frontiers in Plant Science

ORIGINAL RESEARCH
published: 03 March 2022
doi: 10.3389/fpls.2022.831204

Large-Scale Integrative Analysis of Soybean Transcriptome Using an Unsupervised Autoencoder Model

*Lingtao Su[1], Chunhui Xu[2], Shuai Zeng[1], Li Su[2], Trupti Joshi[1,2,3], Gary Stacey[4] and Dong Xu[1,2]**

[1] Department of Electrical Engineering and Computer Science and Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO, United States, [2] Institute for Data Science and Informatics, Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO, United States, [3] Department of Health Management and Informatics and Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO, United States, [4] Division of Plant Sciences and Technology and Biochemistry Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO, United States

Plant tissues are distinguished by their gene expression patterns, which can help identify tissue-specific highly expressed genes and their differential functional modules. For this purpose, large-scale soybean transcriptome samples were collected and processed starting from raw sequencing reads in a uniform analysis pipeline. To address the gene expression heterogeneity in different tissues, we utilized an adversarial deconfounding autoencoder (AD-AE) model to map gene expressions into a latent space and adapted a standard unsupervised autoencoder (AE) model to help effectively extract meaningful biological signals from the noisy data. As a result, four groups of 1,743, 914, 2,107, and 1,451 genes were found highly expressed specifically in leaf, root, seed and nodule tissues, respectively. To obtain key transcription factors (TFs), hub genes and their functional modules in each tissue, we constructed tissue-specific gene regulatory networks (GRNs), and differential correlation networks by using corrected and compressed gene expression data. We validated our results from the literature and gene enrichment analysis, which confirmed many identified tissue-specific genes. Our study represents the largest gene expression analysis in soybean tissues to date. It provides valuable targets for tissue-specific research and helps uncover broader biological patterns. Code is publicly available with open source at https://github.com/LingtaoSu/SoyMeta.

Keywords: soybean, transcriptome analysis, deep learning, autoencoder, tissue-specific gene, gene regulatory network, functional module

OPEN ACCESS

Edited by:
Xiyin Wang,
Agricultural University of Hebei, China

Reviewed by:
Song Li,
Virginia Tech, United States
John Louis Van Hemert,
Corteva Agriscience™, United States

***Correspondence:**
Dong Xu
xudong@missouri.edu

Specialty section:
This article was submitted to Plant Bioinformatics, a section of the journal Frontiers in Plant Science

Received: *08 December 2021*
Accepted: *09 February 2022*
Published: *03 March 2022*

Citation:
Su L, Xu C, Zeng S, Su L, Joshi T, Stacey G and Xu D (2022) Large-Scale Integrative Analysis of Soybean Transcriptome Using an Unsupervised Autoencoder Model. Front. Plant Sci. 13:831204. doi: 10.3389/fpls.2022.831204

INTRODUCTION

The soybean is a valuable source of oil and protein for humans and livestock; it is also very important for soil fertility, given the symbiotic interaction with nitrogen-fixing rhizobia. The development of high-throughput gene expression quantification technologies, initially dominated by microarray platforms and later by RNA-Seq technologies, has contributed to the substantial rise in soybean transcriptome studies. The related data are well presented in several public data repositories, such as the SoyKB (Joshi et al., 2012, 2014, 2017), SoyBase (Grant et al., 2010;

Brown et al., 2021) and the most recently published Soybean Expression Atlas (Machado et al., 2020), covering thousands of gene expression data sets from various tissues, developmental stages and conditions. Typically, several related experiments studying the same tissue are available, providing a rich set of materials for integrative analysis via data pooling and mining. The increasing sample size also enhances the statistical power to obtain a more precise and robust estimate of molecular markers and reduces the noise effects and individual study biases.

Researchers have become increasingly interested in integrating their own data with publicly available data sets to achieve more accurate results and deeper biological understanding. For example, through a large-scale transcriptome meta-analysis, several hub genes involved in soybean oil accumulation processes were revealed in Qi et al. (2018), and a large number of differentially expressed genes (DEGs) related to soybean symbiotic nitrogen fixation were also identified (Yuan et al., 2017). In addition, some similar studies are available (Liu et al., 2015; Huang et al., 2018; Wang J. et al., 2019; Yi et al., 2019). However, most of these investigations explored only a few conditions or developmental stages. Such *ad hoc* approaches can overlook a myriad of interesting transcriptional patterns, which could otherwise be unraveled by integrative methods using a more comprehensive set of samples. Inspired by this, a global co-expression network analysis of 1,072 soybean microarray samples was conducted (Wu et al., 2019), which revealed a gene module that is likely involved in the evolution of nodulation in plants. Kim et al. (2017) constructed the SoyNet database using 734 microarrays and 290 RNA-seq samples. Moreover, by systematically analyzing 1,270 microarray samples generated with Affymetrix gene chips, a nodulation-related co-expression module was uncovered (Wu et al., 2019). More recently, researchers (Sun et al., 2020) identified key regulators and hub genes in each tissue by analyzing a genome-wide transcriptome dataset from eight tissues at three different seed development stages. To elucidate the dynamics of transcriptional regulation across the broad range of samples, tissues, and cultivars, 1,298 publicly available soybean transcriptome samples were collected and analyzed by Machado et al. (2020).

Properly integrating large-scale data sets can help increase statistical power, but expression profiles inherently contain variations introduced by noise, batch effects and conditions unrelated to the biological hypotheses. Although many batch effect adjustment methods have been proposed (Benito et al., 2004; Johnson et al., 2007; Sims et al., 2008; Luo et al., 2010; Xia et al., 2013), they typically cannot handle large-scale data integration (Haibe-Kains et al., 2013). Therefore, the integration of microarray and RNA-seq data sets continues to be a challenging problem (Lazar et al., 2013). However, the emergence of deep learning techniques provides a new perspective and opportunity to solve this problem. The unsupervised learning model can extract patterns from diverse and noisy data without assuming any statistical properties of the data, which makes it well suited for gene expression analysis (Du et al., 2019; Dincer et al., 2020; Li et al., 2020). For example, the adversarial deconfounding autoencoder (AD-AE) model (Dincer et al., 2020) can generate biologically informative expression embeddings that are both robust to confounders and generalizable. The AD-AE model uses an autoencoder network to capture the true signal and a complete adversary network to remove confounder variables for a noise-free and confounder-free representation. Considering the widespread noise in soybean gene expression datasets (Araujo et al., 2017; Cortijo et al., 2019), in this study, we adapted the AD-AE model to analyze collected soybean datasets, considering not only different data sources but also different sequencing platforms.

As more and more tissue-specific gene expression data become available for soybeans (Libault et al., 2010), another important aspect in the large-scale integrative analysis is detecting tissue-specific genes and constructing tissue-specific gene regulatory networks (GRNs). Several computational methods were developed to measure the tissue specificity of gene expression, such as the EE (Yu et al., 2006) in the database TiGER (Liu et al., 2008), the SPM used in the database TiSGeD (Xiao et al., 2010), and the Gini coefficient (Ceriani and Verme, 2012). However, all these methods need to calculate the mean expression value and the expression maximum value for each gene in each tissue as a global measure of the gene's specificity. One disadvantage of such methods, especially when there are a large number of samples for each tissue, is that confounding variation and noise may hinder learning biologically meaningful representations. Autoencoder-based data compression is preferred in this case to efficiently extract the true signal from high dimensional data and to learn latent representations corresponding to biological information of interest (Gupta et al., 2015; Lin et al., 2017; Xie et al., 2017; Ding et al., 2018). Therefore, we propose the use of an autoencoder model for detecting tissue-specific highly expressed genes. As for GRN construction, GENIE3 is a widely used tool (Marbach et al., 2012) and has been successfully applied in constructing the Arabidopsis (Ezer et al., 2017) and maize (Walley et al., 2016) GRNs. However, GENIE3 is very time consuming especially when the gene expression vector is in high dimension.

Furthermore, in GRN construction, without dimension reduction, confounder-based variations often mask true signals of biologically meaningful regulations (Yi et al., 2018; Kinalis et al., 2019). To address these issues, we applied an autoencoder to compress our sample representations in each tissue into a much lower dimension, i.e., to learn a latent space that maps M samples to D dimension ($M \gg D$) such that the biological signals presented in the original expression space can be preserved in the D-dimensional space. Then, we constructed and compared the differential regulatory network for each tissue using the embedded expression matrix produced by the autoencoder model. Moreover, key transaction factors (TFs) were identified for each tissue. We also predicted the functional modules in each differential network by performing clustering analyses.

In this study, after processing, 5,422 high-quality samples were left for analysis (either RNA-seq or microarray data) and were manually separated into case and control ("baseline") groups. Each group contains gene expression profiles of many different tissues and development stages from a wide range of studies. The control expression was obtained under normal, untreated conditions. Each sample was manually curated, and

both microarray and RNA expression levels were mapped and normalized based on original raw sequencing reads. With the data sets and the autoencoder model, we identified highly expressed genes in the soybean leaf, root, seed, and nodule. In combination with GENIE3 and the autoencoder model, tissue-specific GRNs were constructed, and hub TFs were identified. After comparing our newly constructed GRNs with the corresponding control network, a differential correlation network was constructed. This network provided us the opportunity to identify new genes and interactions with significant changes. We also clustered the network modules and provided their functional annotations. **Figure 1** shows the overall framework and our

