Exploring the complexity of soybean (*Glycine max*) transcriptional regulation using global gene co-expression networks

Fabricio Almeida-Silva[1], Kanhu C. Moharana[1], Fabricio B. Machado[1], Thiago M. Venancio[1*]

[1]Laboratório 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

Soybean (*Glycine max* (L.) Merr.) is one of the most important crops worldwide, constituting a major source of protein and edible oil. Gene co-expression networks (GCN) have been extensively used to study transcriptional regulation and evolution of genes and genomes. Here, we report a soybean GCN using 1,284 publicly available RNA-Seq samples from 15 distinct tissues. We found modules that are differentially regulated in specific tissues, comprising processes such as photosynthesis, gluconeogenesis, lignin metabolism, and response to biotic stress. We identified transcription factors among intramodular hubs, which probably integrate different pathways and shape the transcriptional landscape in different conditions. The top hubs for each module tend to encode proteins with critical roles, such as succinate dehydrogenase and RNA polymerase subunits. Importantly, gene essentiality was strongly correlated with degree centrality and essential hubs enriched in genes involved in nucleic acids metabolism and regulation of cell replication. By using a using a guilt-by-association approach, we predicted functions for 93 of 106 hubs without functional description in soybean. Most of the duplicated genes had different transcriptional profiles, supporting their functional divergence, although paralogs originating from whole-genome duplications (WGD) are more often preserved in the same module than those from other mechanisms. Together, our results highlight the importance of GCN analysis in unraveling key functional aspects of the soybean genome, in particular those associated with hub genes and WGD events.

Keywords: transcription factors, gene duplication, gene function, genome evolution, systems biology.

INTRODUCTION

Soybean (*Glycine max* (L.) Merr.) is one of the most important legume crops worldwide, being important for human nutrition, animal feed, and biotechnological applications. It has a paleopolyploid genome that resulted from two whole-genome duplication (WGD) events that happened 58 and 13 million years ago [1]. Because of these WGD events, ~75% of the soybean genes are present in multiple copies [2]. Over the last decade, several transcriptomic studies have investigated soybean transcriptional programs in different tissues and conditions [3-6]. However, most works explore only a few conditions and developmental stages, typically using pathway/function enrichment methods. Such approaches can overlook a myriad of interesting transcriptional patterns, which could otherwise be unraveled by integrative studies using hundreds of samples submitted to systems-level analyses.

A powerful approach to further explore transcriptomic data is the use of co-expression networks. Co-expression networks are undirected graphs comprising nodes representing genes that are connected by edges whenever significantly co-expressed. Co-expression networks have been used to understand transcriptional regulation in many plant species, such as maize (*Zea mays* L.), rice (*Oryza sativa* L.), tomato (*Solanum lycopersicum* L.), thale cress (*Arabidopsis thaliana* (L.) Heynh), and soybean [7-10]. However, most co-expression network analyses focus on case-control experiments, limiting their potential to uncover broader biological patterns [11-13]. A recent global co-expression network analysis of 1,072 soybean microarray samples revealed a gene module that is likely involved in the evolution of nodulation in plants [14]. In spite of these interesting results, hybridization-based transcriptional quantification often suffer from poor reproducibility [15]. Over the past few years, data from hundreds of soybean RNA-seq samples have accumulated and were recently integrated by our group in a large and publicly available soybean transcriptome atlas [16]. Nevertheless, a network analysis using this collection remains to be performed.

Co-expression analyses have recently been used to study the evolution of gene function in plants [17, 18]. After a duplication event, selective constraints are relaxed and gene copies typically undergo two possible fates: i. gene fractionation, when one copy accumulate loss-of-function mutations and eventually erode from the genome and; ii. gene retention via subfunctionalization or neofunctionalization [19]. Although pseudogenization is the most common fate of duplicated genes [20], genes encoding regulatory proteins (e.g. transcription factors – TFs – and signal transduction proteins) or subunits of protein complexes (e. g. proteasome and ribosome) are retained more often than expected [21, 22], mostly because of the maintenance of their stoichiometry [21]. However, in case of local duplications (e.g. tandem duplications), TFs and subunits of protein complexes tend to be fractionated because of dosage sensitivity [23].

Gene retention is thought to be more likely if the newly duplicated genes remain in the genome for an extended window of time, allowing mutation and

selection pressures to rewire gene functions [24]. The vast majority of retained genes undergo subfunctionalization and, more rarely, neofunctionalization [25]. At the transcriptional level, subfunctionalization of a gene pair typically happens through mutations at cis-regulatory regions, resulting in two possible scenarios: i. both genes reduce their transcription, making them both necessary to perform the original function or; ii. genes from the pair become expressed in different tissues or conditions (transcriptional divergence) [26]. Thus, a network-level gene expression analysis could shed light on the possible fates of duplicated genes.

Here, we report a GCN built with data from 1,284 RNA-seq samples to investigate the global transcriptional landscape of soybean in different tissues and conditions. We identified biological processes and metabolic pathways that are enhanced or repressed in particular tissues and pointed TF families that probably control their expression, including candidates for downstream experiments (e.g. ectopic expression). We predicted the function of 93 unannotated hubs using a guilt-by-association approach. Further, we used this large GCN to gain insight into the fate of duplicated genes. We demonstrate that most soybean duplicated genes diverge in their expression profiles. However, duplicate pairs originating from WGDs are more often preserved in the same module than those arising by other mechanisms, suggesting an increased conservation of gene function for WGD-derived copies. We also demonstrate that duplicate pairs in the same module are likely under stronger purifying selection, as revealed by their lower Ka/Ks ratios. Collectively, our findings reveal potentially novel regulators and enlighten our understanding of the fate of duplicated genes during the evolution of the soybean genome.

MATERIAL AND METHODS

Data collection and processing

The normalized expression data (TPM, Transcripts Per Million) for 1,298 samples were downloaded from the recently published Soybean Expression Atlas [16]. Samples were filtered by using the standardized connectivity (Z.k) method in the R package *WGCNA* [27]. For that, we constructed a sample network based on squared Euclidean distances and assessed the overall scaled connectivity (Z.k) of the network. This method is efficient in filtering outlying samples and has been largely used in gene co-expression analyses [28-30]. Samples with Z.k < -2.5 were considered outliers and were filtered out. The resulting 1,284 samples represent 15 different tissues (Supplementary Figure S1). Further, to reduce noise, we excluded genes with TPM < 5 in more than 20% of the samples. Finally, the 30,563 x 1,284 matrix was $\log_2$ transformed by using log (TPM+1).

