Weighted Gene Correlation Network Analysis (WGCNA) of *Arabidopsis* somatic embryogenesis (SE) and identification of key gene modules to uncover SE-associated hub genes

Kithmee K. de Silva
University of Colombo
Jim M. Dunwell
University of Reading
Anushka M. Wickramasuriya (anushka@pts.cmb.ac.lk)
University of Colombo

Research Article

Keywords: Somatic embryogenesis, Arabidopsis, WGCNA, Gene modules, Hub genes, Co-expression network

Posted Date: January 27th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1302556/v1

Abstract

Background

Somatic embryogenesis (SE), which occurs naturally in many plant species, serves as a model to elucidate cellular and molecular mechanisms of embryo patterning in plants. Decoding the regulatory landscape of SE is essential for its further application. Hence, the present study aimed to employ Weighted Gene Correlation Network Analysis (WGCNA) to construct a gene co-expression network (GCN) for *Arabidopsis* SE, and then identify highly correlated gene modules to uncover the hub genes associated with SE that may serve as potential molecular targets.

Results

A total of 17,059 genes with a variance greater than 0.25 quantile were filtered from microarray datasets derived from *Arabidopsis* SE (NCBI Gene Expression Omnibus accession number: GSE48915). This included 1,711 transcription factors and 445 *EMBRYO DEFECTIVE* genes. GCN analysis identified a total of 26 gene modules with the module size ranging from 35 to 3,418 genes using a dynamic cut tree algorithm. The module-trait analysis revealed that four, four, seven and four modules were associated with Stage I (zygotic embryos), Stage II (proliferating tissues at 7 days of induction), Stage III (proliferating tissues at 14 days of induction) and Stage IV (mature somatic embryos), respectively. Further, we identified a total of 260 hub genes based on the degree of intramodular connectivity. Validation of the hub genes using publicly available expression datasets demonstrated that at least 78 hub genes are potentially associated with embryogenesis; of these many genes remain functionally uncharacterized thus far. *In silico* promoter analysis of these genes revealed the presence of EBOXBNNAPA and SEF4MOTIFGM7S motifs; this suggests these genes may play important roles in plant embryo development.

Conclusion

The present study successfully applied WGCNA to construct a GCN for SE in *Arabidopsis* and identified hub genes involved in the development of somatic embryos. These hub genes could be used as molecular targets to further elucidate the molecular mechanisms underlying SE in plants.

Background

The ability to produce embryos from undifferentiated somatic cells *in vitro* is a unique developmental pathway found within the plant kingdom. Since the first report of somatic embryo induction from callus cells of carrot [1, 2], this developmental pathway based on cellular totipotency has been studied extensively due to its biological and scientific significance; it has been recognized as a model system for studying early plant embryogenesis. Until now, most studies have focused on the mechanism of somatic

embryo development at the morphological level [2–4] or the development of optimized protocols for the generation of somatic embryos from a range of explants [5–8].

Somatic embryogenesis (SE) involves a complex signaling network [9]; transcriptional regulation of a set of genes in response to stress caused by plant growth regulators, nutrients, certain stress conditions, and other signaling elements triggers cellular reprogramming and transformation of somatic cells into embryos [10, 11]. In 2007, Zeng and his colleagues [12] developed the first draft gene regulatory network for early SE employing a set of transcriptionally regulated SE-related genes in cotton. Although a set of genes have been identified as markers for the initiation phase of SE [13, 14], for example, *SOMATIC EMBRYOGENESIS RECEPTOR-LIKE KINASE1* (*SERK1*) [15, 16], *LEAFY COTYLEDON* (*LEC*) [17–21], *BABY BOOM* (*BBM*) [22] and *WUSCHEL* (*WUS*) [18, 23], the current scientific knowledge on the underlying regulatory landscape of SE is limited. The use of transcriptomics has uncovered a large number of differentially expressed genes (DEGs) during SE in many crops, including *Arabidopsis* [24], rice [25], bread wheat [26], cotton [27], maize [28], and coconut [29]. However, the functions of many of these genes in SE are still not understood.