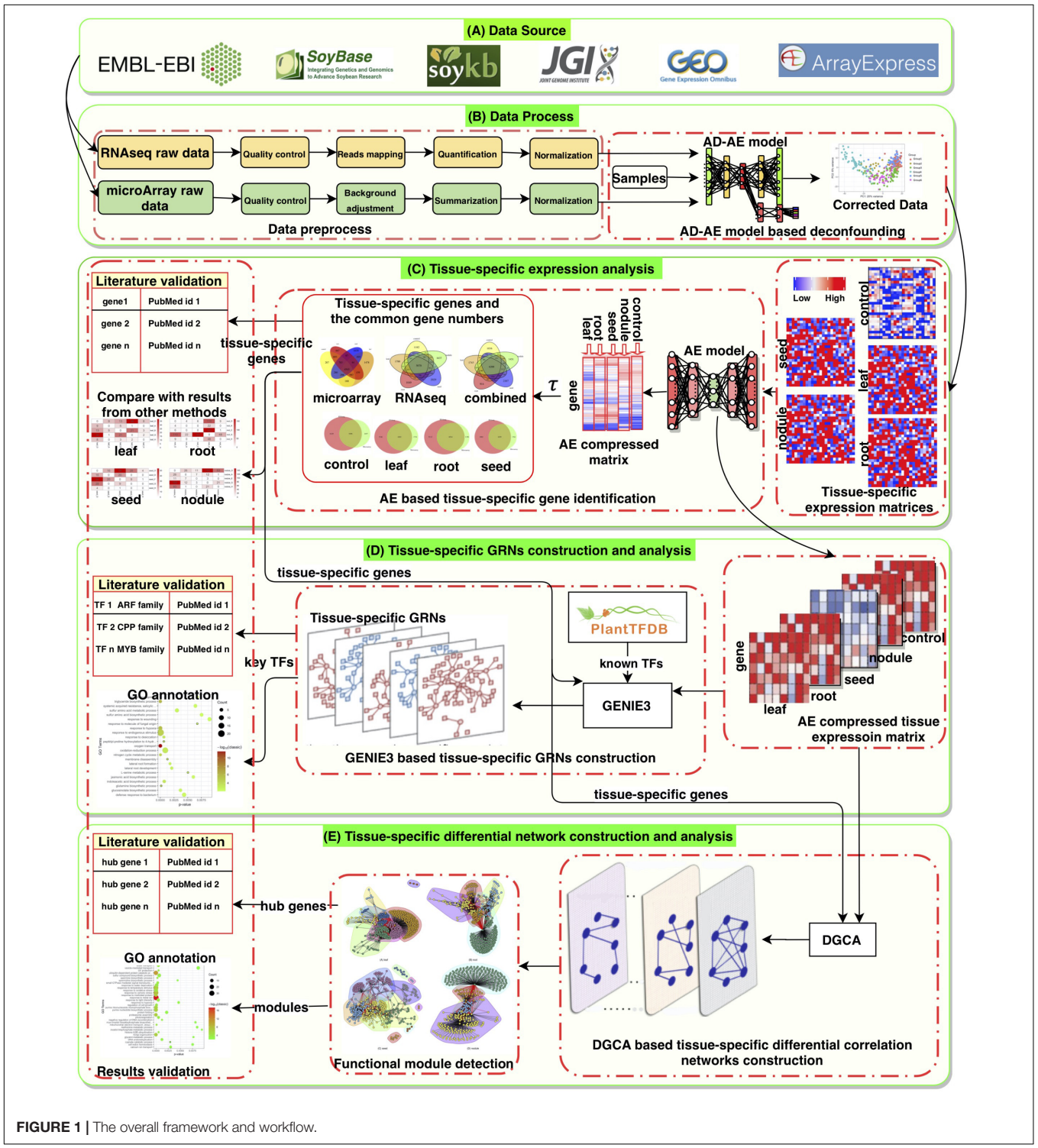


FIGURE 1 | The overall framework and workflow.

workflow, which includes five parts: (A) the data source, (B) the data process, (C) tissue-specific expression analysis, (D) tissue-specific GRN construction and analysis, and (E) tissue-specific differential network construction and analysis. To the best of our knowledge, this is the largest analysis of gene expressions in soybean tissues to date. Our results can provide more accurate targets for future tissue-specific studies and help uncover broader biological patterns.

MATERIALS AND METHODS

Data Collection, Processing, and Normalization

Selecting and pre-processing suitable microarray and RNA-seq datasets, are key issues and steps in conducting large-scale integrating analysis. After a systematic review of soybean-related microarray studies in the literature, we found a large number of samples sequenced by the Affymetrix GPL4592 (Affymetrix Glycine max Genome Array) platform. For easy data integration and normalization, only data generated from the GPL4592 platform were used for microarray datasets in this study. The RNA-Seq samples were mainly downloaded from the NCBI SRA (Leinonen et al., 2011) and the ArrayExpress database (Athar et al., 2019) together with some in-house data. Each dataset was manually checked using the following three criteria: (i) methods for the sequencing experiments, (ii) available raw data (FASTQ or SRA formats), and (iii) detailed sample information to determine whether it was included in the analysis or not. Reads obtained from the same biological sample were combined in a single FASTQ file (or in two files, for paired-end data).

The analysis pipeline is shown in **Figure 1B**, where the microarray and RNA-seq data are processed separately with different tools for quality check, raw data processing and normalization. In detail, for RNA-seq data, we used TrimGalore-0.6.5[1] to trim sequence adapters from the raw reads FASTQ files. The trimmed FASTQ files were then prepared for quality control with FastQC-0.11.8 (Wingett and Andrews, 2018), which provides a quick view on the quality of the raw sequence reads from multiple analyses, ranging from the sequence quality and GC content to library complexity. FastQC-0.11.8 also produces a report in HTML format. Then high-quality reads were aligned to the soybean genome (*Glycine max* Wm82.a4.v1) by STAR (Dobin et al., 2013). As transcripts per million (TPM) (Wagner et al., 2012) normalization is more consistent across technical replicates than other normalization methods. We normalized data using TPM for most of the downstream analysis (Li and Li, 2018), and $\log_2$ transformed raw read counts are used for quality control steps and AD-AE based confounders removal. Datasets with known batch effect information are corrected with the ComBat-Seq (Zhang et al., 2020). For the microarray data type, raw datasets are retrieved with GEOquey (Sean and Meltzer, 2007). After outliers were filtered out, we processed the CEL-type raw data with affy, an R package used to analyze oligonucleotide arrays and manufactured by Affymetrix (Gautier et al., 2004) and the oligo package developed by Carvalho and Irizarry (2010), which serves as a Bioconductor tool that supports R packages. Final data were normalized with GCRMA (Gharaibeh et al., 2008), which converts background-adjusted probe intensities to expression measures using the same normalization and summarization methods as the robust multiarray average (RMA) (Irizarry et al., 2003).

As shown in **Supplementary Figure 1**, 5,422 high-quality samples remained for further analysis, including 3,819 microarray samples and 1,603 RNA-seq samples. For each sample, we manually labeled its cultivar, tissue, development stage, and case-control information. All RNA-seq and microarray data analyzed in this work can be obtained from the European Nucleotide Archive[2] and Gene Expression Omnibus (GEO), respectively.[3] Accession numbers are summarized in **Supplementary File 1**.

Removal of Confounders

As in Dincer et al. (2020), the AD-AE model consists of one standard autoencoder and an adversary network model that takes the embedded layer as input and predicts the confounders. Here, we used the data sources as confounder variates. The autoencoder network consists of an encoder network and a decoder network. The encoder network is defined as $f_{\emptyset} : X \rightarrow Z$, which maps each sample $X \in \mathbb{R}^M$ in the input layer to the embedding layer $Z \in \mathbb{R}^D$, where M is the gene number of each sample. The decoder network tries to reconstruct X with the embedded layer Z. The optimize function is defined in Eqn. (1):

$$min_{\Phi,\psi} \mathrm{E}||x - g_{\psi}\left(f_{\Phi}\left(x\right)\right)||_2^2, \tag{1}$$

where Φ and ψ are the parameters for the encoder and decoder networks, respectively. In this study, after parameters optimization, for all tissue-specific expression data, the embedded layer size is set to 100, and the input layer sizes are the same as the gene number. We used one hidden layer for all the AE models, with the size of the half gene number, resulting in a 50%-dimension reduction. The minibatch size was set to 128, and we trained the model with the Adam optimizer using a learning rate of 0.0001. We applied the ReLU activation to all layers except the last layer, where we applied linear activation. The adversary model maps the embedding Æ to confounders. To reduce the confounding effects, after training, the autoencoder needed to converge to generate an embedding that contains less information about the confounder, and the adversary model needed to converge to reach a random prediction performance.

The adversary model h_v is optimized with the following objective:

$$min_v \mathbb{E}[L(h_v\left(x\right), c)] \tag{2}$$

where c is the confounder, and L is the loss function with categorical cross entropy loss. For the adversarial model, as in Dincer et al. (2020), a fully connected neural network has two hidden layers with 100 hidden nodes in each layer, and this network uses the ReLU activation function. The last layer's

[1]https://github.com/FelixKrueger/TrimGalore

[2]https://www.ebi.ac.uk

[3]https://www.ncbi.nlm.nih.gov/geo/

number of nodes corresponds to the number of data sources, and this layer has softmax activation. First, we trained each model separately. We then used the following objective function for the alternative joint training.

$$min_{\Phi,\psi,v}\mathbb{E}[||x - g_\psi(f_\Phi(x))||_2^2 - \lambda L(h_v(x), c)] \quad (3)$$

We first froze the weights of the adversary model and trained the autoencoder model for one epoch on a randomly selected minibatch of the data using stochastic gradient descent. We then froze the autoencoder model and trained the adversary model for an entire epoch to minimize Eqn. (2). For each dataset, we applied a fivefold cross validation to select the hyperparameters of autoencoder models, which is suitable considering the data set size. When training the model, we used 80% of the data for training with the rest left for validation, and we determined the optimal epochs based on the validation loss. We used the reconstructed data from the autoencoder model as input for our next step analysis. After removing confounders, for each tissue, data sets from different sources were more consistent and datasets from different tissues became more separate.

Tissue-Specific Gene Identification, Gene Regulatory Networks, and Differential Network Construction

Instead of using the mean value across all samples as the gene expression value, we proposed using an autoencoder model to compress the gene expression vector to a lower dimension. The encoder network is defined as $f_\emptyset : Y \rightarrow Z$, which maps each gene $Y \in \mathbb{R}^N$ in the input layer to the embedded layer $Z \in \mathbb{R}^D$, where N is the sample number of each gene. Here the input to the autoencoder model is the gene expression values across all samples. To get genes highly expressed in each tissue for each tissue-specific gene expression matrix, the embedding layer size was set to 1-dimension, the input layer size was set with the same sample number of each tissue. We used one hidden layer for all the AE models, with the size set as half the sample number, which resulted in a 50%-dimension reduction. The minibatch size was set to 12, and we trained the model with the Adam optimizer using a learning rate of 0.0001. We used the ReLU activation for all layers except the last layer, where we applied softplus activation. With the compressed gene expression matrix, we used τ as defined in Eqn. (4) to measure the tissue specificity of each gene.

$$\tau = \frac{\sum_{i=1}^{n}(1 - \widehat{y_i})}{n-1}; \widehat{y_i} = \frac{y_i}{\max_{1 \le i \le n}(y_i)} \quad (4)$$

where n is the tissue number including the control condition. y_i is the expression of the gene in tissue i. We used the 3rd quantile of all τ values as the threshold value to filter highly expressed genes in each tissue and in the control condition.

For constructing tissue-specific GRNs, we used the same autoencoder model as that used to identify tissue-specific genes except that we set the embedding layer size to 64-dimension. However, the embedding layer size for the nodule tissue was set to size 32 (because only 47 samples exist for nodules). A total of 3,747 *Glycine max* known TFs were obtained and in combination with tissue-specific genes identified above, 662, 736, 781, and 617 TFs were highly expressed in leaf, root, seed, and nodule, respectively. Among these TFs, 110, 93, 155, and 110 are uniquely highly expressed in leaf, root, seed, and nodule, respectively. Some genes or TFs are not expressed in some tissues; therefore, they will not have predicted targets. Hence, the compressed expression matrix for each tissue is filtered to remove low variance genes by using the filterGenes function in the DGCA R package (Mckenzie et al., 2016; Zhou et al., 2020). GENIE3 ran under the default parameters setting and restricted the candidate regulators to the filtered tissue-specific TFs. By utilizing the embedded data sets, the filtered tissue-specific TFs and the GENIE3 (Huynh-Thu et al., 2010) algorithm, four tissue-specific (leaf, root, seed, and nodule) GRNs were constructed. To find TFs that play important roles in each GRN, we ranked all TFs based on their degree.

