The state of the art in soybean transcriptomics resources and gene coexpression networks

Fabricio Almeida-Silva¹, Kanhu C. Moharana¹, Thiago M. Venancio¹*

¹Laboratório de Química e Função de Proteínas e Peptídeos, Centro de Biociências e Biotecnologia, Universidade Estadual do Norte Fluminense Darcy Ribeiro, Campos dos Goytacazes, RJ, Brazil.

*TMV: Laboratório de Química e Função de Proteínas e Peptídeos, Centro de Biociências e Biotecnologia, Universidade Estadual do Norte Fluminense Darcy Ribeiro. Av. Alberto Lamego 2000, P5, sala 217, Campos dos Goytacazes, RJ, Brazil. Email: thiago.venancio@gmail.com
ABSTRACT
In the past decade, over 3000 samples of soybean transcriptomic data have accumulated in public repositories. Here, we review the state of the art in soybean transcriptomics, highlighting the major microarray and RNA-seq studies that investigated soybean transcriptional programs in different tissues and conditions. Further, we propose approaches for integrating such big data using gene coexpression network and outline important web resources that may facilitate soybean data acquisition and analysis, contributing to the acceleration of soybean breeding and functional genomics research.

Keywords: legumes, functional genomics, gene expression, hub genes, modules.
Introduction

Soybean [Glycine max (L.) Merr.] is a major economic crop and is an important source of edible protein and oil. Besides, its ability to host symbiotic diazotrophic bacteria contributes to sustainable agricultural practices by reducing the use of chemical fertilizers because of bacteria-mediated fixation of atmospheric nitrogen (N₂) (Hungria et al. 2015). An important milestone in soybean genomics-based research was the availability of high-throughput genomic and transcriptomic sequencing data, which can help identify and characterize genes and pathways underlying plant growth and development. Over the past years, there has been a significant increase in soybean transcriptomic studies (Libault et al. 2010a; Severin et al. 2010; Garg and Jain 2013a; O’Rourke et al. 2017), which explored the transcriptional landscape in several tissues (Irsigler et al. 2007; Libault et al. 2010b; Severin et al. 2010; Liu et al. 2018), developmental conditions (Lu et al. 2016; Gazara et al. 2019; Yang et al. 2019), and genotypes (Goettel et al. 2014; Prince et al. 2015). This short review aims to summarize major transcriptomic studies and discuss recent advances in coexpression network analyses for soybean, outlining important web resources to facilitate data integration.

The major transcriptomic datasets available in soybean were generated by hybridization-based (i.e., microarrays) and sequence-based (i.e., RNA-seq) methods. Although microarrays are relatively easy to conduct and inexpensive, dependence on oligo probes limits the identification of novel transcripts. In addition, there are technical limitations that reduce reproducibility of gene expression estimates, such as incorrect probe-to-gene assignments (Dai et al. 2005), non-specific probe hybridization (Kroll et al. 2008), and low accuracy and precision (Draghi et al. 2006; Jaksik et al. 2015). On the other hand, RNA-seq uses second and third generation sequencing methods to estimate transcriptional profiles independent of genomic information (Lee et al. 2002). In addition to the numerous technical advantages, absolute transcript levels from RNA-seq are also more correlated with the proteomic profiles (Fu et al. 2009). Thus, this method is considered the best technique available for transcriptomics studies, particularly in non-model species (Garg and Jain 2013b).

Microarray-based studies

Initially (i.e., prior to 2010), soybean microarray studies led to a considerable accumulation of data in public repositories such as Gene Expression Omnibus (GEO). Many of those studies explored differential gene expression during pathogen attack (Moy et al. 2004; Zou et al. 2005; Ithal et al. 2007; Li et al. 2008). Besides, expression profiles in nodules have been investigated to understand the complex mechanism of symbionts survival in the host plant (Brechenmacher et al. 2008). Thibaud-Nissen et al. reported important transcript changes associated with somatic embryogenesis in soybean that could be useful for tissue culture (Thibaud-Nissen et al. 2003), while Hudson explored the role of circadian clock in gene expression regulation and identified important circadian clock-controlled genes (Hudson 2010). A list of important soybean transcriptomic studies using microarrays are shown in Table 1.
RNA-seq-based studies