Network reconstruction and module detection

The signed network was inferred by using the R package *WGCNA* [27]. We assessed the scale-free topology fit to choose the most suitable β power. The lowest β power for which the network reached a scale-free topology (R^2 = 0.8) was 17

(Supplementary Figure S2). Pairwise correlations between genes were calculated by using the Pearson Correlation Coefficient (PCC). The PCC matrix was then transformed to an adjacency matrix, which was used to construct a topological overlap matrix (TOM). The TOM considers both PCC and shared neighbors to determine similarity, which amplifies disparity between gene pairs, making it effective for module detection. Genes were clustered into modules with hierarchical clustering based on dissimilarity (1 – TOM). Finally, we calculated pairwise similarity between modules and merged the ones that had PCC ≥ 0.8. Minimum module size was set to 30. The R package *igraph* (http://igraph.org) [31] was used for extraction and visualization of subgraphs.

Hub gene inference and functional enrichment

Hub genes are those with the highest number of connections (i. e. degree) in a network. Intramodular connectivity (kWithin) was used as degree to detect intramodular hubs. The choice of thresholds for hub identification is somewhat arbitrary. While some groups consider the top 10% genes with the highest degree as hubs [32, 33], others use module membership (MM) to define such genes [34]. MM is a measure of how close a given gene is to its module, and it can be calculated by correlating a gene to the module eigengene. Here, we combined both parameters and considered hubs the top 10% genes with the highest intramodular connectivity and with MM ≥ 0.8. Enrichment analysis for GO terms, pathways and conserved protein domains was performed in PhytoMine (https://phytozome.jgi.doe.gov/phytomine/begin.do), and REVIGO was used to remove redundancy [35]. Soybean TFs were downloaded from PlantTFDB [36], and TF enrichment was determined by performing an one-sided Fisher's exact test with the R package *bc3net* [37]

Analysis of upstream cis-regulatory elements

We searched for regulatory motifs in the regions between -1000bp and +200bp relative to the transcription start site, as done in *A. thaliana* [38]. These regions were extracted with *bedtools* [39]. A *de novo* motif discovery was performed by using MEME [40]. As background, we used an order 0 Hidden Markov Model (HMM) with all expressed genes, using the script *fasta-get-markov* available in MEME suite.

Simulated networks, paralogous genes and selection analysis

We simulated 10,000 degree-preserving networks in R through successive permutations of node labels to verify the probability of finding our results by chance. Paralogous gene pairs derived from dispersed, tandem, proximal, transposed, and whole-genome duplications were download from the Plant Duplicate Gene Database [41]. We used Ks thresholds to subdivide WGD-derived pairs into the WGDs that happened 13 mya (Ks ≤ 0.4) and 58 mya (Ks > 0.4). For simplicity, a smaller number of pairs potentially arising from older WGD events were combined in this latter

category. Ka, Ks and Ka/Ks ratios were computed using the Perl script *calculate_Ka_Ks_pipe.pl*, available in [41]. Paralog pairs in the same module were interpreted as having similar transcriptional profiles. Pairs were interpreted as having different transcriptional profiles if: i. genes were distributed in different modules; ii. only one gene from the pair was present in the network or; iii. both genes were in the *grey* module, which stores genes that were not assigned to modules.

RESULTS AND DISCUSSION

Degree centrality analysis reveals that hubs are enriched in essential genes

After hierarchically clustering the genes based on TOM dissimilarity and merging the modules with 80% or more similarity, we generated a GCN containing 29 modules (Figure 1; Supplementary Table S1). The largest and smallest modules were *chocolate3* (n = 6,250) and *steelblue3* (n = 30), respectively. The *grey* module, which stores unassigned genes, had 4,766 genes. We selected the top hub from each module and analyzed their functions in PhytoMine. Top hub genes tend to encode proteins of paramount importance, such as succinate dehydrogenase (Glyma.19G08600), an RNA polymerase subunit (Glyma.06G217400), DNA topoisomerase (Glyma.14G137100), and cellulose synthase (Glyma.06G324300) (Supplementary Table S2).

Jeong *et al* found that hubs tend to be essential in yeast (*S. cerevisiae*) protein-protein interaction networks [42]. We hypothesized that degree centrality in a GCN is also associated with gene essentiality in soybean. Essential genes in plants are often referred to as embryonic lethal genes (i. e. genes that compromise embryo viability when not expressed). *A. thaliana* essential genes have been diligently catalogued over the years by Meinke's group, who recently published a curated list of 510 *EMBRYO-DEFECTIVE (EMB)* genes [43], out of which 481 have soybean orthologs (1,010 soybean genes). Importantly, the soybean orthologs of the essential genes (Supplementary Table S3) are significantly overrepresented among GCN hubs (Fisher's exact test; p = 5.8e-05). Further, essential genes that are hubs are enriched in functions related to nucleic acids metabolism, translation and regulation of cell division, such as replication factors, ribosomal proteins, and elongation factors.

Next, we investigated if any of our modules was individually enriched for essential genes and essential hub genes. The modules *darkmagenta, yellow2, lightskyblue1, lightblue1, chocolate3, darkgoldenrod3*, and *darkviolet* were significantly enriched in essential genes (Fisher's exact test; p = 7.8e-05, 2.3e-19, 4.5e-02, 5.5e-09, 7.8e-06, 5.8e-09, and 6.0e-03, respectively). A functional enrichment analysis revealed that, respectively, these modules are related to photosynthesis and gluconeogenesis; translation and ribosome biogenesis; RNA synthesis and processing; RNA metabolism; DNA replication; protein ubiquitination and regulation of transcription and; *darkviolet* had the enriched protein domain acetyl-CoA carboxylase alpha subunit. The latter three modules were also enriched with essential hubs (Fisher's exact test; p = 9.6e-08, 1.3e-04, 4.8e-02, respectively).