Understanding interactions that specifically exist in one tissue in comparison with those in the control conditions is important to the understanding of tissue-specific gene regulation. Therefore, we constructed the differential co-expression network of each tissue based on paired tissues and control expression data produced by the autoencoder model. Different from differential expression, differential co-expression operates on the level of gene pairs rather than individual genes. To deal with this, the Fisher z-transformation is employed as in Eqns. (5 and 6).

$$z = \frac{1}{2} log_e\left(\frac{1 + \rho}{1 - \rho}\right) \quad (5)$$

$$dz = \frac{(z_1 - z_2)}{\sqrt{|s_{z_1}^2 - s_{z_2}^2|}} \quad (6)$$

where ρ is the Pearson correlation coefficients, z_1 and z_2 corresponding to tissue-specific and control values, respectively, and s_z^2 refers to the variances of the z-score. Using the difference in z-scores, dz, a two-sided p-value can be calculated using the standard normal distribution and by adjusting the p-values for multiple hypotheses tests with the conservative Benjamini-Hochberg p-value adjustment method. Gene pairs can then be ranked based on the relative strengths of their differential correlation. To determine the top-ranked significantly changed hub genes, we sought to compare the differential correlation from paired tissues and the control genes co-expression network. To get hub genes, we calculated the average change in correlation for each gene with all others and set top-ranked genes as hub genes. This paper also presents a series of detailed tests to determine the difference in mean z-scores and adjust the p-value by random permutation samples 100 times. We used the DGCA (Mckenzie et al., 2016) R package to conduct the corresponding analyses in this study. DGCA offers a convenience function for extracting gene lists corresponding to the differential correlation, converting the resulting gene symbols to inputs for gene ontology enrichment testing and detecting functional modules. The whole pipeline is shown in **Supplementary Figure 2**.

RESULTS

Tissue-Specific Gene Expression

By utilizing the autoencoder model, the collected soybean expression data were processed and their PCA plots were shown in **Supplementary Figure 3**. As shown in **Supplementary Figure 3**, after data reconstruction, tissue-specific expression signatures of data can still be maintained. By utilizing the autoencoder model and the τ index (between 0 and 1), genes with a τ index close to 1 were more specifically expressed in one tissue, while genes with a τ index closer to 0 are equally expressed across all tissues studied (Yanai et al., 2005).

We identified highly expressed genes in the leaf, root, seed, nodule and genes under the baseline condition, using both microarray and RNA-seq data sets. We combined the compressed 1-dimension expression value of each tissue and its control together and compared the global expression patterns among tissues to identify highly expressed tissue-specific genes. A gene is filtered as a tissue-specific gene only if it is highly expressed in that tissue compared to values of other tissues and the control mixture, which includes samples that do not involve exposure to the treatment or intervention from any studies (**Supplementary Figures 4, 5**). The Venn plot of combined tissue-specific genes in both microarray and RNA-seq data sets is shown in **Figure 2C**. In summary, we detected 1,743, 914, 2,107, and 1,451 genes highly expressed in leaf, root, seed, and nodule, respectively. The detailed gene list is shown in **Supplementary File 2**. Our method outperforms traditional methods in two ways: (1) with the autoencoder model, we can more accurately detect tissue-specific genes in each tissue, as shown in **Figures 2A–G**; (2) more common tissue-specific genes can be detected between the microarray and RNA-seq data types relative to traditional methods. The Venn plot of common tissue-specific genes between the microarray and RNA-seq identified using the traditional method is shown in **Figure 3**, which shows much fewer common tissue-specific genes between the microarray and RNA-seq data types.

To verify the accuracy of all the tissue-specific genes, we conducted a comprehensive literature search and finally collected 2,108, 1,908, 1,516, and 2,122 tissue-specific genes in total for leaf, root, seed, and nodule, respectively (**Supplementary File 2**), which are used as the benchmark datasets (Libault et al., 2009, 2010; Severin et al., 2010; Asakura et al., 2012; Jones and Vodkin, 2013; Brown and Hudson, 2015; Machado et al., 2020; Moisseyev et al., 2020). **Supplementary Figure 6** shows the common genes between our results (denoted as O) and genes found in the eight benchmark studies (from A to H, with some detailed information from each study shown in **Supplementary File 2**). Many of these tissue-specific genes from the eight studies can always be detected by our method in contrast with other methods which usually have no gene in common, even with similar tissue-specific gene numbers for comparison. Different from other methods, we considered not only gene expression in specific tissues but also their expression under the control condition; hence, the genes detected by our methods are more likely highly expressed tissue-specific genes. To further verify the accuracy of our results, the top-10 ranked genes (genes with highest expression values in a tissue in comparison with other tissues and the baseline condition) of each tissue were searched in PubMed, as shown in **Supplementary Table 1**. Most of these top-ranked genes have direct PubMed publications that support their high expression in the corresponding tissue. A detailed tissue-specific expression of each gene can also be viewed through the link[4] from the Soybean Expression Atlas. Considering the top-10 ranked nodule genes in **Supplementary Table 1**, we independently searched their average expression in various soybean tissues using the locus name (genome version: Gmax_275_Wm82.a2.v1), which confirmed their nodule-specific expression.

Functional Annotation of the Top-Ranked Tissue-Specific Genes

To further validate the accuracy of the tissue-specific genes, we performed a GO functional enrichment analysis of the top 20 ranked tissue-specific genes using the TopGO package (Alexa et al., 2006). Only the top 20 significant GO terms of the unique genes are reported, except for nodules where 23 GO terms are reported. As shown in **Supplementary File 2**, many of the leaf-specific genes are those responding to red/blue light, such as Glyma.08G173700, Glyma.11G221000, Glyma.13G046200, Glyma.18G036400, Glyma.19G046600, and Glyma.19G046800. Among these genes, Glyma.11G221000 and Glyma.18G036400 also function in responding to cold and in the defense response to bacterial pathogens. Genes Glyma.13G046200, Glyma.19G046600, and Glyma.19G046800 also take part in the carbon fixation process. In a leaf, Glyma.14G061500 has a function in water transport and also in response to water deprivation. Another top-ranked gene is Glyma.13G347700, which is enriched in the processes of lateral root formation, responding to the abscisic acid, and it has a defense response to a bacterial pathogen. The Glyma.09G044200 gene takes part in processes of positive regulation of microtubule depolymerization, and it responds to cadmium ion and the regulation of multicellular organism growth (Berkowitz et al., 2008).

Many genes in seed are significantly enriched in lipid storage, embryo development, and seed germination processes (such as Glyma.10G246300 and Glyma.09G044200). Genes highly expressed in nodules are enriched in processes of nodulation and nitrogen fixation, which is consistent with previous knowledge (Elhady et al., 2020). Among the nodule-specific genes, Glyma.04G079200 responds to hypoxia and oxygen transport. Glyma.11G238800 participates in the pathogen defense response, and it is also highly expressed in the nodule tissue of *Lotus japonicus* (Guenther and Roberts, 2000). Glyma.10G199000 is highly expressed during nodule development (Marcker et al., 1984). Therefore, many of these tissue-specific genes have been experimentally verified, and others are good candidates for further exploration.

Differential Network and Hub Genes

An early study (Sonawane et al., 2017) shows that network edges have higher tissue specificity than network nodes. Therefore,

[4]http://venanciogroup.uenf.br/cgi-bin/gmax_atlas/index.cgi

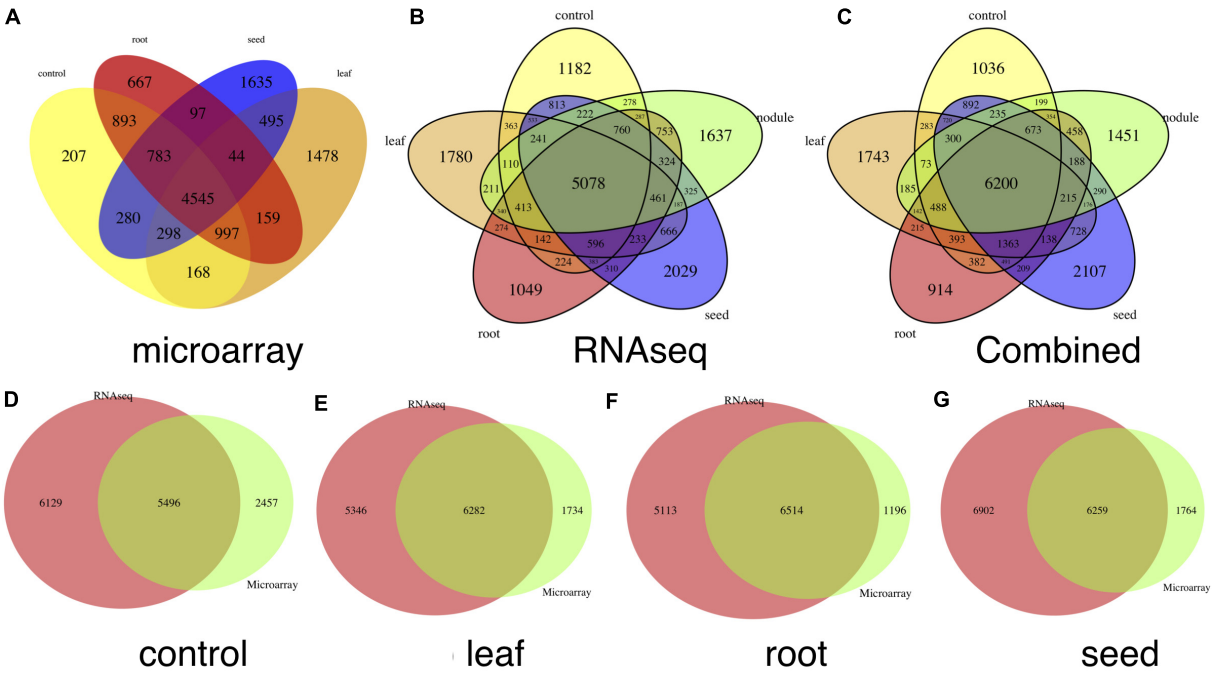


FIGURE 2 | Venn plots of tissue-specific genes. **(A)** Number of tissue-specific genes in each tissue and control condition detected in the microarray data sets, **(B)** numbers of issue-specific genes in each tissue and control condition in the RNA-seq data sets, and **(C)** number of issue-specific genes in each tissue and control condition in combination with **(A,B)**. **(D–G)** Show common and tissue-specific genes detected in both microarray and RNA-seq data sets in control, leaf, root, and seed tissues, respectively.

comparing interaction changes between genes in a tissue and genes in the control group is important for understanding tissue-specific gene expression. We used the autoencoder compressed gene expression matrixes and the paired control gene expression matrix for tissue-specific differential network construction. Since genes with low average expression levels or low variance in expression levels are less likely to have biologically relevant differences between conditions, they were filtered out before further analysis as per previous protocol in studies (Mckenzie et al., 2016). We calculated the Pearson correlation value between any pair of genes in each tissue and also the corresponding value in the control condition; then, the significance of the difference between the two values was tested and the *p*-values were adjusted by the Benjamini-Hochberg method (Benjamini and Hochberg, 1995). We ranked the differential edges based on the adjusted *p*-values, and the top 1,000 ranked edges were selected for each tissue for the network construction. Results of each network after fast greedy clustering (Clauset et al., 2004) are shown in **Figure 4**, where there are 7, 4, and 10 modules in leaf, root, and seed differential networks, respectively. The nodule results are independently shown in the nodule section. We performed the biological process enrichment analysis for the modules in these tissues using the TopGO package (Alexa et al., 2006). Detailed gene module information and the GO annotation bubble plot are shown in **Supplementary File 3**.

