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Research* **43**:D447–D452.

Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, Santos A, Doncheva NT, Roth A, Bork P, Jensen LJ, von Mering C. 2017. The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Research* **45**:D362–D368.

Udvardi MK, Kakar K, Wandrey M, Montanari O, Murray JD, Andriankaja A, Zhang JY, Benedito VA, Hofer JMI, Chueng F. 2007. Update on legume transcription factors legume transcription factors : global regulators of plant development and response to the environment. *Plant Physiology* **144**:583–549.

Veerabagu, M, Elgass K, Kirchler T, Huppenberger T, Harter K, Chaban C, Mira-Rodado V. 2012. The arabidopsis b-type response regulator 18 homomerizes and positively regulates cytokinin responses. *The Plant Journal* **72**:721–731.

Vermeirssen V, Joshi A, Michoel T, Bonnet E, Casneuf T, Van de Peer Y. 2009. Transcription regulatory networks in caenorhabditis elegans inferred through reverse-engineering of gene expression profiles constitute biological hypotheses for metazoan development. *Molecular Biosystems* **5**:1817–1830.

Wang L, Mai YX, Zhang YC, Luo Q, Yang HQ. 2010. MicroRNA171c-targeted SCL6-II, SCL6-III, and SCL6-IV genes regulate shoot branching in arabidopsis. *Molecular Plant* **3**:794–806.

Xie J, Ma A, Zhang Y, Liu B, Cao S, Wang C, Xu J, Zhang C, Ma Q. 2019. QUBIC2: a novel and robust biclustering algorithm for analyses and interpretation of large-scale RNA-Seq data. *Bioinformatics* **36**:1143–1149.

Xulvi-Brunet R, Li H. 2010. Co-expression networks: graph properties and topological comparisons. *Bioinformatics* **26**:205–214.

Zhang Y, Xie J, Yang J, Fennell A, Zhang C, Ma Q. 2017. QUBIC: a bioconductor package for qualitative biclustering analysis of gene co-expression data. *Bioinformatics* **33**:450–452.

Zhu H, Chen T, Zhu M, Fang Q, Kang H, Hong Z, Zhang Z. 2008. A novel ARID DNA-binding protein interacts with SymRK and is expressed during early nodule development in *Lotus japonicus*. *Plant Physiology* **148**:337–347.

Zhu M, Dahmen JL, Stacey G, Cheng J. 2013. Predicting gene regulatory networks of soybean nodulation from RNA-Seq transcriptome data. *BMC Bioinformatics* **14**:278.