Due to its many advantages (Fu et al. 2009; Jaksik et al. 2015), RNA-seq quickly replaced microarrays as the method of choice for transcriptomics. To date, several studies have explored the spatiotemporal changes occurring in various soybean tissues using RNA-seq. The two first studies were reported by Libault et al., who sequenced cDNA obtained from 14 tissues and provided experimental support for soybean gene annotations (Libault et al. 2010b) and; Severin et al., who investigated the transcriptomes from seven tissues and seven seed developmental stages and observed that tissues were grouped in three large clusters with underground, aerial and seed parts samples (Severin et al. 2010). While these studies focused on genome annotation and major differential gene expression patterns across a broad range of tissues, many other research groups targeted important specialized tissues or organs (e.g., seeds).

Seeds are the primary source of nutrient reserves for germination and early seedling growth. Jones and Vodkin examined the expression changes of soybean seed development from four days post-fertilization to the mature seed (Jones and Vodkin 2013). In another study, Goettel et al. sequenced the transcriptomes of soybean seeds from nine lines varying in oil composition and content and reported transcript nucleotide polymorphisms associated with oil quality traits (Goettel et al. 2014). In an original approach, Lambirth et al. investigated the physiological effects of expressing some recombinant proteins at high levels in soybean, such as human thyroglobulin and myelin basic protein (Lambirth et al. 2015).

In addition to the cotyledons, the embryonic axis plays key regulatory roles during seed germination and early seedling growth. Bellieny-Rabelo et al. explored the transcriptional landscape of soybean embryonic axes during germination and described the activation of several biochemical pathways that are critical for germination, such as lipid mobilization and glyoxylate cycle (Bellieny-Rabelo et al. 2016). To understand the role of the key hormone gibberellin (GA) in regulating seed germination, Gazara et al. analyzed the transcriptomes of embryonic axes from seeds germinating in the presence of paclobutrazol, a GA biosynthesis inhibitor (Gazara et al. 2019).

The development of soybean cultivars capable of tolerating diverse biotic and abiotic stresses is a major challenge, as plant responses to such stresses involve many complex pathways. Hence, transcriptome profiling provides an opportunity to identify genes involved in stress tolerance (Deshmukh et al. 2014). Transcription factors (TFs) often play a critical role in regulating stress response. Belamkar et al. characterized the expression of HD-Zip TFs under dehydration and salt stress (Belamkar et al. 2014). In another experiment, Prince et al. compared the transcriptomes of two lines contrasting for drought tolerance and found several genes associated with biotic stress that were up-regulated in the drought-tolerant genotype (Prince et al. 2015). Further, this study also identified genomic variations in aquaporin genes of the drought-tolerant line that could facilitate marker-assisted selection in soybeans with improved drought tolerance.
From genes to networks: network modeling from transcriptomic data

The large number of soybean gene expression studies has resulted in the accumulation of a huge amount of data on public repositories, offering opportunities to perform more integrative and complex analyses, such as the construction of gene coexpression networks (GCNs). In GCNs, nodes represent genes and edges represent the strength of association between gene pairs. GCN-based studies can be used to extract biological signals from the system as a whole and can often identify genes that are important regulators of various biological processes, but individually lack high statistical significance or have small expression changes in differential gene expression analyses (Elo et al. 2007).

GCN inference consists in identifying densely connected gene clusters using similarity measures and clustering algorithms. The gene-gene similarity measures can be correlation-based, such as Pearson Correlation Coefficient (PCC, parametric) and Spearman Correlation Coefficient (SCC, non-parametric), or rank-based, such as Highest Reciprocal Rank (HRR), Mutual Information (MI) and Partial Correlations (PC) (Liesecke et al. 2018). From an $n \times n$ similarity matrix, various different clustering algorithms (e.g., hierarchical clustering and graph-based clustering) can identify coexpression modules (i.e., groups of coexpressed genes). In plant biology, WGCNA (Weighted Gene Coexpression Network Analysis) (Langfelder and Horvath 2008) has become the most popular algorithm, and this trend also stands true for soybean GCNs (Table 2). This algorithm calculates gene-gene expression similarities from an adjacency matrix (PCC or SCC raised to a power $\beta$ to amplify disparities) and attributes association strengths between gene pairs using both correlation coefficients and shared neighbors. Module detection is then performed using dynamic tree cutting, which identifies gene clusters from hierarchical clustering dendrograms (Langfelder and Horvath 2008).