To get hub genes in each network, we calculated each gene's average gain or loss of correlation in the data set with all others (see details in "Materials and Methods" section). The top-ranked hub genes (genes with biggest gain or loss of correlation) of each network in **Figure 4** are shown in **Supplementary Table 2**. Many of these genes in **Supplementary Table 2** are known for their expression in the corresponding tissues. For example, Glyma.11G111400 is a fructose-bisphosphate aldolase protein, which is a key plant enzyme involved in glycolysis and the Calvin cycle in wheat and corn leaves (Lv et al., 2017). The root hub gene Glyma.07G248600 is a C2 domain-containing protein. Many C2 domain-containing proteins, such as CaSRC2-1, are known to be highly expressed in roots (Kim et al., 2008). The seed hub gene Glyma.13G295200 is a zinc finger CCCH domain-containing protein, which is known for its association with seed oil accumulation (Li et al., 2017).

Tissue-Specific Gene Regulatory Networks and Transaction Factors

GENIE3 (Huynh-Thu et al., 2010) was used for tissue-specific GRN construction (see the "Materials and Methods" section for details) after we compressed each tissue's expression matrix to a low dimension. To find the optimal dimension size, we tuned our model using the DREAM 5 challenge *E. coli* expression data and the corresponding benchmark network with varying dimension sizes from 12 to 24, 32, 64, and 96, as well as using the original sample size. According to the area under the receiver operator characteristic (ROC) curves, the optimal dimension size was 64. For an expression matrix with more than 64 samples

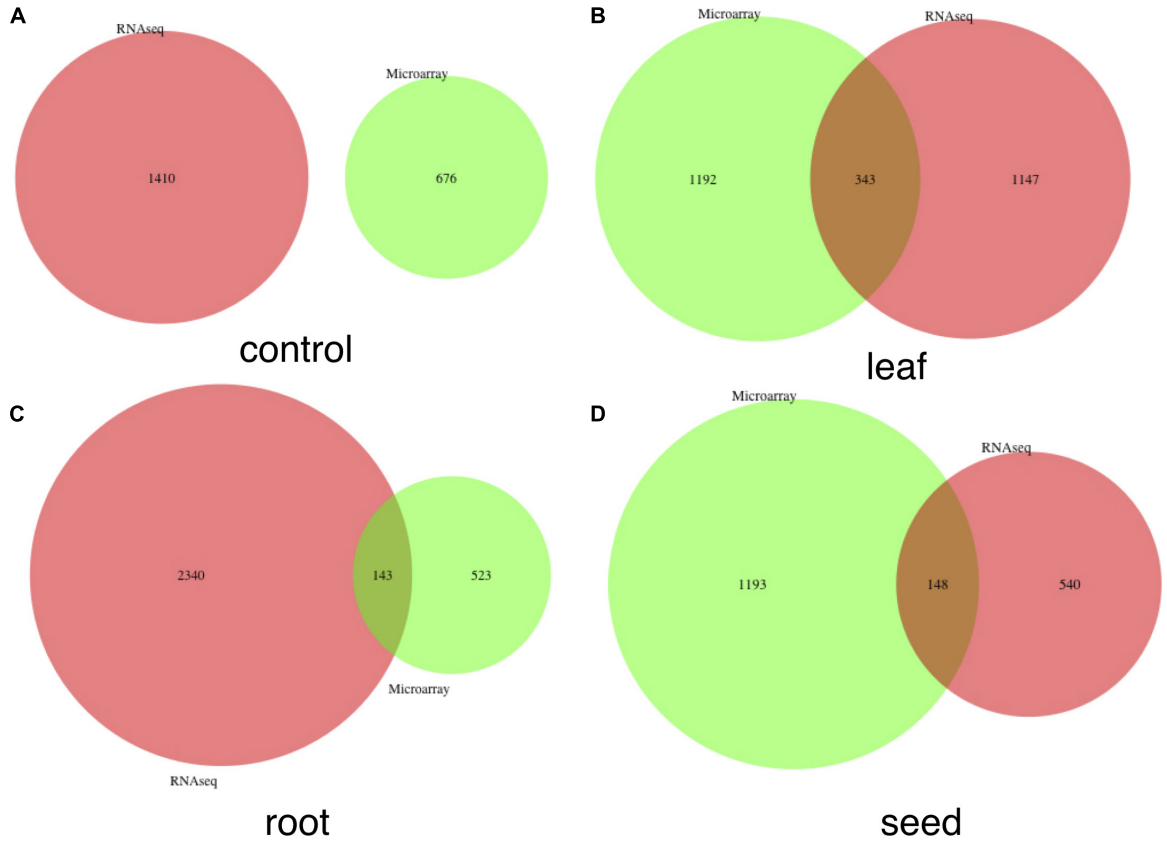


FIGURE 3 | Venn plots of common genes detected in both microarray and RNA-seq data sets without the autoencoder model. These genes were detected directly with the τ index. **(A–D)** Correspond to numbers of common genes in the control, leaf, root, and seed, respectively.

(leaf, root, seed, and control) we compressed them to the 64-dimension size. Other sample data (nodule) were compressed to the 32-dimension size due to having fewer samples than 64. For candidate regulators, we downloaded 3,747 soybean TFs from the PlantTFDB database (Jin et al., 2017). We combined the detected tissue-specific genes, 110, 93, 155, and 110, and from them, we found uniquely highly expressed TFs in leaf, root, seed, and nodule, respectively. Other parameters of GENIE3 were set as the default values. Prior to the analysis, we filtered out all the genes with low variance across samples as in Sun et al. (2020) by using the filterGenes function in the DGCA R package (Mckenzie et al., 2016). Because we only constructed tissue-specific GRNs, for each tissue only tissue-specific genes were considered, resulting in 660, 632, 737, and 582 genes left for leaf, root, seed, and nodule, respectively. For better visualization and comparison, each GRN was constrained to include only the top 1,000 high-weight edges calculated by GENIE3 (Walley et al., 2016; Huang et al., 2018). As shown in **Figure 5**, the top 5 central nodes in each network are all classified as biologically essential for the corresponding tissue.

The TF with the highest degree (number of connecting edges) in each GRN means it likely regulates many other genes or TFs, and such TFs are important in tissue-specific gene expression. We ranked all network nodes based on their degree. **Supplementary Table 3** shows detailed information on the top-ranked TFs in the four tissues. Two root TFs are from the WRKY transcription factor family. Genes in the WRKY family play important roles in plants responding to microbial pathogens (Yang et al., 2017). The WRKY transcription factor genes are highly expressed in hairy roots and can enhance the resistance of soybean to the oomycete pathogen *Phytophthora sojae* (Cui et al., 2019). Another two root TFs are from the GRAS family, which broadly participates in many critical processes such as signal transductions, root radial elongations, axillary shoot meristem formations, and stress responses in plants (Bolle et al., 2000; Li et al., 2018). Overexpressing the GRAS family gene in hairy roots can improve the resistance of soybeans to drought and salt stresses (Wang et al., 2020). Three BZIP family TFs are top-ranked in seed. The BZIP genes play important roles in seed maturation and storage protein gene regulation (Lara et al., 2003; Wang Z. H. et al., 2019). The other two seed TFs are from the MIKC and MADS family, which often play potentially essential roles in seed development (Fan et al., 2013). Three nodule-specific TFs are from the C3H transcription factor family, one from NAC and one from the ERF family. The NAC family proteins are highly expressed during early symbiotic events in *Medicago truncatula* and *Sinorhizobium meliloti* (Lohar et al., 2007). The C3H and ERF family TFs are also important in the symbiosis process of *Lotus japonicus* (Asamizu et al., 2008).

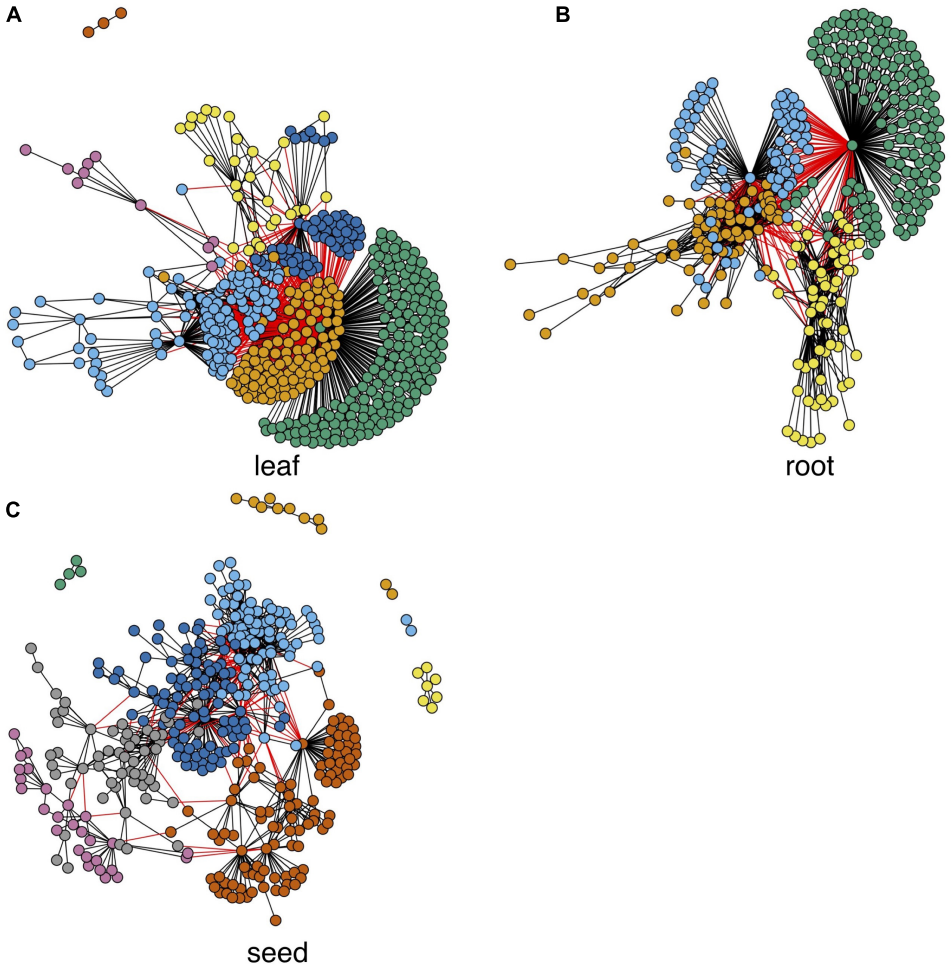


FIGURE 4 | Tissue-specific differential networks and their functional modules. Each node represents a gene, and each edge represents a differential edge. Modules are marked by nodes in different colors.

Nodule-Specific Genes and Gene Regulatory Network

The symbiosis between the rhizobium and soybean is a much cheaper and more effective agronomic practice for ensuring an adequate supply of nitrogen. Due to its agricultural importance, many efforts are underway to identify the underlying molecular mechanisms of the symbiosis process. As a result, many genes have been identified, such as the LysM receptor-like kinases, NFP and LYK3 (Radutoiu et al., 2003), which mediate the perception of Nod factors. SymRK is required for root nodule development (Chen et al., 2012) and miR393j-3p limits nodule development through repression of the nodulin gene ENOD93 (Yan et al., 2015). A higher percentage of NIN-like and C2H2 nodule-specific TFs has been reported (Libault et al., 2010; Severin et al., 2010). Furthermore, over 200 nodulins (organ-specific plant proteins induced during symbiotic nitrogen fixation) have been experimentally validated (Roy et al., 2020). However, considering the complexity of the symbiosis process, more genes are likely involved and need to be discovered. **Figure 6** shows many nodule-specific highly expressed genes and their corresponding GRNs based on our method.