192 These findings show that these gene clusters are essential to embryo viability and
193 could be experimentally validated in the future. Overall, our findings corroborate the
194 well-established role of hub genes as essential players of the cell physiology [44, 45].
195
196 **Module hubs uncover biological processes associated with specific tissues**
197 We selected the hub genes from each module and performed an enrichment analysis
198 for GO terms, pathways and protein domains (see methods for details). Information
199 on the number of hubs per module are in Supplementary Table S4. Only 14 of 29
200 modules (48.3%) had enriched GO terms, pathways or protein domains, and we
201 identified the function of nine of the 14 modules based on enriched biological
202 processes (GO-BP) and pathways (Supplementary Table S5). Further, we identified
203 module-tissue relationships by correlating each module eigengene to each of the 15
204 tissues available in our dataset (Figure 2; Supplementary Table S6), some of which
205 are discussed below.
206 Genes in the module *aliceblue* are related to lignin and phenylpropanoid
207 metabolism, such as laccases (Glyma.01G112600, Glyma.03G077900,
208 Glyma.08G359100), peroxidases (Glyma.01G163100, Glyma.11G049600,
209 Glyma.11G078400, Glyma.12G195500), xylem cysteine proteinase
210 (Glyma.04G014700, Glyma.06G014800), and cytochrome b561
211 (Glyma.02G306800). Module-tissue correlation showed that they are up-regulated in
212 leaf, root, and hypocotyl (r = 0.33, 0.31, 0.22; p = 1e-34, 3e-29, 1e-15, respectively)
213 and down-regulated in cotyledons, embryo and seeds (r = -0.21, -0.29, -0.33; p =
214 1e-14, 8e-26, 2e-34, respectively). Lignins are polymers made up of monolignols,
215 which are polymerized into lignin by laccases and peroxidases [46]. Because of their
216 strong covalent bonds to cellulose in secondary cell walls, lignins provide mechanical
217 support to plant organs such as stems and leaves, which explains the up-regulation
218 of *aliceblue* genes in this tissue [47]. In addition, lignins are difficult for herbivores
219 to digest, constituting a defense mechanism. Lignin is also well known for its role in
220 the stability of vascular tissues such as xylem [48]. The xylem cysteine proteinase
221 encoded by Glyma.04G014700, also in *aliceblue*, is likely involved in the autolysis of
222 tracheary elements during xylogenesis in roots [49]. Lignin is primarily synthesized
223 in hypocotyls during seedling establishment, as shown in growing hemp hypocotyls
224 [50]. Since lignin biosynthesis begins in the hypocotyl at the seedling stage, such
225 genes are also expected to be down-regulated in embryo, seeds and cotyledons,
226 which are mostly represented by developing seeds in our dataset. Thus, our results
227 recapitulate some of the physiological and anatomical roles of lignin.
228 Hub genes in the module *antiquewhite4* encode proteins related to the
229 tricarboxylic acid (TCA) cycle and aerobic respiration, such as isocitrate
230 dehydrogenase (Glyma.13G144900, Glyma.10G058100, Glyma.16G118900) and
231 ureide permeases (Glyma.01G058500 and Glyma.02G116300). These genes are up-
232 regulated in leaf (r = 0.53; p = 5e-93) and down-regulated in seed, embryo and
233 cotyledon (r = -0.43, -0.43, -0.2; p = 7e-60, 6e-60, 2e-13, respectively). As most

seed-related tissues in our dataset represent developing seeds, these genes are expected to be down-regulated, since seed reserves are accumulated during development. Storage proteins and other nutrients are then mobilized during germination to promote early seedling development [51]. The up-regulation of TCA cycle-related genes in leaf is likely because the TCA cycle is one of the major sources of redox equivalents for oxidative phosphorylation, which provides ATP for sucrose synthesis [52].

In *blue4,* we found genes related to transcriptional regulation and signaling, such as ABC transporters (Glyma.13G361900, Glyma.07G015800), ferrochelatase (Glyma.04G050400), phosphatases (Glyma.06G290200, Glyma.08G071900), kinases (Glyma.04G088800, Glyma.06G090800, Glyma.13G073900, Glyma.14G001300, Glyma.02G311200), TFs (Glyma.03G127600, Glyma.02G232600, Glyma.07G023300, Glyma.14G200200, Glyma.09G005700, Glyma.12G212300, Glyma.13G289400), and adenylate kinase (Glyma.10G048200). *Blue4* genes were slightly up-regulated in leaf and root (r = 0.21 and 0.24; p = 6e-14, 5e-18, respectively) and down-regulated in embryo (r = -0.34; p = 6e-36). We analyzed the annotation of *blue4* orthologs in *A. thaliana* and found that this module is likely related to defense against pathogens, especially fungi. Among these orthologs, we found the ABC transporters PEN3 and PIS1 (AT1G59870 and AT3G53480), which participate in salicylic acid-dependent resistance against direct pathogen penetration and in secretion of phenolic compounds involved in Fe uptake [53]. Absorbed Fe ions can then be used by the ferrochelatase FC1 (AT5G26030) to supply heme for defensive hemoproteins [54]. Similarly, the TFs WRKY33 and TGA6 from this module are involved in the activation of salicylic acid-responsive genes and pathways mediating response to fungal pathogens, respectively [54, 55]. Innate immune responses rely on the detection of microbe-associated molecular patterns (MAMPs) by pattern recognition receptors (PRRs). Most *A. thaliana* PRRs are leucine-rich repeat (LRR)-receptor kinases [56, 57]. *Blue4* also harbor orthologs of *A. thaliana* innate immunity genes, such as an LRR-receptor kinase (AT5G01950) and an interleukin-1 receptor-associated kinase (AT1G14370). We also found an ortholog of the receptor kinase CERK1 (AT3G21630), which is involved in response to chitooligosaccharides, a very common plant MAMP [56, 58]. Hence, the up-regulation of these genes in leaf and root is coherent, whereas their down-regulation in embryo is likely because of its physical protection by other organs and tissues [59].

Hubs from *darkmagenta* encode proteins related to photosynthesis and gluconeogenesis, such as phosphoribulokinase (Glyma.01G010200), glycine transaminase (Glyma.01G026700), photosystem II enhancer protein 1 (Glyma.02G061100), NADPH-dependent thioredoxin reductase 3 (Glyma.02G185300), fructose-bisphosphate aldolase (Glyma.04G008300), glyceraldehyde-3-phosphate dehydrogenase (Glyma.04G015900), and fructose-1,6-biphosphatase (Glyma.07G142700). These genes are remarkably up-regulated in leaf (r = 0.8; p = 7e-292) and down-regulated in embryo, root, seed, and seed coat (r =

-0.28, -0.29, -0.34, and -0.28; p = 4e-25, 2e-26, 4e-36, 2e-25, respectively). This is expected because in leaves, photosynthesis is responsible for CO_2 fixation in triose phosphate, which feeds sucrose synthesis [60]. Hence, our findings support a strong transcriptional co-regulation of sucrose synthesis and photosynthesis genes. In root and seed-related parts, however, the down-regulation of these genes is due to their very low photosynthetic potential as compared with other plant parts, making them predominantly heterotrophic [61-64].

Finally, hubs from *orange2* are enriched in genes involved in cytochrome C biogenesis and assembly (e.g. Glyma.06G254100, Glyma.12G147300). This module is up-regulated in leaf (r = 0.59; p = 2e-119) and down-regulated in seed-related parts (embryo and seed coat, r = -0.35, -0.32; p = 1e-37, 6e-32, respectively). The *Arabidopsis* ortholog (AT5G54290) of these genes encodes CcdA, a thylakoid membrane cytochrome C biogenesis protein that transfers reducing equivalents from stromal thioredoxins to the thylakoid lumen [65]. Many proteins located in the thylakoid lumen depend on redox regulation, and it has been shown that the *Arabidopsis ccda* mutant is incapable of assembling the cytochrome b_6f in thylakoids [66].

Functional prediction of unannotated hubs

Guilt-by-association (GBA) approaches have been widely used for function inference [67-70]. In the context of GCN, it assumes that genes with similar expression patterns are co-regulated and more likely to share the same pathway or biological process. We explored the GCN and found 106 unannotated (i.e. genes without brief description in PhytoMine) hubs, of which 93 were distributed in the nine modules for which we could assign some biological function, as described above. By using a GBA approach, we predicted functions for all these 93 hubs (Supplementary Table S7), which might unveil interesting players in soybean transcriptional regulation.