Further, GCNs can be integrated with TF-DNA interactions to reconstruct gene regulatory networks (GRNs), which depict interactions between TFs and their target genes. This approach assumes that coexpressed genes are co-regulated and, thus, are more likely to share a conserved motif (i.e., TF binding site) in their promoter sequences (Ko and Brandizzi 2020). From a coexpression module, GRNs can be inferred with two main approaches: (i) mapping known TF motifs to the promoter sequences of its genes via enrichment methods, or (ii) de novo motif detection followed by TF mapping (Bailey et al. 2009; Ko and Brandizzi 2020). Information on each TF’s motifs can be obtained by ChIP-seq (chromatin immunoprecipitation followed by deep sequencing), but this approach can identify motifs of one TF at a time, limiting its application in a genome-wide scale (Kulkarni and Vandepoele 2020). Besides, less than 20 ChIP-Seq experiments have been performed for soybean genes, what reduces the accuracy of soybean GRN modelling.

Currently, the major bottleneck of GCN inference is the lack of a standard pipeline to process data and model networks, especially for large heterogenous data sets. This can lead to a considerably high number of false positives that can bias biological interpretations in highly heterogenous data sets. Despite the challenges, the number of studies reporting soybean GCNs has increased significantly (Table 2). Although only a few soybean GCN studies have been conducted, this is certainly a promising approach to explore the vast...
amount of accumulated RNA-seq data in the coming years. GCNs can be used to answer many biological questions, and the most prominent of them are discussed below.

**Gene coexpression networks for the inference of gene function**

As a consequence of extensive gene duplications, inferring gene function in plants exclusively through homology mapping is extremely challenging, as paralogs often lack functional conservation (Mutwil et al. 2011). GCNs can help address this issue by deploying guilt-by-association (GBA) approaches, which assume that functionally related genes tend to be co-regulated and, hence, coexpressed. Almeida-Silva et al. have recently predicted the function of 93 hub genes without functional description in soybean by using GBA from a global soybean GCN (Almeida-Silva et al. 2020). Likewise, Mutwil et al. proposed an integration of sequence data with GCN comparison across seven plant species to transfer knowledge of gene function from one species to another (Mutwil et al. 2011). As a small fraction of soybean genes have had their functions experimentally characterized, GCN-based function inference can be used complementarily to other approaches to accelerate soybean functional genomics approaches.

**Gene coexpression networks for evolutionary analyses**

The use of GCNs to study the evolution of genes and genomes has gained popularity over the past few years. In this sense, Almeida-Silva et al. analyzed the distribution of paralog pairs across modules of a global soybean GCN and observed that most of the paralogous genes tend to diverge at the transcriptional level after any mode of gene duplication, but particularly upon local (e.g., tandem or proximal) duplication events (Almeida-Silva et al. 2020). From a global microarray-based GCN, Wu et al. identified a nodulation-related module and compared its genes with their homologs in other species, such as *Medicago truncatula*, *Phaseolus vulgaris* and *Lotus japonicus* (Wu et al. 2019). Wisecaver et al. integrated coexpression data for several species and identified gene clusters likely involved in the biosynthesis of specialized metabolites (Wisecaver et al. 2017). Yu et al. inferred GCNs for 4 species and identified conserved and species-specific modules, as well as their functions based on enrichment analyses (Yu et al. 2017). Lu et al. analyzed genes involved in seed weight and oil content in wild and cultivated soybean and identified different connectivity patterns between the two species, suggesting a transcriptional rewiring during soybean domestication (Lu et al. 2016).