Figure 6A shows that all the detected nodule-specific genes are highly expressed only in nodules that compare with this type of gene in other tissues. Of the top 10 ranked nodule specific genes, eight also exist in the Soybean Expression Atlas database, except for the Glyma.10G198800 and Glyma.10G199100 genes. All eight nodule-specific genes are exclusively highly expressed in nodules and have no expression value in any other tissues, as shown in **Figure 6B**. Most of the top-ranked genes identified by our method match PubMed publications that support their high expression in the corresponding tissue as shown in **Figure 6C**. The bubble plot of the enriched biological processes of these genes is shown in **Figure 6D**.

According to the GO enrichment analysis results, many biological processes are related to the symbiosis process. For instance, copper is an essential nutrient for symbiotic nitrogen fixation, and cellular copper ion homeostasis is an important process in rhizobia-infected nodule cells (Senovilla et al., 2018). In *Lotus japonicus*, the expression of two thiamine biosynthesis

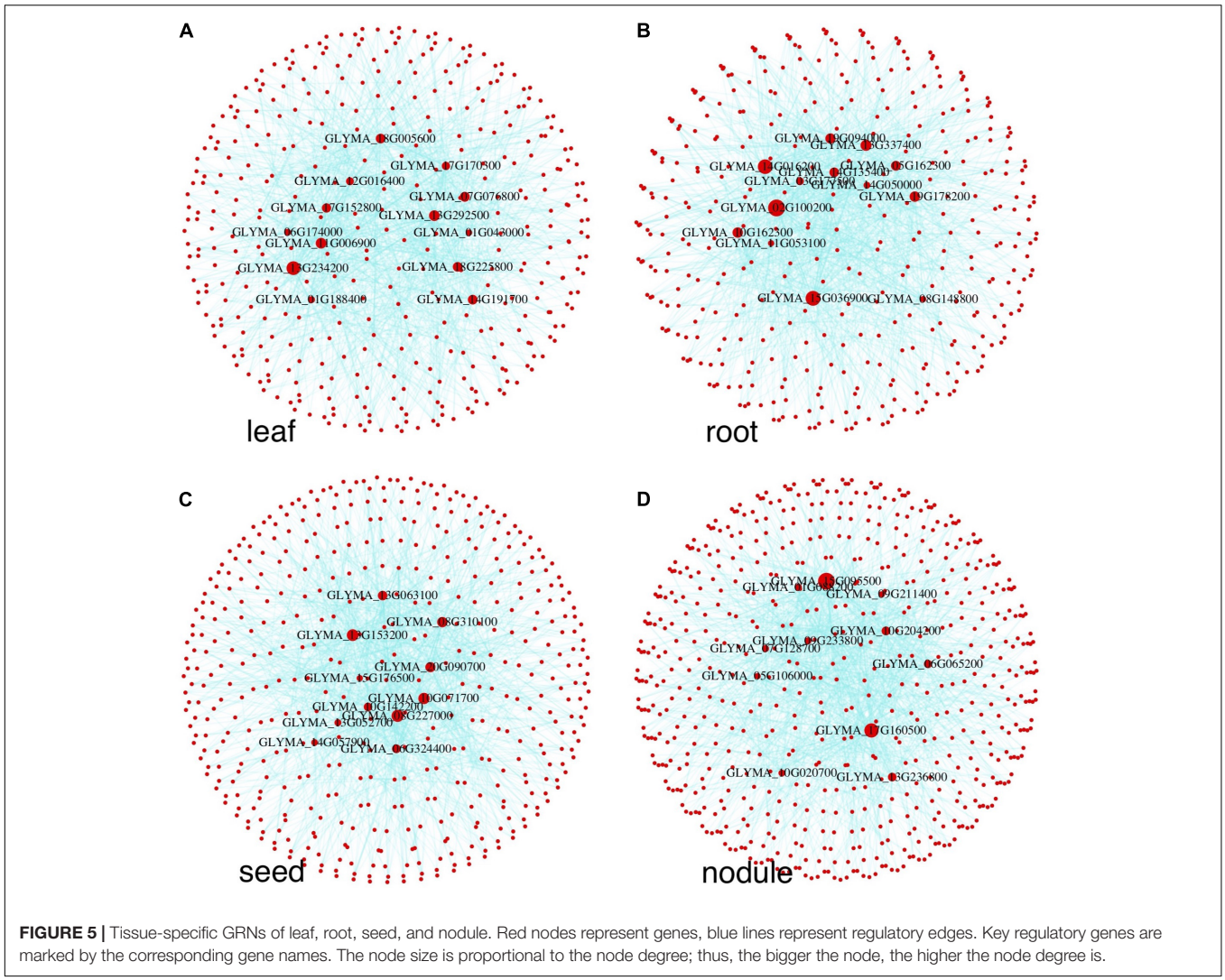


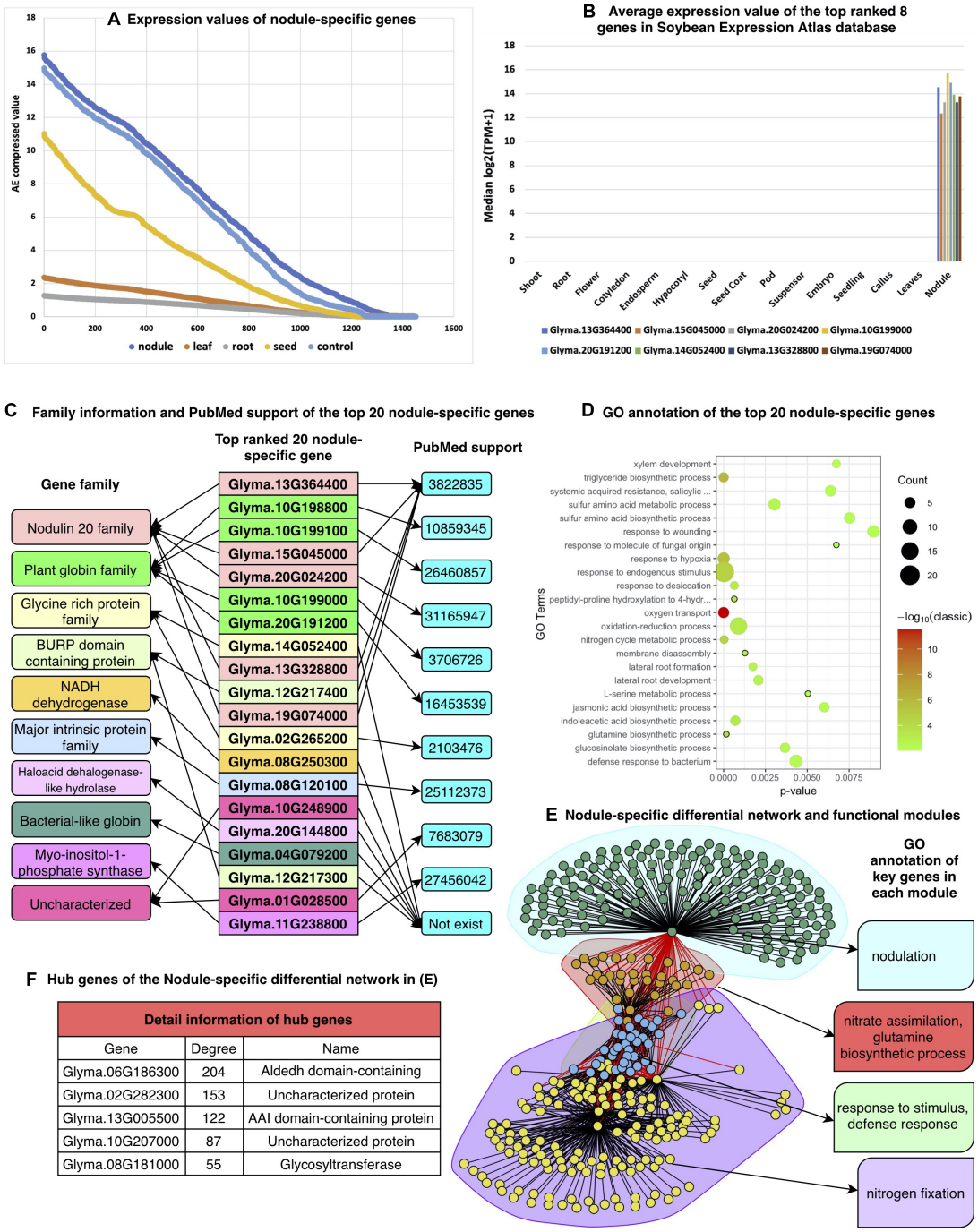
FIGURE 5 | Tissue-specific GRNs of leaf, root, seed, and nodule. Red nodes represent genes, blue lines represent regulatory edges. Key regulatory genes are marked by the corresponding gene names. The node size is proportional to the node degree; thus, the bigger the node, the higher the node degree is.

genes, THI1 and THIC, is enhanced by inoculation with rhizobia but not by inoculation with arbuscular mycorrhizal fungi, and thiamine biosynthesis genes can promote nodule growth (Nagae et al., 2016). The ratios of sucrose/fructose in nodules can be changed in response to nitrate, indicating that nitrate affects sugar concentration in nodules (Streeter, 1981). Some genes are enriched in the process of responding to fructose, sucrose, etc. Other biological processes like the response to oxidative stress and response to hypoxia have all been shown to be involved in the soybean symbiosis process (Van Heerden et al., 2008; Zilli et al., 2011; Pucciariello et al., 2019).

As noted in the previous section, we constructed the nodule-specific differential regulatory network and found its functional modules. As shown in **Figure 6E**, four modules were detected, which are related to (1) nodulation, (2) nitrate assimilation as part of the glutamine biosynthetic process, (3) response to stimulus in a defense mode, and (4) nitrogen fixation, and all of them are well known for their relationship to the symbiosis process. Furthermore, except for the hub genes as in **Figure 6F**, many other genes are also related to the nodule function.

Of the top 100 ranked nodule-specific highly expressed genes, many symbiosis-related genes are identified, including four leghemoglobin gene Glyma.10G198800, Glyma.10G199100, Glyma.10G199000, and Glyma.20G191200, as well as nine nodulin genes (Glyma.13G364400, Glyma.15G045000, Glyma.20G024200, Glyma.13G328800, Glyma.19G074000, Glyma.06G216500, Glyma.02G204500, Glyma.17G073400, and Glyma.08G076800). Furthermore, some top-ranked TFs are also identified as symbiosis related. These include Glyma.01G159200 from the NIN-like family, Glyma.15G173300, Glyma.17G051400, and Glyma.09G014100 from the NF-YA/NF-YB family, seven genes (Glyma.13G094400, Glyma.05G106000, Glyma.06G303100, Glyma.12G100600, Glyma.15G069300, Glyma.17G065800, and Glyma.17G160500) from the MYB family and Glyma.01G101800, Glyma.02G144400, Glyma.14G110900, and Glyma.18G042300 from the C2H2 family (**Supplementary File 2**).