We also found that 28 out of 29 modules had annotated top hubs. The top hub in the module *antiquewhite4* (Glyma.19G074200) encoded a protein of unknown function. We performed a functional enrichment analysis of its direct neighbors ($r \geq 0.7$) and found that they are involved in response to auxin (p = 1.19e-3), which is likely the function of this unannotated gene. These genes encode auxin-responsive GH3 family proteins (Glyma.03G256200 and Glyma.19G258800), and SAUR-like auxin-responsive proteins (Glyma.04G231100, Glyma.06G134000, Glyma.08G004100, Glyma.19G254000).

The soybean genome remains far from being completely annotated. While 80.71% of soybean genes have some, although often minimal, functional annotation in PhytoMine's brief description, the remaining 19.29% encode proteins of unknown function. Network-based approaches like the one reported here might help us understand the functions of many interesting genes, in particular those potentially associated with agronomic traits. As hub genes are typically major players of the cell physiology, our results can be relevant for the soybean genomics community.

318
319 **Enrichment of transcription factor families uncovers major regulators of**
320 **important biological processes**
321 We applied a Fisher's exact test and found that 10 of the 29 modules are enriched in
322 TFs (Supplementary Table S8), suggesting that the functions involved in these
323 modules require a more complex transcriptional regulatory architecture. We also
324 found 6 modules enriched in specific TF families (FDR ≤ 0.05; Table 1). We
325 hypothesize that the overrepresented TF families in each module have key regulators
326 of the biological processes enriched in the module (Table 1). We also performed a *de*
327 *novo* motif discovery in promoter regions of genes in each module with enriched TF
328 families (Supplementary Figure S3), although none of the identified motifs matched
329 those from the same family in JASPAR and other Arabidopsis databases available in
330 Tomtom (MEME suite) [71]. Because of their regulatory roles, TFs are often reported
331 as important targets for developing transgenic lines with increased efficiency in
332 several biological processes and metabolic pathways [72-74]. Hence, the genes
333 reported in Table 1 could be interesting targets for plant transformation studies.
334
335 **Network topology and the possible fate of soybean duplicated genes**
336 We sought to explore the constructed GCN to explore the fates of duplicate genes.
337 All soybean duplicated gene pairs were downloaded from Plant Duplicate Gene
338 Database [41]. Duplicated gene pairs were classified in tandem, proximal,
339 transposed, dispersed and whole-genome duplication (TD, PD, TRD, DD and WGD,
340 respectively). Based on the Ks distribution, WGD-derived duplicates were subdivided
341 into WGD 58 mya and WGD 13 mya, representing WGD events that happened 58 and
342 13 million years ago, respectively (Supplementary Figure S4). The co-occurrence of
343 duplicates in modules can help understand their fates after duplication, as each
344 module has a unique expression profile (Supplementary Figure S5). In 27.2% of the
345 gene pairs (22,003/80,946), both genes were excluded during data filtering (see
346 methods for details) and are thus not present in the network, while in 27.6%
347 (22,355/80,946) only one gene of the pair was present in the network. Finally, both
348 genes were present in the network for 45.2% of the pairs (36,589/80,946).
349 Expectedly, these three categories have distinct gene expression levels
350 (Supplementary Figure S6).
351 We analyzed the co-occurrence of duplicated genes and the prevalence of their
352 modes of duplication in modules. Duplicate pairs in which both genes were absent in
353 the network were not considered. The co-occurrence of duplicate pairs in the same
354 module was interpreted as an indicative of functional conservation, as they reflect
355 similar expression profiles after duplication. On the other hand, when duplicate pairs
356 were in different modules or when only one gene from the pair was present, we
357 interpreted it as an indicative of functional divergence. In order to assess the
358 significance of these results, we created 10,000 degree-preserving simulated
359 networks (Figure 3; see methods for details). We observed that most pairs displayed

divergent expression profiles or signs of fractionation (Figure 3). Strikingly, WGD pairs, particularly those from the most recent WGD, showed greater levels of retention when compared to other modes of gene duplication, both in the same and in different modules. These observations are possibly correlated with previous reports on the increased retention of WGD duplicates involved in intricate systems (e.g. regulatory genes) and protein complexes [23].

As previously shown, subfunctionalization is by far the most common fate of retained duplicated genes [75], although it is not totally clear how this process happens at the transcriptional level. Our findings suggest that the most common fate of duplicated genes is functional divergence, as 59,4% to 91.7% of the duplicates do not belong to the same module. Nevertheless, the frequency of genes in the same module is higher than the expected by chance, supporting that the transcriptional similarity of part of the duplicates is under selective pressure. This finding provides new insights into how retained copies are regulated throughout the soybean genome.

Finally, we hypothesized that paralog pairs with similar expression profiles (i.e same module) are under stronger purifying selection as compared to paralog pairs with different expression profiles (i.e. members of the pair in different modules, both in *grey* module or only one gene from the pair present in the network). We observed that, except for proximal and tandem duplication-derived pairs, paralog pairs with similar expression profiles have significantly lower Ka/Ks ratios as compared to those with different expression profiles (Wilcoxon test; $p < 0.05$; Figure 3C). This result suggests that duplicates that are retained in the same module are under stronger negative selection, both at transcriptional and sequence levels. These pairs could be involved in systems that rely on tightly regulated stoichiometric balance of its components.

CONCLUSION

Here, we employed a thorough GCN analysis to explore the systems-wide transcriptional landscape of soybean tissues. We explored the functions of several modules and hubs using enrichment and GBA methods. Several of the identified modules are strongly associated with specific tissues, giving important clues about their physiological roles. We also found modules enriched in TFs, which probably correspond to systems dependent on more intricate transcriptional regulation architectures. Finally, we used our GCN to explore evolution of duplicate genes in soybean and found that most of the retained duplicated genes tend to diverge in expression, although genes from WGD are more often found in the same modules than those duplicated by other mechanisms. The GCN reported here provides insights into the roles of several soybean genes and co-expression modules, which might be important not only in the elucidation of gene functions, but also in exploring genes involved in relevant agronomic traits.

AUTHOR CONTRIBUTIONS

Conceived the study: FA-S, KCM and TMV. Data analysis: FA-S, KCM and FBM. Funding, project coordination and infrastructure: TMV. Manuscript writing: FA-S and TMV.

ACKNOWLEDGEMENTS

This work was supported by funding from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES, Finance code 001), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), and Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ). The authors declare no conflict of interest.