**Gene coexpression networks uncover genes likely involved in processes of agronomic relevance**

The most common use of GCN aims at identifying genes involved in a biological process or pathway of interest. For that, researchers rely on two main strategies. The first one consists in using guide genes, which are genes known to be involved in the process, most often with reliable experimental support. Using a guide-directed approach, Bian et al. constructed a GCN of differentially expressed TFs under salt stress and identified a hub TFs involved in
stress tolerance (Bian et al. 2020). Lima et al. identified TFs potentially related to seed longevity (Lima et al. 2017), while Liu et al. identified hubs that are strongly connected to genes differentially expressed under salt stress, several of which are involved in phytohormone signaling (Liu et al. 2018). Dias et al. identified genes that are coexpressed with previously reported drought stress-related genes (Dias et al. 2016).

The second approach consists in calculating module-trait correlations, which can identify gene clusters associated with a particular phenotype. Wu et al. identified a nodulation-related module whose central genes were experimentally validated as critical for nodulation (Wu et al. 2019). Yang et al. found a module whose genes are likely involved in oil accumulation in seeds (Yang et al. 2019). Similarly, Qi et al. identified hub genes related to seed oil composition (Qi et al. 2018). Du et al. identified hub genes involved in seed set and size (Du et al. 2017). Wisecaver et al. found groups of genes putatively related to the biosynthesis of specialized metabolites in four species (Wisecaver et al. 2017). Almeida-Silva et al. identified pathways that are enhanced or inhibited in specific tissues in a global analysis (Almeida-Silva et al. 2020). Lu et al. identified genes that were differentially expressed and differentially connected between early and mid-maturation in seeds, many of which are likely involved in seed weight and oil content (Lu et al. 2016).

**Soybean transcriptomic resources**

There are several available web resources for obtaining soybean transcriptomic data, such as Phytozome (https://phytozome.jgi.doe.gov/pz/portal.html), Soyseq (https://soybase.org/soyseq/), Soybean eFP Browser (http://bar.utoronto.ca/efpsoybean/cgi-bin/efpWeb.cgi) and LegumeInfo (https://legumeinfo.org/lis_expression/all). However, these resources have limited number of samples. To address this issue and facilitate the systematic acquisition and analysis of large-scale soybean transcriptomic data, Machado et al. have created a comprehensive Soybean Expression Atlas that outperforms the pre-existing ones, with 1,298 RNA-seq samples that can be easily explored in a web interface (http://venanciogroup.uenf.br/cgi-bin/qmax_atlas/index.cgi) (Machado et al. 2020). Through this user-friendly interface, users can visualize the expression levels of genes of interest in several soybean tissues as expression maps, heatmaps or bar plots. Additionally, expression data (TPM-normalized or raw read counts) can be downloaded by tissue, PubMed ID or DOI. A short tutorial on how to use the Soybean Expression Atlas for data download and analysis is described in Supporting Information File S1.

Furthermore, web resources containing pre-computed soybean GCNs have also been built. PlaNet (http://aranet.mpimp-golm.mpg.de) can be used to analyze co-functional networks within and across species of photosynthetic organisms (Mutwil et al. 2011). ATTED-II (https://atted.jp) is a coexpression database that was originally developed for Arabidopsis thaliana, but now includes data for 9 plant species, of which 7 have both microarray and RNA-seq data (Obayashi et al. 2018). Finally, a collection of soybean context-specific networks entitled SoyCSN (http://soykb.org/SoyCSN/) have recently been developed and made available for users (Wang et al. 2019). Our group plans to integrate
the recently reported GCN (Almeida-Silva et al. 2020) in the Soybean Expression Atlas described above.

Publicly available soybean expression data and future prospects

With the decreasing cost of sequencing technologies, the number of RNA-seq experiments deposited on NCBI’s Sequence Read Archive (SRA) has been growing exponentially. Not surprisingly, the top 3 countries in number of soybean RNA-seq samples (China, USA and Brazil, respectively) are also among the top 4 global leaders in soybean production (http://soystats.com/international-world-soybean-production/), indicating a correlation between soybean production and investments in soybean research (Fig. 1a). Expectedly, many RNA-seq samples have accumulated on SRA since the Soybean Expression Atlas has been released (Fig. 1b; information from October 2\textsuperscript{nd} 2020). However, the quality of the new samples is unknown, as we are yet to incorporate them in the next update of the Soybean Expression Atlas. For instance, 15% of the samples processed for the first version of the atlas were excluded in the quality check steps of our pipeline.