Many genes from these gene families are known to function in the symbiosis process. For example, a higher percentage of NIN-like and C2H2 nodule-specific TFs have been reported



F Hub genes of the Nodule-specific differential network in (E)

Detail information of hub genes		
Gene	Degree	Name
Glyma.06G186300	204	Aldedh domain-containing
Glyma.02G282300	153	Uncharacterized protein
Glyma.13G005500	122	AAI domain-containing protein
Glyma.10G207000	87	Uncharacterized protein
Glyma.08G181000	55	Glycosyltransferase

FIGURE 6 | Results of nodule-specific gene analysis. **(A)** Expression values of 1,451 nodule-specific genes compared with those in other tissues, **(B)** validation of the expression patterns of top-ranked nodule-specific genes from the Soybean Expression Atlas database, **(C)** detailed gene family and PubMed support for the top-20 ranked nodule genes. Genes with the same color are in the same gene family. **(D)** GO biological process annotation of the top 20 nodule-specific genes, **(E)** differential correlation network of the nodule-specific genes and key function of each module, and **(F)** hub genes of each module.

(Libault et al., 2010; Severin et al., 2010). Furthermore, NIN-like and C2H2 TFs are important in nitrate signaling (Konishi and Yanagisawa, 2013) and symbiosome differentiation during nodule development (Sinharoy et al., 2013). Therefore, our newly identified TFs are more likely to play important roles in the symbiosis process. We also found four nodule-specific ERF TFs (Glyma.03G112000, Glyma.07G114000, Glyma.09G072000, and Glyma.09G233800) that are essential for nodule differentiation and development (Vernie et al., 2008). Next, in the top 100 ranked nodule-specific highly expressed genes, we analyzed the

12 nodule-related module hub genes (**Supplementary File 2**) in the nodule-specific differential network. Notably, two of these genes (Glyma.18G041100 and Glyma.11G215500) are glutamine synthetase genes, which are tightly controlled enzymes located at the core of nitrogen metabolism. Glutamine synthetase catalyzes the first step in nitrogen assimilation, which is the ATP-dependent condensation of ammonium with glutamate (Seabra and Carvalho, 2015). The hub gene Glyma.17G045800 is a sucrose synthase gene, which has an essential function in the nitrogen fixation process (Gordon et al., 1999). Another hub gene, Glyma.08G181000, has been verified to function in the isoflavone biosynthetic pathway (Gupta et al., 2017). Due to the significant expression changes of these genes in the nodule tissue, they are likely to play important roles in the symbiosis process.

DISCUSSION AND CONCLUSION

Transcriptome data is still the main data source for researchers to obtain useful information about plant biological processes and to identify key biomarker genes related to specific phenotypes. Each experimental study investigates gene expression in a range from a few samples to hundreds of samples. Due to the sample condition or method difference, even for the same plant tissue, different labs can obtain quite different results. Therefore, large-scale integration analysis of all the available datasets is needed to help us better understand gene expression from the systematic view. However, several factors make such analysis difficult; for instance, different studies utilize different platforms. The microarray platform is the most popular at first, but is dominated by the RNA-seq later. Besides, different studies use different kinds of plant samples under different treatments—or samples are collected at different development stages and time points. Over the years, several meta-studies have been conducted, but the scale and depth have been limited.

In this study, we systematically collected more than 7,000 raw sequencing data sets, which were mapped and processed in a uniform way. To our knowledge, this is the largest transcriptome analysis of soybeans until now. In the data normalization and processing, we proposed utilizing the unsupervised autoencoder model. Two features of unsupervised learning make it well suited to gene expression analysis. The first feature is the ability to train informative models without supervision, as it is challenging to obtain a high number of expression samples with coherent labels. Although many new expression profiles are released daily, the portion of the datasets with labels of interest is often too small. A second feature of unsupervised learning is: models are trained to extract patterns from the data without imposed hypotheses or restrictions. This aspect can be key to unlocking biological mechanisms unknown to the scientific community. To minimize the difference between data from different sources, we proposed using the unsupervised AD-AE machine learning model, which can efficiently remove confounders with the collected data sets. Because each tissue has many samples, to extract important signals from noises, the autoencoder model can efficiently compress expression values to a lower dimension. With the normalized and processed data sets, we analyzed highly expressed tissue-specific genes in leaf, root, seed and nodule. Besides, we constructed the tissue-specific GRNs and differential correlation networks based on these networks, and we identified key TFs, functional modules, and hub genes. According to our analysis, many identified genes have had the tissue-specific expression. The results were integrated into SoyKB. These tissue-specific genes may help researchers test hypotheses in downstream experiments and functional genomics studies. However, several limitations exist in this study. Although many tissues and development stages were involved in the collected datasets, here we only showed results for seed, root, leaf, and, nodules as these four tissues occupied the most samples. More attention will be paid to other tissue analysis in the future.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

LTS, DX, TJ, and GS contributed to the conception and design of this study. LTS, CX, SZ, and LS processed the datasets. LTS, CX, and LS performed the data analyses. LTS wrote the first draft of the manuscript. All authors contributed to manuscript revision and also read and approved the submitted version.

FUNDING

This work was supported by the National Science Foundation Plant Genome Program (grant no. #IOS-1734145) and National Institutes of Health (R35-GM126985). This work used the high-performance computing infrastructure provided by the Research Computing Support Services at the University of Missouri.

ACKNOWLEDGMENTS

We would like to thank Carla Roberts for thoroughly proofreading this manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.831204/full#supplementary-material

Supplementary Figure 1 | Sample number distribution. **(A)** Sample number distribution of all the collected samples, **(B)** sample number of each tissue of RNAseq datasets, and **(C)** sample number of each tissue represented by microarray datasets.

Supplementary Figure 2 | Detailed pipeline of tissue-specific gene identification, GRNs, and differential network construction.

Supplementary Figure 3 | PCA plots of the AD-AE reconstructed datasets. **(A)** The PCA plot of the reconstructed microarray control dataset, **(B)** the PCA plot of the reconstructed RNA-seq control dataset, **(C)** the PCA plot of the reconstructed microarray case dataset, with samples from leaf, root, and seed, and **(D)** the PCA plot of the RNA-seq case dataset, with samples from leaf, root, seed, and nodule. Each color represents a different tissue, and the same shape represents samples from the same tissue.

Supplementary Figure 4 | Expression heatmap of highly expressed tissue-specific genes detected in the microarray data sets. **(A–D)** Correspond to tissue-specific genes in control, leaf, root, and seed, respectively.

Supplementary Figure 5 | Expression heatmap of highly expressed tissue-specific genes detected in the RNA-seq data sets. **(A–E)** Correspond to tissue-specific genes in control, leaf, root, seed, and nodule, respectively.

Supplementary Figure 6 | (1) Tissue-specific genes collected from eight benchmark studies **(A–H)** and results from our method and (2) heatmap of common gene numbers between results from different studies. O represents our own results. Leaf_O represents leaf genes detected by our method and leaf_A represents leaf genes detected in study **(A)**. Others are correspondingly defined in the figure.

Supplementary File 1 | Detailed information of all samples used in this study. Microarray_Control shows all microarray samples in control condition. Microarray_Case shows all microarray samples with various treatments. RNA-seq_Control shows all RNA-seq samples in the control condition. RNA-seq_Case shows all RNA-seq samples with various treatments. RNA-seq SRAruninfo shows all the SRA IDs in the RNA-seq raw sequencing data that we used. Microarray GEOid shows all the GEO IDs in the microarray raw sequencing data that we used. In-house sample data show all the sample information provided by our collaborators.

Supplementary File 2 | Detailed results of all tissue-specific genes.

Supplementary File 3 | Gene and GO annotation information of all the functional modules in **Figure 5**.

REFERENCES

Alexa, A., Rahnenführer, J., and Lengauer, T. (2006). Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* 22, 1600–1607. doi: 10.1093/bioinformatics/btl140

Araujo, I. S., Pietsch, J. M., Keizer, E. M., Greese, B., Balkunde, R., Fleck, C., et al. (2017). Stochastic gene expression in *Arabidopsis thaliana*. *Nat. Commun.* 8:2132.

Asakura, T., Tamura, T., Terauchi, K., Narikawa, T., Yagasaki, K., Ishimaru, Y., et al. (2012). Global gene expression profiles in developing soybean seeds. *Plant Physiol. Biochem.* 52, 147–153. doi: 10.1016/j.plaphy.2011.12.007

Asamizu, E., Shimoda, Y., Kouchi, H., Tabata, S., and Sato, S. (2008). A positive regulatory role for LjERF1 in the nodulation process is revealed by systematic analysis of nodule-associated transcription factors of *Lotus japonicus*. *Plant Physiol.* 147, 2030–2040. doi: 10.1104/pp.108.118141

Athar, A., Fullgrabe, A., George, N., Iqbal, H., Huerta, L., Ali, A., et al. (2019). ArrayExpress update – from bulk to single-cell expression data. *Nucleic Acids Res.* 47, D711–D715. doi: 10.1093/nar/gky964

Benito, M., Parker, J., Du, Q., Wu, J., Xiang, D., Perou, C. M., et al. (2004). Adjustment of systematic microarray data biases. *Bioinformatics* 20, 105–114. doi: 10.1093/bioinformatics/btg385

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Berkowitz, O., Jost, R., Pollmann, S., and Masle, J. (2008). Characterization of TCTP, the translationally controlled tumor protein, from *Arabidopsis thaliana*. *Plant Cell* 20, 3430–3447. doi: 10.1105/tpc.108.061010

Bolle, C., Koncz, C., and Chua, N. H. (2000). PAT1, a new member of the GRAS family, is involved in phytochrome A signal transduction. *Genes Dev.* 14, 1269–1278.

Brown, A. V., and Hudson, K. A. (2015). Developmental profiling of gene expression in soybean trifoliate leaves and cotyledons. *BMC Plant Biol.* 15:169. doi: 10.1186/s12870-015-0553-y

Brown, A. V., Conners, S. I., Huang, W., Wilkey, A. P., Grant, D., Weeks, N. T., et al. (2021). A new decade and new data at SoyBase, the USDA-ARS soybean genetics and genomics database. *Nucleic Acids Res.* 49, D1496–D1501. doi: 10.1093/nar/gkaa1107

Carvalho, B. S., and Irizarry, R. A. (2010). A framework for oligonucleotide microarray preprocessing. *Bioinformatics* 26, 2363–2367. doi: 10.1093/bioinformatics/btq431

Ceriani, L., and Verme, P. (2012). The origins of the Gini index: extracts from VariabilitA e MutabilitA (1912) by Corrado Gini. *J. Econ. Inequality* 10, 421–443. doi: 10.1007/s10888-011-9188-x

Chen, T., Zhu, H., Ke, D., Cai, K., Wang, C., Gou, H., et al. (2012). A MAP kinase kinase interacts with SymRK and regulates nodule organogenesis in *Lotus japonicus*. *Plant Cell* 24, 823–838. doi: 10.1105/tpc.112.095984

Clauset, A., Newman, M. E. J., and Moore, C. (2004). Finding community structure in very large networks. *Phys. Rev. E* 70:06611.