REFERENCES

1. Gill, N., et al., *Molecular and chromosomal evidence for allopolyploidy in soybean.* Plant Physiol, 2009. **151**(3): p. 1167-74.
2. Schmutz, J., et al., *Genome sequence of the palaeopolyploid soybean.* Nature, 2010. **463**(7278): p. 178-83.
3. Bellieny-Rabelo, D., et al., *Transcriptome analysis uncovers key regulatory and metabolic aspects of soybean embryonic axes during germination.* Sci Rep, 2016. **6**: p. 36009.
4. Collakova, E., et al., *Metabolic and Transcriptional Reprogramming in Developing Soybean (Glycine max) Embryos.* Metabolites, 2013. **3**(2): p. 347-72.
5. Gazara, R.K., et al., *Transcriptional landscape of soybean (Glycine max) embryonic axes during germination in the presence of paclobutrazol, a gibberellin biosynthesis inhibitor.* Sci Rep, 2019. **9**(1): p. 9601.
6. Severin, A.J., et al., *RNA-Seq Atlas of Glycine max: a guide to the soybean transcriptome.* BMC Plant Biol, 2010. **10**: p. 160.
7. Huang, J., et al., *Construction and Optimization of a Large Gene Coexpression Network in Maize Using RNA-Seq Data.* Plant Physiol, 2017. **175**(1): p. 568-583.
8. Jiang, Z., et al., *Differential Coexpression Analysis Reveals Extensive Rewiring of Arabidopsis Gene Coexpression in Response to Pseudomonas syringae Infection.* Sci Rep, 2016. **6**: p. 35064.
9. Ozaki, S., et al., *Coexpression analysis of tomato genes and experimental verification of coordinated expression of genes found in a functionally enriched coexpression module.* DNA Res, 2010. **17**(2): p. 105-16.
10. Pereira Lima, J.J., et al., *Molecular characterization of the acquisition of longevity during seed maturation in soybean.* PLoS One, 2017. **12**(7): p. e0180282.

11. Gamboa-Tuz, S.D., et al., *Transcriptomics and co-expression networks reveal tissue-specific responses and regulatory hubs under mild and severe drought in papaya (Carica papaya L.).* Sci Rep, 2018. **8**(1): p. 14539.
12. Tan, M., et al., *Co-expression network analysis of the transcriptomes of rice roots exposed to various cadmium stresses reveals universal cadmium-responsive genes.* BMC Plant Biol, 2017. **17**(1): p. 194.
13. You, Q., et al., *Co-expression network analyses identify functional modules associated with development and stress response in Gossypium arboreum.* Sci Rep, 2016. **6**: p. 38436.
14. Wu, Z., et al., *A global coexpression network of soybean genes gives insights into the evolution of nodulation in nonlegumes and legumes.* New Phytol, 2019. **223**(4): p. 2104-2119.
15. Jaksik, R., et al., *Microarray experiments and factors which affect their reliability.* Biol Direct, 2015. **10**: p. 46.
16. Brum Machado, F., et al., *Systematic analysis of 1,298 RNA-Seq samples and construction of a comprehensive soybean (Glycine max) expression atlas.* Plant J, 2020.
17. Hansen, B.O., et al., *Elucidating gene function and function evolution through comparison of co-expression networks of plants.* Front Plant Sci, 2014. **5**: p. 394.
18. Romero-Campero, F.J., et al., *A contribution to the study of plant development evolution based on gene co-expression networks.* Front Plant Sci, 2013. **4**: p. 291.
19. Roulin, A., et al., *The fate of duplicated genes in a polyploid plant genome.* Plant J, 2013. **73**(1): p. 143-53.
20. Rensing, S.A., *Gene duplication as a driver of plant morphogenetic evolution.* Curr Opin Plant Biol, 2014. **17**: p. 43-8.
21. Birchler, J.A., et al., *Dosage balance in gene regulation: biological implications.* Trends Genet, 2005. **21**(4): p. 219-26.
22. Papp, B., C. Pal, and L.D. Hurst, *Dosage sensitivity and the evolution of gene families in yeast.* Nature, 2003. **424**(6945): p. 194-7.
23. Moharana, K.C. and T.M. Venancio, *Polyploidization events shaped the transcription factor repertoires in legumes (Fabaceae).* Plant J, 2020.
24. Teufel, A.I., et al., *The Many Nuanced Evolutionary Consequences of Duplicated Genes.* Mol Biol Evol, 2019. **36**(2): p. 304-314.
25. Panchy, N., M. Lehti-Shiu, and S.H. Shiu, *Evolution of Gene Duplication in Plants.* Plant Physiol, 2016. **171**(4): p. 2294-316.
26. Braasch, I., et al., *Reply to: 'Subfunctionalization versus neofunctionalization after whole-genome duplication'.* Nat Genet, 2018. **50**(7): p. 910-911.
27. Langfelder, P. and S. Horvath, *WGCNA: an R package for weighted correlation network analysis.* BMC Bioinformatics, 2008. **9**: p. 559.
28. Koh, F.M., et al., *Emergence of hematopoietic stem and progenitor cells involves a Chd1-dependent increase in total nascent transcription.* Proc Natl Acad Sci U S A, 2015. **112**(14): p. E1734-43.
29. Levine, M.E., P. Langfelder, and S. Horvath, *A Weighted SNP Correlation Network Method for Estimating Polygenic Risk Scores.* Methods Mol Biol, 2017. **1613**: p. 277-290.