Interestingly, there was a five- and eight-fold increase in the number of seed and root samples, respectively. New seed samples comprise mostly studies targeting seed oil biosynthesis, suggesting this may be a promising area in soybean genomics-based research. Recently deposited root RNA-seq samples are mostly represented by microbial inoculation experiments (both pathogenic and plant growth-promoting microorganisms), demonstrating increasing global efforts to better understand soybean association with microorganisms. The Soybean Expression Atlas is currently being updated to include these new RNA-seq samples, and we hope to release new versions yearly. This will help accelerate soybean genomics-based research by providing researchers with easy download and analysis of large-scale expression data.

Acknowledgements

This work was supported by Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ; grants E-26/203.309/2016 and E-26/203.014/2018), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES; Finance Code 001), and Conselho Nacional de Desenvolvimento Científico e Tecnológico. The funding agencies had no role in the design of the study and collection, analysis, and interpretation of data and in writing.
REFERENCES

Almeida-Silva F, Moharana KC, Machado FB, and Venancio TM. 2020. Exploring the complexity of soybean (Glycine max) transcriptional regulation using global gene co-expression networks. *Planta* 252: 1–12.

Bailey TL, Boden M, Buske FA, et al. 2009. MEME Suite: Tools for motif discovery and searching. *Nucleic Acids Res* 37: 202–8.

Belamkar V, Weeks NT, Bharti AK, et al. 2014. Comprehensive characterization and RNA-Seq profiling of the HD-Zip transcription factor family in soybean (Glycine max) during dehydration and salt stress. *BMC Genomics* 15: 950.

Bellieny-Rabelo D, Oliveira EAG de, Ribeiro E da S, et al. 2016. Transcriptome analysis uncovers key regulatory and metabolic aspects of soybean embryonic axes during germination. *Sci Rep* 6: 36009.

Bian XH, Li W, Niu CF, et al. 2020. A class B heat shock factor selected for during soybean domestication contributes to salt tolerance by promoting flavonoid biosynthesis. *New Phytol* 225: 268–83.

Brechenmacher L, Kim M-Y, Benitez M, et al. 2008. Transcription Profiling of Soybean Nodulation by Bradyrhizobium japonicum. *Mol Plant-Microbe Interact* 21: 631–45.

Dai M, Wang P, Boyd AD, et al. 2005. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res* 33: 1–9.

Deshmukh R, Sonah H, Patil G, et al. 2014. Integrating omic approaches for abiotic stress tolerance in soybean. *Front Plant Sci* 5.

Dhaubhadel S, Gijzen M, Moy P, and Farhangkhoee M. 2007. Transcriptome Analysis Reveals a Critical Role of CHS7 and CHS8 Genes for Isoflavonoid Synthesis in Soybean Seeds. *Plant Physiol* 143: 326–38.

Dias LP, Oliveira-Busatto LA de, and Bodanese-Zanettini MH. 2016. The differential expression of soybean [Glycine max (L.) Merrill] WRKY genes in response to water deficit. *Plant Physiol Biochem* 107: 288–300.

Draghici S, Khatri P, Eklund AC, and Szallasi Z. 2006. Reliability and reproducibility issues in DNA microarray measurements. *Trends Genet* 22: 101–9.

Du J, Wang S, He C, et al. 2017. Identification of regulatory networks and hub genes controlling soybean seed set and size using RNA sequencing analysis. *J Exp Bot* 68: erw460.

Elo LL, Järvenpää H, Orešič M, et al. 2007. Systematic construction of gene coexpression networks with applications to human T helper cell differentiation process. *Bioinformatics* 23: 2096–103.

Fu X, Fu N, Guo S, et al. 2009. Estimating accuracy of RNA-Seq and microarrays with
proteomics. *BMC Genomics* 10: 1–9.