Cortijo, S., Aydin, Z., Ahnert, S., and Locke, J. C. (2019). Widespread inter-individual gene expression variability in *Arabidopsis thaliana*. *Mol. Syst. Biol.* 15:e8591. doi: 10.15252/msb.20188591

Cui, X. X., Yan, Q., Gan, S. P., Xue, D., Wang, H. T., Xing, H., et al. (2019). GmWRKY40, a member of the WRKY transcription factor genes identified from *Glycine max* L., enhanced the resistance to *Phytophthora sojae*. *BMC Plant Biol.* 19:598. doi: 10.1186/s12870-019-2132-0

Dincer, A. B., Janizek, J. D., and Lee, S. I. (2020). Adversarial deconfounding autoencoder for learning robust gene expression embeddings. *Bioinformatics* 36, I573–I582. doi: 10.1093/bioinformatics/btaa796

Ding, J. R., Condon, A., and Shah, S. P. (2018). Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. *Nat. Commun.* 9:2002.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. doi: 10.1093/bioinformatics/bts635

Du, J. C., Jia, P. L., Dai, Y. L., Tao, C., Zhao, Z. M., and Zhi, D. G. (2019). Gene2vec: distributed representation of genes based on co-expression. *BMC Genomics* 20:82. doi: 10.1186/s12864-018-5370-x

Elhady, A., Hallmann, J., and Heuer, H. (2020). Symbiosis of soybean with nitrogen fixing bacteria affected by root lesion nematodes in a density-dependent manner. *Sci. Rep.* 10:1619.

Ezer, D., Shepherd, S. J. K., Brestovitsky, A., Dickinson, P., Cortijo, S., Charoensawan, V., et al. (2017). The G-Box transcriptional regulatory code in *Arabidopsis*. *Plant Physiol.* 175, 628–640. doi: 10.1104/pp.17.01086

Fan, C. M., Wang, X., Wang, Y. W., Hu, R. B., Zhang, X. M., Chen, J. X., et al. (2013). Genome-wide expression analysis of soybean MADS genes showing potential function in the seed development. *PLoS One* 8:e62288. doi: 10.1371/journal.pone.0062288

Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. (2004). AFFY – analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* 20, 307–315. doi: 10.1093/bioinformatics/btg405

Gharaibeh, R. Z., Fodor, A. A., and Gibas, C. J. (2008). Background correction using dinucleotide affinities improves the performance of GCRMA. *BMC Bioinformatics* 9:452. doi: 10.1186/1471-2105-9-452

Gordon, A. J., Minchin, F. R., James, C. L., and Komina, O. (1999). Sucrose synthase in legume nodules is essential for nitrogen fixation. *Plant Physiol.* 120, 867–877. doi: 10.1104/pp.120.3.867

Grant, D., Nelson, R. T., Cannon, S. B., and Shoemaker, R. C. (2010). SoyBase, the USDA-ARS soybean genetics and genomics database. *Nucleic Acids Res.* 38, D843–D846. doi: 10.1093/nar/gkp798

Guenther, J. F., and Roberts, D. M. (2000). Water-selective and multifunctional aquaporins from *Lotus japonicus* nodules. *Planta* 210, 741–748. doi: 10.1007/s004250050675

Gupta, A., Wang, H. H., and Ganapathiraju, M. (2015). "Learning structure in gene expression data using deep architectures, with an application to gene clustering," in *Proceedings 2015 IEEE International Conference on Bioinformatics and Biomedicine*, Washington, DC, 1328–1335.

Gupta, O. P., Nigam, D., Dahuja, A., Kumar, S., Vinutha, T., Sachdev, A., et al. (2017). Regulation of isoflavone biosynthesis by miRNAs in two contrasting soybean genotypes at different seed developmental stages. *Front. Plant Sci.* 8:567. doi: 10.3389/fpls.2017.00567

Haibe-Kains, B., El-Hachem, N., Birkbak, N. J., Jin, A. C., Beck, A. H., Aerts, H. J. W. L., et al. (2013). Inconsistency in large pharmacogenomic studies. *Nature* 504, 389–393. doi: 10.1038/nature12831

Huang, J., Zheng, J., Yuan, H., and Mcginnis, K. (2018). Distinct tissue-specific transcriptional regulation revealed by gene regulatory networks in maize. *BMC Plant Biol.* 18:111. doi: 10.1186/s12870-018-1329-y

Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., and Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. *PLoS One* 5:e12776. doi: 10.1371/journal.pone.0012776

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., et al. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4, 249–264. doi: 10.1093/biostatistics/4.2.249

Jin, J., Tian, F., Yang, D. C., Meng, Y. Q., Kong, L., Luo, J., et al. (2017). PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Res.* 45, D1040–D1045. doi: 10.1093/nar/gkw982

Johnson, W. E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8, 118–127. doi: 10.1093/biostatistics/kxj037

Jones, S. I., and Vodkin, L. O. (2013). Using RNA-Seq to profile soybean seed development from fertilization to maturity. *PLoS One* 8:e59270. doi: 10.1371/journal.pone.0059270

Joshi, T., Fitzpatrick, M. R., Chen, S., Liu, Y., Zhang, H., Endacott, R. Z., et al. (2014). Soybean knowledge base (SoyKB): a web resource for integration of soybean translational genomics and molecular breeding. *Nucleic Acids Res.* 42, D1245–D1252. doi: 10.1093/nar/gkt905

Joshi, T., Patil, K., Fitzpatrick, M. R., Franklin, L. D., Yao, Q., Cook, J. R., et al. (2012). Soybean Knowledge Base (SoyKB): a web resource for soybean translational genomics. *BMC Genomics* 13 Suppl 1:S15. doi: 10.1186/1471-2164-13-S1-S15

Joshi, T., Wang, J., Zhang, H., Chen, S., Zeng, S., Xu, B., et al. (2017). The evolution of soybean knowledge base (SoyKB). *Methods Mol. Biol.*1533, 149–159. doi: 10.1007/978-1-4939-6658-5_7

Kim, E., Hwang, S., and Lee, I. (2017). SoyNet: a database of co-functional networks for soybean Glycine max. *Nucleic Acids Res.* 45, D1082–D1089. doi: 10.1093/nar/gkw704

Kim, Y. C., Kim, S. Y., Choi, D., Ryu, C. M., and Park, J. M. (2008). Molecular characterization of a pepper C2 domain-containing SRC2 protein implicated in resistance against host and non-host pathogens and abiotic stresses. *Planta* 227, 1169–1179. doi: 10.1007/s00425-007-0680-2

Kinalis, S., Nielsen, F. C., Winther, O., and Bagger, F. O. (2019). Deconvolution of autoencoders to learn biological regulatory modules from single cell mRNA sequencing data. *BMC Bioinformatics* 20:379. doi: 10.1186/s12859-019-2952-9

Konishi, M., and Yanagisawa, S. (2013). *Arabidopsis* NIN-like transcription factors have a central role in nitrate signalling. *Nat. Commun.* 4:1617. doi: 10.1038/ncomms2621

Lara, P., Oñate-Sánchez, L., Abraham, Z., Ferrándiz, C., Díaz, I., Carbonero, P., et al. (2003). Synergistic activation of seed storage protein gene expression in *Arabidopsis* by ABI3 and two bZIPs related to OPAQUE2. *J. Biol. Chem.* 278, 21003–21011. doi: 10.1074/jbc.M210538200

Lazar, C., Meganck, S., Taminau, J., Steenhoff, D., Coletta, A., Molter, C., et al. (2013). Batch effect removal methods for microarray gene expression data integration: a survey. *Brief Bioinform.* 14, 469–490. doi: 10.1093/bib/bbs037

Leinonen, R., Sugawara, H., Shumway, M., and International Nucleotide Sequence Database Collaboration (2011). The sequence read archive. *Nucleic Acids Res.* 39, D19–D21. doi: 10.1093/nar/gkq1019

Li, P., Zhang, B., Su, T. B., Li, P. R., Xin, X. Y., Wang, W. H., et al. (2018). BrLAS, a GRAS Transcription factor from brassica rapa, is involved in drought stress tolerance in transgenic *Arabidopsis*. *Front. Plant Sci.* 9:1792. doi: 10.3389/fpls.2018.01792

Li, Q. T., Lu, X., Song, Q. X., Chen, H. W., Wei, W., Tao, J. J., et al. (2017). Selection for a zinc-finger protein contributes to seed oil increase during soybean domestication. *Plant Physiol.* 173, 2208–2224. doi: 10.1104/pp.16.01610

Li, W. V., and Li, J. J. (2018). Modeling and analysis of RNA-seq data: a review from a statistical perspective. *Quant. Biol.* 6, 195–209. doi: 10.1007/s40484-018-0144-7

Li, X. J., Wang, K., Lyu, Y. F., Pan, H. Z., Zhang, J. X., Stambolian, D., et al. (2020). Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis. *Nat. Commun.* 11:2338. doi: 10.1038/s41467-020-15851-3

Libault, M., Farmer, A., Joshi, T., Takahashi, K., Langley, R. J., Franklin, L. D., et al. (2010). An integrated transcriptome atlas of the crop model Glycine max, and its use in comparative analyses in plants. *Plant J.* 63, 86–99. doi: 10.1111/j.1365-313X.2010.04222.x

Libault, M., Joshi, T., Takahashi, K., Hurley-Sommer, A., Puricelli, K., Blake, S., et al. (2009). Large-scale analysis of putative soybean regulatory gene expression identifies a Myb gene involved in soybean nodule development. *Plant Physiol.* 151, 1207–1220. doi: 10.1104/pp.109.144030

Lin, C., Jain, S., Kim, H., and Bar-Joseph, Z. (2017). Using neural networks for reducing the dimensions of single-cell RNA-Seq data. *Nucleic Acids Res.* 45:e156. doi: 10.1093/nar/gkx681

Liu, X., Yu, X. P., Zack, D. J., Zhu, H., and Qian, J. (2008). TiGER: A database for tissue-specific gene expression and regulation. *BMC Bioinformatics* 9:271. doi: 10.1186/1471-2105-9-271

Liu, Z. Z., Yao, D., Zhang, J., Li, Z. L., Ma, J., Liu, S. Y., et al. (2015). Identification of genes associated with the increased number of four-seed pods in soybean (*Glycine max* L.) using transcriptome analysis. *Genet. Mol. Res.* 14, 18895–18912. doi: 10.4238/2015.December.28.39

Lohar, D. P., Haridas, S., Gantt, J. S., and Vandenbosch, K. A. (2007). A transient decrease in reactive oxygen species in roots leads to root hair deformation in the legume-rhizobia symbiosis. *New Phytol.* 173, 39–49. doi: 10.1111/j.1469-8137.2006.01901.x

Luo, J., Schumacher, M., Scherer, A., Sanoudou, D., Megherbi, D., Davison, T., et al. (2010). A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data. *Pharmacogenomics J.* 10, 278–291. doi: 10.1038/tpj.2010.57

Lv, G. Y., Guo, X. G., Xie, L. P., Xie, C. G., Zhang, X. H., Yang, Y., et al. (2017). Molecular characterization, gene evolution, and expression analysis of the fructose-1, 6-bisphosphate Aldolase (FBA) gene family in wheat (*Triticum aestivum* L.). *Front. Plant Sci.* 8:1030. doi: 10.3389/fpls.2017.01030

Machado, F. B., Moharana, K. C., Almeida-Silva, F., Gazara, R. K., Pedrosa-Silva, F., Coelho, F. S., et al. (2020). Systematic analysis of 1298 RNA-Seq samples and construction of a comprehensive soybean (*Glycine max*) expression atlas. *Plant J.* 103, 1894–1909. doi: 10.1111/tpj.14850

Marbach, D., Costello, J. C., Kuffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., et al. (2012). Wisdom of crowds for robust gene network inference. *Nat. Methods* 9, 796–804. doi: 10.1038/Nmeth.2016

Marcker, A., Lund, M., Jensen, E. O., and Marcker, K. A. (1984). Transcription of the soybean leghemoglobin genes during nodule development. *EMBO J.* 3, 1691–1695.