Gene co-expression networks (GCNs) are increasingly used to understand the interactions among a set of transcriptionally regulated genes. There are many types of co-expression networks: signed/unsigned co-expression networks and weighted/un-weighted co-expression networks [30]. In the present study, we have focused on weighted network construction as it is likely to produce more robust findings than un-weighted networks [31]. Weighted Correlation Network Analysis (WGCNA) is one of the most popular clustering packages for GCN analysis [31, 32] and the first tool to be employed to construct GCNs from RNA-sequencing (RNA-seq) data. This co-expression tool is easy to use and can be used to find clusters (modules) of highly correlated genes and to identify biologically relevant associations between phenotypes/sample traits and modules from expression data [30]. Recently, WGCNA has been effectively used to identify stage-specific gene expression clusters associated with key stages of *Arabidopsis* zygotic embryo development [33]. In addition, this approach has been successfully used to discover the regulatory landscape of SE in rice [25] and several other biological pathways in plants [34–36]. Here we have analyzed, a transcriptome dataset covering four somatic embryo developmental stages in *Arabidopsis* using WGCNA to understand better the system-level functionality of the transcriptionally regulated genes in dicot SE.

Results

Hierarchical clustering of somatic embryo transcriptomes

In the present study, transcriptome datasets generated through microarray experiments were retrieved from the NCBI Gene Expression Omnibus (GEO) database (GSE48915) covering four somatic embryo developmental stages (with two replicates for each stage), referred to herein as Stage I (zygotic embryos), II (proliferating tissues at 7 days of induction), III (proliferating tissues at 14 days of induction) and IV (mature somatic embryos). The hierarchical clustering of samples (Fig. 1a) confirmed that the sample

replicates of each stage have a higher degree of correlation with each other than with other developmental stages; sample outliers were not detected in the dataset. The clustering heatmap clearly distinguished four discrete clusters of related expression patterns corresponding to the stages of somatic embryo development (Fig. 1b). Further, Stage I showed a poor correlation with the other three stages. This suggests that Stage I may have a distinct expression profile as compared to other somatic embryo developmental stages.

Filtering of genes for the GCN construction and downstream analysis

As recommended by Langfelder and Horvath (2008) [32] genes were filtered by the variance for the GCN construction; filtering genes for variance greater than 0.25 quantile identified a total of 17,059 genes (Fig. 2a; see Additional file 1: Table S1). This included 445 *EMBRYO-DEFECTIVE* (*EMB*) genes [37], 10 SE marker genes [38] and 1,711 *Arabidopsis* transcription factors (TFs; 65.3%).

In addition, DEGs were identified by a pairwise ratio of expression between consecutive stages of development. A total of 2,244 genes were identified by threshold filtering based on log Fold Change (FC) ($log_2\ FC \geq 2.0 \mid log_2\ FC \leq 2.0$) and $p\text{-}value < 0.05$. 64 *EMB* genes [37], four SE marker genes [38] and 458 TFs were present within the DEGs identified (see Additional file 1: Table S2). Only *Photosystem II Subunit Q* gene (AT4G05180) was differentially expressed throughout SE (Fig. 2b).

Construction of GCN

The expression profiles of the filtered 17,059 genes were used to construct a scale-free gene expression network with a soft threshold of 15 (Fig. 3a). The dynamic hierarchical clustering approach integrated with the WGCNA pipeline distinguishes groups of genes with co-expression patterns and clusters them into network modules. In total, 26 distinct co-expression gene modules were detected with the module size ranging from 35 to 3,418 genes (Fig. 3b and c); each module was assigned with a unique colour. The module comprising most genes was the turquoise (3,418 genes) followed by the blue (2,973 genes) and brown (2,437 genes) (Fig. 3b). The expression profiles of co-expressed genes clustered in each module were summarized as ‘module eigengenes’ (MEs). Among the filtered genes, 13 genes that failed to fit within a distinct group were assigned to the grey module and removed from the downstream analysis. Module preservation analysis indicated high module preservation, confirming that the modules generated here can also be found in diverse independent datasets (Fig. 3b). Each module was exported and visualized using Cytoscape (Fig. 3d).