Garg R and Jain M. 2013a. Transcriptome Analyses in Legumes: A Resource for Functional Genomics. *Plant Genome* 6: 0.

Garg R and Jain M. 2013b. RNA-Seq for Transcriptome Analysis in Non-model Plants. 1069: 43–58.

Gazara RK, Oliveira EAG de, Rodrigues BC, *et al.* 2019. Transcriptional landscape of soybean (*Glycine max*) embryonic axes during germination in the presence of paclobutrazol, a gibberellin biosynthesis inhibitor. *Sci Rep* 9: 1–12.

Goettel W, Xia E, Upchurch R, *et al.* 2014. Identification and characterization of transcript polymorphisms in soybean lines varying in oil composition and content. *BMC Genomics* 15: 299.

Hudson KA. 2010. The Circadian Clock-controlled Transcriptome of Developing Soybean Seeds. *Plant Genome* 3: 3–13.

Hungria M, Nogueira MA, and Araujo RS. 2015. Soybean seed co-inoculation with *Bradyrhizobium* spp. and *Azospirillum brasilense*: a new biotechnological tool to improve yield and sustainability. *Am J Plant Sci* 06: 811–7.

Irsigler AST, Costa MDL, Zhang P, *et al.* 2007. Expression profiling on soybean leaves reveals integration of ER- and osmotic-stress pathways. *BMC Genomics* 8: 431.

Ithal N, Recknor J, Nettleton D, *et al.* 2007. Parallel Genome-Wide Expression Profiling of Host and Pathogen During Soybean Cyst Nematode Infection of Soybean. *Mol Plant-Microbe Interact* 20: 293–305.

Jaksik R, Iwanaszko M, Rzeszowska-Wolny J, and Kimmel M. 2015. Microarray experiments and factors which affect their reliability. *Biol Direct* 10: 46.

Jones SI and Vodkin LO. 2013. Using RNA-Seq to profile soybean seed development from fertilization to maturity. *PLoS One* 8: e59270.

Ko DK and Brandizzi F. 2020. Network-based approaches for understanding gene regulation and function in plants. *Plant J* 104: 302–17.

Kroll KM, Barkema GT, and Carlon E. 2008. Modeling background intensity in DNA microarrays. *Phys Rev E - Stat Nonlinear, Soft Matter Phys* 77: 1–8.

Kulkarni SR and Vandepoele K. 2020. Inference of plant gene regulatory networks using data-driven methods: A practical overview. *Biochim Biophys Acta - Gene Regul Mech* 1863: 194447.

Lambirth KC, Whaley AM, Blakley IC, *et al.* 2015. A Comparison of transgenic and wild type soybean seeds: analysis of transcriptome profiles using RNA-Seq. *BMC Biotechnol* 15: 89.
Langfelder P and Horvath S. 2008. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9: 559.

Lee J-M, Williams ME, Tingey S V, and Rafalski AJ. 2002. DNA array profiling of gene expression changes during maize embryo development. *Funct Integr Genomics* 2: 13–27.

Li Y, Zou J, Li M, *et al.* 2008. Soybean defense responses to the soybean aphid. *New Phytol* 179: 185–95.

Libault M, Farmer A, Brechenmacher L, *et al.* 2010a. Complete Transcriptome of the Soybean Root Hair Cell, a Single-Cell Model, and Its Alteration in Response to Bradyrhizobium japonicum Infection. *Plant Physiol* 152: 541–52.

Libault M, Farmer A, Joshi T, *et al.* 2010b. An integrated transcriptome atlas of the crop model Glycine max, and its use in comparative analyses in plants. *Plant J* 63: 86–99.

Liesecke F, Daudu D, Bernonville RD De, *et al.* 2018. Ranking genome-wide correlation measurements improves microarray and RNA-seq based global and targeted co-expression networks. *Sci Rep* 8: 1–16.

Lima JJP, Buitink J, Lalanne D, *et al.* 2017. Molecular characterization of the acquisition of longevity during seed maturation in soybean. *PLoS One* 12.

Liu A, Xiao Z, Li M, *et al.* 2018. Transcriptomic reprogramming in soybean seedlings under salt stress. *Plant Cell Environ* 42: 98–114.