30. Patel, H., R.J.B. Dobson, and S.J. Newhouse, *A Meta-Analysis of Alzheimer's Disease Brain Transcriptomic Data.* J Alzheimers Dis, 2019. **68**(4): p. 1635-1656.
31. Csardi, G. and T. Nepusz, *The igraph software package for complex network research.* InterJournal, complex systems, 2006. **1695**(5): p. 1-9.
32. Bi, D., et al., *Gene expression patterns combined with network analysis identify hub genes associated with bladder cancer.* Comput Biol Chem, 2015. **56**: p. 71-83.
33. Borin, G.P., et al., *Gene Co-expression Network Reveals Potential New Genes Related to Sugarcane Bagasse Degradation in Trichoderma reesei RUT-30.* Front Bioeng Biotechnol, 2018. **6**: p. 151.
34. Mack, K.L., et al., *Gene Expression Networks Across Multiple Tissues Are Associated with Rates of Molecular Evolution in Wild House Mice.* Genes (Basel), 2019. **10**(3).
35. Supek, F., et al., *REVIGO summarizes and visualizes long lists of gene ontology terms.* PLoS One, 2011. **6**(7): p. e21800.
36. Jin, J., et al., *PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants.* Nucleic Acids Res, 2017. **45**(D1): p. D1040-D1045.
37. de Matos Simoes, R. and F. Emmert-Streib, *Bagging statistical network inference from large-scale gene expression data.* PLoS One, 2012. **7**(3): p. e33624.
38. Yu, C.P., J.J. Lin, and W.H. Li, *Positional distribution of transcription factor binding sites in Arabidopsis thaliana.* Sci Rep, 2016. **6**: p. 25164.
39. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features.* Bioinformatics, 2010. **26**(6): p. 841-2.
40. Bailey, T.L., et al., *MEME SUITE: tools for motif discovery and searching.* Nucleic Acids Res, 2009. **37**(Web Server issue): p. W202-8.
41. Qiao, X., et al., *Gene duplication and evolution in recurring polyploidization-diploidization cycles in plants.* Genome Biology, 2019. **20**.
42. Jeong, H., et al., *Lethality and centrality in protein networks.* Nature, 2001. **411**(6833): p. 41-2.
43. Meinke, D.W., *Genome-wide identification of EMBRYO-DEFECTIVE (EMB) genes required for growth and development in Arabidopsis.* New Phytol, 2019.
44. Han, J.D., et al., *Evidence for dynamically organized modularity in the yeast protein-protein interaction network.* Nature, 2004. **430**(6995): p. 88-93.
45. Luscombe, N.M., et al., *Genomic analysis of regulatory network dynamics reveals large topological changes.* Nature, 2004. **431**(7006): p. 308-12.
46. Heldt, H.-W., *Plant Biochemistry*. 2011.
47. Liu, Q., L. Luo, and L. Zheng, *Lignins: Biosynthesis and Biological Functions in Plants.* Int J Mol Sci, 2018. **19**(2).
48. Bhuiyan, N.H., et al., *Gene expression profiling and silencing reveal that monolignol biosynthesis plays a critical role in penetration defence in wheat against powdery mildew invasion.* J Exp Bot, 2009. **60**(2): p. 509-21.
49. Zhao, C., et al., *XND1, a member of the NAC domain family in Arabidopsis thaliana, negatively regulates lignocellulose synthesis and programmed cell death in xylem.* Plant J, 2008. **53**(3): p. 425-36.

50. Behr, M., et al., *Insights into the molecular regulation of monolignol-derived product biosynthesis in the growing hemp hypocotyl.* BMC Plant Biol, 2018. **18**(1): p. 1.
51. Rajjou, L., et al., *Seed germination and vigor.* Annu Rev Plant Biol, 2012. **63**: p. 507-33.
52. Kromer, S., *Respiration during Photosynthesis.* Annual Review of Plant Physiology and Plant Molecular Biology, 1995. **46**: p. 45-70.
53. Marmagne, A., et al., *A high content in lipid-modified peripheral proteins and integral receptor kinases features in the Arabidopsis plasma membrane proteome.* Molecular & Cellular Proteomics, 2007. **6**(11): p. 1980-1996.
54. Nagai, S., et al., *Induction of isoforms of tetrapyrrole biosynthetic enzymes, AtHEMA2 and AtFC1, under stress conditions and their physiological functions in arabidopsis.* Plant Physiology, 2007. **144**(2): p. 1039-1051.
55. Libault, M., et al., *Identification of 118 Arabidopsis transcription factor and 30 ubiquitin-ligase genes responding to chitin, a plant-defense elicitor.* Molecular Plant-Microbe Interactions, 2007. **20**(8): p. 900-911.
56. Belkhadir, Y., et al., *Brassinosteroids modulate the efficiency of plant immune responses to microbe-associated molecular patterns.* Proceedings of the National Academy of Sciences of the United States of America, 2012. **109**(1): p. 297-302.
57. Ronald, P.C. and B. Beutler, *Plant and Animal Sensors of Conserved Microbial Signatures.* Science, 2010. **330**(6007): p. 1061-1064.
58. Osman, G.H., et al., *Development of insect resistant maize plants expressing a chitinase gene from the cotton leaf worm, Spodoptera littoralis.* Scientific Reports, 2015. **5**.
59. Raviv, B., et al., *The Dead Can Nurture: Novel Insights into the Function of Dead Organs Enclosing Embryos.* International Journal of Molecular Sciences, 2018. **19**(8).
60. Ward, J.M., et al., *Sucrose transport in higher plants.* International Review of Cytology - a Survey of Cell Biology, Vol 178, 1998. **178**: p. 41-71.
61. Adams, C.A. and R.W. Rinne, *Carbon-Dioxide Fixation in Developing Soybean Seeds.* Plant Physiology, 1979. **63**(5): p. 17-17.
62. Asokanthan, P.S., et al., *The photosynthetic potential of canola embryos.* Physiologia Plantarum, 1997. **101**(2): p. 353-360.
63. Harvey, D.M., C.L. Hedley, and R. Keely, *Photosynthetic and Respiratory Studies during Pod and Seed Development in Pisum-Sativum-L.* Annals of Botany, 1976. **40**(169): p. 993-1001.
64. Rolletschek, H., H. Weber, and L. Borisjuk, *Energy status and its control on embryogenesis of legumes. Embryo photosynthesis contributes to oxygen supply and is coupled to biosynthetic fluxes.* Plant Physiology, 2003. **132**(3): p. 1196-1206.
65. Motohashi, K. and T. Hisabori, *CcdA Is a Thylakoid Membrane Protein Required for the Transfer of Reducing Equivalents from Stroma to Thylakoid Lumen in the Higher Plant Chloroplast.* Antioxidants & Redox Signaling, 2010. **13**(8): p. 1169-1176.

566 66. Page, M.L.D., et al., *A homolog of prokaryotic thiol disulfide transporter CcdA is*
567 *required for the assembly of the cytochrome b(6)f complex in Arabidopsis*
568 *chloroplasts.* Journal of Biological Chemistry, 2004. **279**(31): p. 32474-32482.
569 67. Usadel, B., et al., *Co-expression tools for plant biology: opportunities for*
570 *hypothesis generation and caveats.* Plant Cell Environ, 2009. **32**(12): p. 1633-51.
571 68. Shahan, R., et al., *Consensus Coexpression Network Analysis Identifies Key*
572 *Regulators of Flower and Fruit Development in Wild Strawberry.* Plant Physiol,
573 2018. **178**(1): p. 202-216.
574 69. Petereit, J., et al., *petal: Co-expression network modelling in R.* BMC Syst Biol,
575 2016. **10 Suppl 2**: p. 51.
576 70. De Smet, R., et al., *Coordinated Functional Divergence of Genes after Genome*
577 *Duplication in Arabidopsis thaliana.* Plant Cell, 2017. **29**(11): p. 2786-2800.
578 71. Sandelin, A., et al., *JASPAR: an open-access database for eukaryotic transcription*
579 *factor binding profiles.* Nucleic Acids Res, 2004. **32**(Database issue): p. D91-4.
580 72. Rabara, R.C., P. Tripathi, and P.J. Rushton, *The Potential of Transcription Factor-*
581 *Based Genetic Engineering in Improving Crop Tolerance to Drought.* Omics-a
582 Journal of Integrative Biology, 2014. **18**(10): p. 601-614.
583 73. Sharma, V., et al., *An apple transcription factor, MdDREB76, confers salt and*
584 *drought tolerance in transgenic tobacco by activating the expression of stress-*
585 *responsive genes.* Plant Cell Reports, 2019. **38**(2): p. 221-241.
586 74. Wang, N.N., et al., *The cotton WRKY transcription factor (GhWRKY33) reduces*
587 *transgenic Arabidopsis resistance to drought stress.* Scientific Reports, 2019. **9**.
588 75. Freeling, M., *The evolutionary position of subfunctionalization, downgraded.*
589 Genome Dyn, 2008. **4**: p. 25-40.