Mckenzie, A. T., Katsyv, I., Song, W. M., Wang, M., and Zhang, B. (2016). DGCA: A comprehensive R package for differential gene correlation analysis. *BMC Syst. Biol.* 10:106. doi: 10.1186/s12918-016-0349-1

Moisseyev, G., Park, K., Cui, A., Freitas, D., Rajagopal, D., Konda, A. R., et al. (2020). RGPDB: database of root-associated genes and promoters in maize, soybean, and sorghum. *Database (Oxford)* 2020:baaa038. doi: 10.1093/database/baaa038

Nagae, M., Parniske, M., Kawaguchi, M., and Takeda, N. (2016). The thiamine biosynthesis gene THI1 promotes nodule growth and seed maturation. *Plant Physiol.* 172, 2033–2043. doi: 10.1104/pp.16.01254

Pucciariello, C., Boscari, A., Tagliani, A., Brouquisse, R., and Perata, P. (2019). Exploring legume-rhizobia symbiotic models for waterlogging tolerance. *Front. Plant Sci.* 10:578. doi: 10.3389/fpls.2019.00578

Qi, Z., Zhang, Z., Wang, Z., Yu, J., Qin, H., Mao, X., et al. (2018). Meta-analysis and transcriptome profiling reveal hub genes for soybean seed storage composition during seed development. *Plant Cell Environ.* 41, 2109–2127. doi: 10.1111/pce.13175

Radutoiu, S., Madsen, L. H., Madsen, E. B., Felle, H. H., Umehara, Y., Gronlund, M., et al. (2003). Plant recognition of symbiotic bacteria requires two LysM receptor-like kinases. *Nature* 425, 585–592. doi: 10.1038/nature02039

Roy, S., Liu, W., Nandety, R. S., Crook, A., Mysore, K. S., Pislariu, C. I., et al. (2020). Celebrating 20 years of genetic discoveries in legume nodulation and symbiotic nitrogen fixation([OPEN]). *Plant Cell* 32, 15–41. doi: 10.1105/tpc.19.00279

Seabra, A. R., and Carvalho, H. G. (2015). Glutamine synthetase in Medicago truncatula, unveiling new secrets of a very old enzyme. *Front. Plant Sci.* 6:578. doi: 10.3389/fpls.2015.00578

Sean, D., and Meltzer, P. S. (2007). GEOquery: a bridge between the gene expression omnibus (GEO) and BioConductor. *Bioinformatics* 23, 1846–1847. doi: 10.1093/bioinformatics/btm254

Senovilla, M., Castro-Rodriguez, R., Abreu, I., Escudero, V., Kryvoruchko, I., Udvardi, M. K., et al. (2018). Medicago truncatula copper transporter 1 (MtCOPT1) delivers copper for symbiotic nitrogen fixation. *New Phytol.* 218, 696–709. doi: 10.1111/nph.14992

Severin, A. J., Woody, J. L., Bolon, Y. T., Joseph, B., Diers, B. W., Farmer, A. D., et al. (2010). RNA-Seq atlas of glycine max: a guide to the soybean transcriptome. *BMC Plant Biol.* 10:160. doi: 10.1186/1471-2229-10-160

Sims, A. H., Smethurst, G. J., Hey, Y., Okoniewski, M. J., Pepper, S. D., Howell, A., et al. (2008). The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets - improving meta-analysis and prediction of prognosis. *BMC Med. Genomics* 1:42. doi: 10.1186/1755-8794-1-42

Sinharoy, S., Torres-Jerez, I., Bandyopadhyay, K., Kereszt, A., Pislariu, C. I., Nakashima, J., et al. (2013). The C2H2 transcription factor regulator of symbiosome differentiation represses transcription of the secretory pathway gene VAMP721a and promotes symbiosome development in *Medicago truncatula*. *Plant Cell* 25, 3584–3601. doi: 10.1105/tpc.113.114017

Sonawane, A. R., Platig, J., Fagny, M., Chen, C. Y., Paulson, J. N., Lopes-Ramos, C. M., et al. (2017). Understanding tissue-specific gene regulation. *Cell Rep.* 21, 1077–1088. doi: 10.1016/j.celrep.2017.10.001

Streeter, J. G. (1981). Effect of nitrate in the rooting medium on carbohydrate composition of soybean nodules. *Plant Physiol.* 68, 840–844. doi: 10.1104/pp.68.4.840

Sun, S., Yi, C., Ma, J., Wang, S., Peirats-Llobet, M., Lewsey, M. G., et al. (2020). Analysis of spatio-temporal transcriptome profiles of soybean (*Glycine max*) tissues during early seed development. *Int. J. Mol. Sci.* 21:7603. doi: 10.3390/ijms21207603

Van Heerden, P. D., Kiddle, G., Pellny, T. K., Mokwala, P. W., Jordaan, A., Strauss, A. J., et al. (2008). Regulation of respiration and the oxygen diffusion barrier in soybean protect symbiotic nitrogen fixation from chilling-induced inhibition and shoots from premature senescence. *Plant Physiol.* 148, 316–327. doi: 10.1104/pp.108.123422

Vernie, T., Moreau, S., De Billy, F., Plet, J., Combier, J. P., Rogers, C., et al. (2008). EFD Is an ERF transcription factor involved in the control of nodule number and differentiation in *Medicago truncatula*. *Plant Cell* 20, 2696–2713. doi: 10.1105/tpc.108.059857

Wagner, G. P., Kin, K., and Lynch, V. J. (2012). Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.* 131, 281–285. doi: 10.1007/s12064-012-0162-3

Walley, J. W., Sartor, R. C., Shen, Z., Schmitz, R. J., Wu, K. J., Urich, M. A., et al. (2016). Integration of omic networks in a developmental atlas of maize. *Science* 353, 814–818. doi: 10.1126/science.aag1125

Wang, J., Hossain, M. S., Lyu, Z., Schmutz, J., Stacey, G., Xu, D., et al. (2019). SoyCSN: Soybean context-specific network analysis and prediction based on tissue-specific transcriptome data. *Plant Direct* 3:e00167. doi: 10.1002/pld3.167

Wang, T. T., Yu, T. F., Fu, J. D., Su, H. G., Chen, J., Zhou, Y. B., et al. (2020). Genome-wide analysis of the GRAS gene family and functional identification of GmGRAS37 in drought and salt tolerance. *Front. Plant Sci.* 11:604690. doi: 10.3389/fpls.2020.604690

Wang, Z. H., Yan, L. Y., Wan, L. Y., Huai, D. X., Kang, Y. P., Shi, L., et al. (2019). Genome-wide systematic characterization of bZIP transcription factors and their expression profiles during seed development and in response to salt stress in peanut. *BMC Genomics* 20:51. doi: 10.1186/s12864-019-5434-6

Wingett, S. W., and Andrews, S. (2018). FastQ screen: a tool for multi-genome mapping and quality control. *F1000Res* 7:1338. doi: 10.12688/f1000research.15931.2

Wu, Z. H., Wang, M. R., Yang, S. Y., Chen, S. C., Chen, X., Liu, C., et al. (2019). A global coexpression network of soybean genes gives insights into the evolution of nodulation in nonlegumes and legumes. *New Phytol.* 223, 2104–2119. doi: 10.1111/nph.15845

Xia, J., Fjell, C. D., Mayer, M. L., Pena, O. M., Wishart, D. S., and Hancock, R. E. (2013). INMEX–a web-based tool for integrative meta-analysis of expression data. *Nucleic Acids Res.* 41, W63–W70. doi: 10.1093/nar/gkt338

Xiao, S. J., Zhang, C., Zou, Q., and Ji, Z. L. (2010). TiSGeD: a database for tissue-specific genes. *Bioinformatics* 26, 1273–1275. doi: 10.1093/bioinformatics/btq109

Xie, R., Wen, J., Quitadamo, A., Cheng, J. L., and Shi, X. H. (2017). A deep auto-encoder model for gene expression prediction. *BMC Genomics* 18:845. doi: 10.1186/s12864-017-4226-0

Yan, Z., Hossain, M. S., Arikit, S., Valdes-Lopez, O., Zhai, J. X., Wang, J., et al. (2015). Identification of microRNAs and their mRNA targets during soybean nodule development: functional analysis of the role of miR393j-3p in soybean nodulation. *New Phytol.* 207, 748–759. doi: 10.1111/nph.13365

Yanai, I., Benjamin, H., Shmoish, M., Chalifa-Caspi, V., Shklar, M., Ophir, R., et al. (2005). Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* 21, 650–659. doi: 10.1093/bioinformatics/bti042

Yang, Y., Zhou, Y., Chi, Y. J., Fan, B. F., and Che, Z. X. (2017). Characterization of soybean WRKY gene family and identification of soybean WRKY genes that promote resistance to soybean cyst nematode. *Sci. Rep.* 7:17804. doi: 10.1038/s41598-017-18235-8

Yi, F., Gu, W., Chen, J., Song, N., Gao, X., Zhang, X., et al. (2019). High temporal-resolution transcriptome landscape of early maize seed development. *Plant Cell* 31, 974–992. doi: 10.1105/tpc.18.00961

Yi, H. C., You, Z. H., Huang, D. S., Li, X., Jiang, T. H., and Li, L. P. (2018). A deep learning framework for robust and accurate prediction of ncRNA-protein interactions using evolutionary information. *Mol. Therapy Nucleic Acids* 11, 337–344. doi: 10.1016/j.omtn.2018.03.001

Yu, X. P., Lin, J., Zack, D. J., and Qian, J. A. (2006). Computational analysis of tissue-specific combinatorial gene regulation: predicting interaction between transcription factors in human tissues. *Nucleic Acids Res.* 34, 4925–4936. doi: 10.1093/nar/gkl595

Yuan, S. L., Li, R., Chen, H. F., Zhang, C. J., Chen, L. M., Hao, Q. N., et al. (2017). RNA-Seq analysis of nodule development at five different developmental stages of soybean (*Glycine max*) inoculated with Bradyrhizobium japonicum strain 113-2. *Sci. Rep.* 7:42248. doi: 10.1038/srep42248

Zhang, Y., Parmigiani, G., and Johnson, W. E. (2020). ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genom. Bioinform.* 2:lqaa078. doi: 10.1093/nargab/lqaa078

Zhou, P., Li, Z., Magnusson, E., Gomez Cano, F., Crisp, P. A., Noshay, J. M., et al. (2020). Meta gene regulatory networks in maize highlight functionally relevant regulatory interactions. *Plant Cell* 32, 1377–1396. doi: 10.1105/tpc.20.00080

Zilli, C. G., Cruz, D. M. S., Polizio, A. H., Tomaro, M. L., and Balestrasse, K. B. (2011). Symbiotic association between soybean plants and *Bradyrhizobium japonicum* develops oxidative stress and heme oxygenase-1 induction at early stages. *Redox Rep.* 16, 49–55. doi: 10.1179/174329211x13020951739811

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.