Lu X, Li QT, Xiong Q, *et al.* 2016. The transcriptomic signature of developing soybean seeds reveals the genetic basis of seed trait adaptation during domestication. *Plant J* 86: 530–44.

Machado FB, Moharana KC, Almeida-Silva F, *et al.* 2020. Systematic analysis of 1298 RNA-Seq samples and construction of a comprehensive soybean (Glycine max) expression atlas. *Plant J*.

Moy P, Qutob D, Chapman BP, *et al.* 2004. Patterns of Gene Expression Upon Infection of Soybean Plants by Phytophthora sojae. *Mol Plant-Microbe Interact* 17: 1051–62.

Mutwil M, Klie S, Tohge T, *et al.* 2011. PlaNet: Combined sequence and expression comparisons across plant networks derived from seven species. *Plant Cell* 23: 895–910.

O’Rourke JA, Charlson D V, Gonzalez DO, *et al.* 2007. Microarray analysis of iron deficiency chlorosis in near-isogenic soybean lines. *BMC Genomics* 8: 476.

O’Rourke JA, Graham MA, and Whitham SA. 2017. Soybean Functional Genomics: Bridging the Genotype-to-Phenotype Gap. : 151–70.

Obayashi T, Aoki Y, Tadaka S, *et al.* 2018. ATTED-II in 2018: A Plant Coexpression Database Based on Investigation of the Statistical Property of the Mutual Rank Index.
Plant Cell Physiol 59: e3.

Prince SJ, Joshi T, Mutava RN, et al. 2015. Comparative analysis of the drought-responsive transcriptome in soybean lines contrasting for canopy wilting. Plant Sci 240: 65–78.

Qi Z, Zhang Z, Wang Z, et al. 2018. Meta-analysis and transcriptome profiling reveal hub genes for soybean seed storage composition during seed development. Plant Cell Environ 41: 2109–27.

Severin AJ, Woody JL, Bolon Y-T, et al. 2010. RNA-Seq Atlas of Glycine max: A guide to the soybean transcriptome. BMC Plant Biol 10: 160.

Thibaud-Nissen F, Shealy RT, Khanna A, and Vodkin LO. 2003. Clustering of microarray data reveals transcript patterns associated with somatic embryogenesis in soybean. Plant Physiol 132: 118–36.

Wang J, Hossain MS, Lyu Z, et al. 2019. SoyCSN: Soybean context-specific network analysis and prediction based on tissue-specific transcriptome data. Plant Direct 3: 1–15.

Wisecaver JH, Borowsky AT, Tzin V, et al. 2017. A Global Coexpression Network Approach for Connecting Genes to Specialized Metabolic Pathways in Plants. Plant Cell 29: 944–59.

Wu Z, Wang M, Yang S, et al. 2019. A global coexpression network of soybean genes gives insights into the evolution of nodulation in nonlegumes and legumes. New Phytol 223: 2104–19.

Yang S, Miao L, He J, et al. 2019. Dynamic transcriptome changes related to oil accumulation in developing soybean seeds. Int J Mol Sci 20.

Yu H, Jiao B, and Liang C. 2017. Systematic analysis of RNA-seq-based gene co-expression. bioRxiv: 1–22.

Zou J, Rodriguez-Zas S, Aldea M, et al. 2005. Expression Profiling Soybean Response to Pseudomonas syringae Reveals New Defense-Related Genes and Rapid HR-Specific Downregulation of Photosynthesis. Mol Plant-Microbe Interact 18: 1161–74.
Figures

Figure 1. Main descriptive statistics of RNA-seq samples on NCBI’s SRA. (A) Distribution of samples by country. Numbers preceding country names indicate their rank in number of samples. Country colors vary according to the log₂ of the number of samples. (B) Number of SRA samples by tissue. For each sample, dark blue bars represent BioSamples that have been analyzed in the Soybean Expression Atlas (http://venanciogroup.uenf.br/cgi-bin/gmax_atlas/index.cgi), while light blue bars represent BioSamples that were deposited after the release of the Soybean Expression Atlas (verified on October 2nd 2020). Some samples were not assigned to any tissue (i.e., “unknown”) because submitters did not include this information in the SRA metadata. Both country and tissue information were extracted from the SRA metadata.
### Tables