Figures

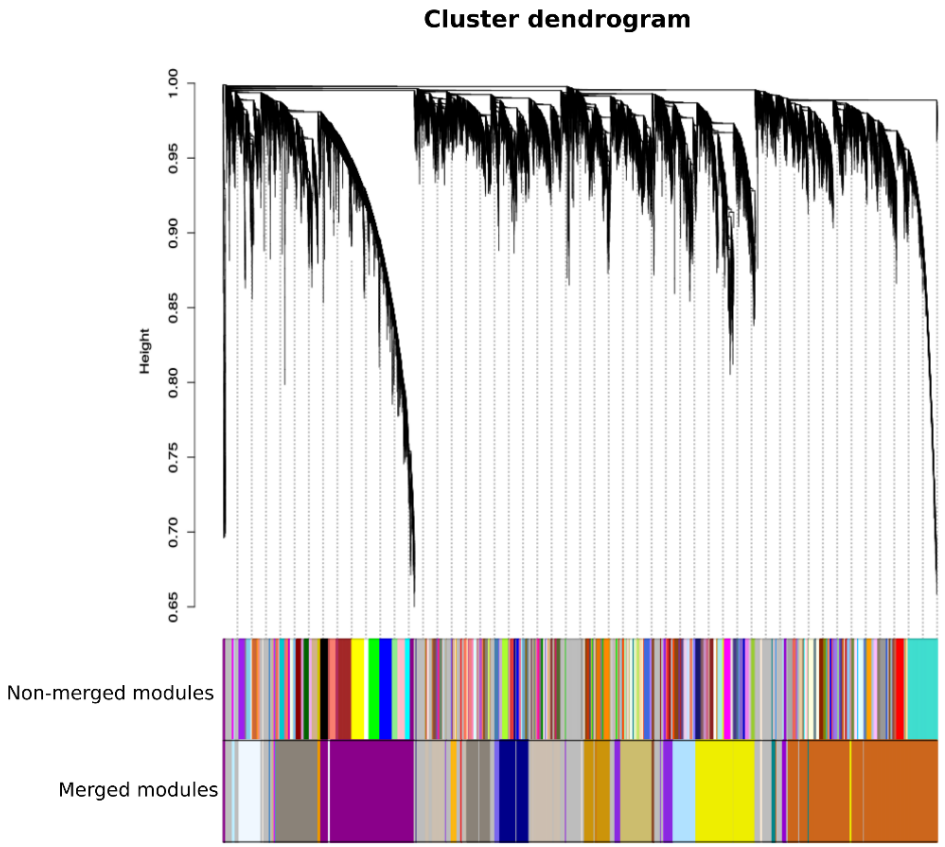

Figure 1. Dendrogram and module colors before and after merging similar modules. In the WGCNA standard pipeline, each module is given a color. Modules were merged based on similarity of their eigengenes. The module eigengene is the first principal component of a principal component analysis, which summarizes the expression of the module. Modules with eigengenes more than 80% similar (r = 0.8) were merged.

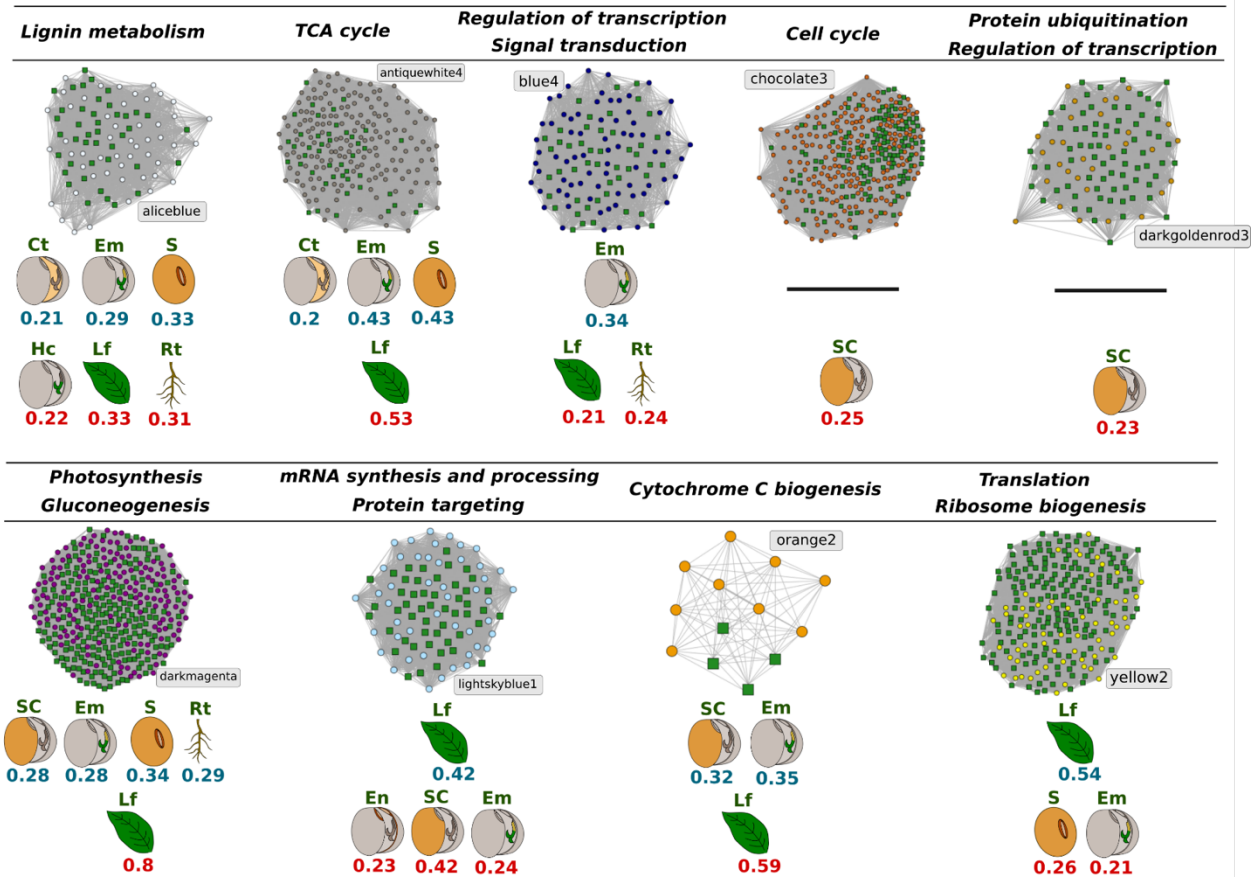

Figure 2. Hubs from modules with enriched biological processes (GO) and pathways. For each subgraph, green squares represent hubs that are TFs. Below each subgraph, we illustrate tissues where the genes tend to be down-regulated (blue) and up-regulated (red). Up- and down-regulation in each tissue were obtained by calculating Pearson correlation coefficients between module eigengenes and a binary matrix m_{ij} containing 1 if the sample *i* corresponds to the tissue *j,* and 0 otherwise. P-values were derived by using the function corPvalueStudent() in WGCNA. Given the heterogeneity of the data, correlations were considered significant if $r \geq 0.2$ and $p < 0.05$. Ct, cotyledon; Em, embryo; S, seed; Hc, hypocotyl; Lf, leaf; Rt, root; SC, seed coat; En, endosperm.