**Table 1. Main microarray-based studies in soybean based on number of citations.**

| Reference                                  | Contribution                                                                                                                                 |
|--------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------|
| (Moy et al. 2004)                          | Identified soybean genes that were strongly upregulated during infection with *Phytophthora sojae* and hypothesized that pathogen transits from biotrophy to necrotrophy between 12 and 24 hours after infection |
| (Ithal et al. 2007)                        | Identified 429 differentially expressed genes between uninfected and soybean cyst nematode-infected root tissues                               |
| (Zou et al. 2005)                          | Reported significant changes in transcript abundance for 3,897 genes in response to *Pseudomonas syringae* infection                           |
| (Dhaubhadel et al. 2007)                   | Reported a transcriptome profiling of soybean embryos during development and indicated differences in isoflavonoid content                   |
| (Li et al. 2008)                           | Reported 40 genes that showed specific responses in a soybean aphid-resistant cultivar                                                       |
| (Irsigler et al. 2007)                     | Reported a list of candidate regulatory components that may integrate the osmotic- and ER-stress signaling pathways in plants                  |
| (Brechennacher et al. 2008)                | Explored the complexity of regulatory elements that are likely essential for the development of the symbiosis in root nodules                  |
| (O’Rourke et al. 2007)                     | Reported 43 differentially expressed genes between iron efficient and inefficient genotypes                                                |
Table 2. Gene coexpression network studies in soybean. For the objectives, 1: Inference of gene function; 2: Evolutionary analysis and/or network comparison; 3: Guide-directed discovery of genes involved in a biological process of interest; 4: Module-trait association to identify gene clusters involved in a biological process of interest.

| Reference                  | Algorithm | Objective | # of samples | # of modules | # of genes | Experiment |
|----------------------------|-----------|-----------|--------------|--------------|------------|------------|
| (Almeida-Silva et al. 2020)| WGCNA     | 1,2,4     | 1284         | 29           | 30,563     | RNA-seq    |
| (Bian et al. 2020)         | MR        | 3         | 12           | -            | NA         | RNA-seq    |
| (Yang et al. 2019)         | WGCNA     | 4         | 15           | 12           | 28,072     | Microarray |
| (Wu et al. 2019)           | WGCNA     | 2,4       | 1072         | 156          | 29,047     | y          |
| (Qi et al. 2018)           | WGCNA     | 4         | 45           | 46           | 47,265     | RNA-seq    |
| (Liu et al. 2018)          | WGCNA     | 3         | 18 / 18      | 21 / 24      | 15,494     | RNA-seq    |
| (Du et al. 2017)           | WGCNA     | 4         | 21           | 12           | 7,359      | RNA-seq    |
| (Lima et al. 2017)         | PCC + MR + ClusterONE | 3       | 10           | NA           | 15,497 / y RNA-seq |
| (Wisecaver et al. 2017)    | WGCNA     | 2,4       | 410          | 330†         | 47,965     | RNA-seq    |
| (Yu et al. 2017)           | MCL + Kappa | 2,4     | 375          | 1065         | 38,289     | RNA-seq    |
| (Dias et al. 2016)         | MR        | 3         | 255          | -            | 300        | Microarray |
| (Lu et al. 2016)           | WGCNA     | 2,4       | 40           | 6            | 2680       | RNA-seq    |
| (Mutwil et al. 2011)       | HRR + HCCA | 1,2     | 171          | 422          | NA         | y          |

*NA: Not specified by the authors; †, average number of modules from 10 different networks.

ClusterONE: graph-based clustering method; MCL: Markov clustering; Kappa: statistic to measure agreement between two or more classes of qualitative data; HCCA: Heuristic Cluster Chiseling Algorithm.
Figure 1

A

B

Number of SRA samples by tissue

Tissue

Accepted Manuscript

Downloaded from https://academic.oup.com/insilicoplants/advance-article/doi/10.1093/insilicoplants/diab005/6119548 by guest on 30 January 2021