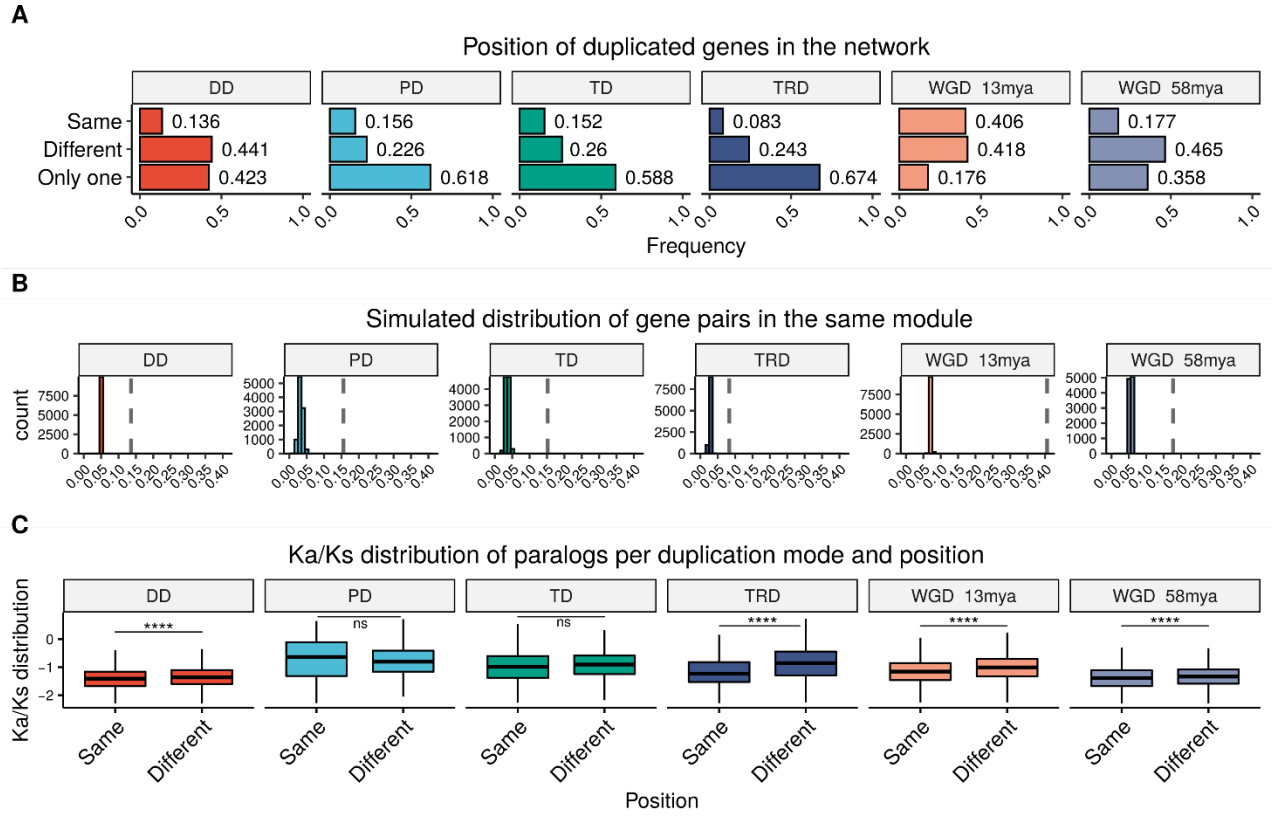

Figure 3. Distribution and Ka/Ks analysis of duplicate pairs in the network. A. We observed that the minority of the duplicate pairs derived from all modes of duplication were in the same module in the GCN, supporting functional divergence as the most common fate of duplicated genes. B. We created 10,000 degree-preserving simulated networks by permuting the node labels of the GCN. Next, we compared the distribution of frequencies in which duplicate pairs were in the same module (histograms) with that observed in the real GCN (dashed lines). When both genes from a pair fell into the *grey* module, we treated them as being in different modules, as *grey* stores unassigned genes. C. Distribution of Ka/Ks ratios per mode of duplication and presence in the same module. Ka/Ks values were log transformed using log(Ka/Ks + 0.1) for better visualization. Except for proximal and tandem duplication-derived gene pairs, Ka/Ks ratios are significantly lower for paralog pairs within the same module, supporting that they are under stronger purifying selection at both transcriptional and sequence levels. DD, dispersed duplications; PD, proximal duplications; TD, tandem duplications; TRD, transposed duplications; WGD, whole-genome duplications; ns, no significant difference; ****, $p<0.0001$.

Tables

Table 1. Overrepresented TF families in soybean co-expression modules. One-sided Fisher's exact test with Benjamini-Hochberg correction was used. TF families were considered significantly overrepresented if FDR < 0.05.

Module name	TF family	FDR	Module function	Gene ID
aliceblue	WOX	8.83e-4	Lignin metabolism	Glyma.04G040900, Glyma.14G084600
antiquewhite4	TALE	1.27e-3	TCA cycle and aerobic respiration	Glyma.09G095700, Glyma.10G081700, Glyma.12G188800, Glyma.13G312900, Glyma.17G104800
blue4	WRKY	2.98e-6	Response to biotic stress	Glyma.02G232600, Glyma.07G023300, Glyma.08G218600, Glyma.09G005700, Glyma.12G212300, Glyma.13G289400, Glyma.14G200200, Glyma.15G003300, Glyma.15G110300
darkmagenta	CO-like	7.28e-3	Photosynthesis and gluconeogenesis	Glyma.05G233700, Glyma.07G091400
lightskyblue1	MYB-related	5.41e-4	mRNA synthesis and processing	Glyma.09G257700, Glyma.13G128500, Glyma.15G236400, Glyma.18G234900
yellow2	C2H2	1.04e-2	Translation and ribosome biogenesis	Glyma.11G189500, Glyma.12G084700, Glyma.12G181400