Identification of stage-related modules

The relationships between the gene modules and different somatic embryo developmental stages were determined by assessing the Pearson correlation coefficient (r) between the MEs and developmental stages. Module-trait correlation analyses revealed that multiple modules are related to SE (Fig. 4a). A

total of 18 modules were significantly associated with the somatic embryo developmental stages ($|r| > 0.8$ and $p\text{-value} \leq 0.01$; Fig. 4), and these modules were “stage-specific”, i.e., the module was significantly associated with only one particular developmental stage of SE; tan, turquoise, dark-orange and green to Stage I, grey60, magenta, brown, and light-yellow to Stage II, green-yellow, dark-gray, dark-green, orange, blue, light-green and light-cyan to Stage III, and pink, dark-turquoise, salmon and yellow to Stage IV. Gene significance, the correlation between modular gene expression and each stage, is shown in Fig. 4b.

Functional enrichment analysis of ‘stage-specific’ gene modules

Gene ontology (GO) enrichment analysis performed on ‘stage-specific’ modules showed that the genes in green and turquoise modules which exhibited a significant association with Stage-I were mainly enriched in the biological processes being involved in post-embryonic development, hormone-mediated signaling pathway, biosynthesis pathways (sterol and fatty acids), DNA methylation and transcription regulation. Genes in brown, light-yellow and magenta modules, which showed significant association with Stage II, were mainly enriched in the biological processes involved in root and shoot development, ATP synthesis, response to the metal ions and DNA replication whereas genes in blue and light-cyan modules, which showed significant association with Stage III, were enriched for the biological processes involved in transition post-embryonic and seed development, hormone- and sugar-mediated signaling pathways, cell differentiation, protein modification and RNA processing. Moreover, the yellow module, which showed a significant relationship to Stage IV, was mainly enriched in biological processes involved in ion transport, post-embryonic development, signal transduction, lipid localization, response to oxidative and water stress as well as response to phytohormones (abscisic acid, gibberellin, cytokinin and jasmonic acid).

Analysis of hub genes

Hub genes are nodes in a network often hypothesized to be functionally significant due to their high degree of intra-modular connectivity. A total of 260 genes (top 10 genes of each module with high connectivity) were identified as potential hub genes; the hub gene with the highest degree of connectivity in each module is given in Table 1 (the complete list of hub genes is given in Additional file 2: Table S3). GO enrichment analysis of the hub genes revealed that they are mainly enriched for biological processes such as metabolic processes (mRNA and cellular amino acid), oxidation-reduction, protein folding and post-embryonic development.

Among the hub genes, only 234 genes were functionally annotated; of these, 13 were TFs: *AUXIN RESPONSE FACTOR 9 (ARF9)*, *FLOWERING BHLH 4 (FBH4)*, *BASIC HELIX-LOOP-HELIX 39 (BHLH39)*, *BASIC LEUCINE-ZIPPER 44 (bZIP44)*, *bZIP19*, *ZIM-LIKE 2 (ZML2)*, *AT5G60820*, *AT4G01270*, *KANADI 3 (KAN3)*, *HOMEODOMAIN GLABROUS 4 (HDG4)*, *CELL DIVISION CYCLE 5 (CDC5)*, *NAC DOMAIN CONTAINING PROTEIN 80 (NAC080)* and *SALT TOLERANCE*

(STO)). In addition, five genes encoding transposable elements (i.e., *AT2G11560*, *AT3G33066*, *AT5G32430*, *AT3G42820* and *AT4G28900*) were identified.

-

In silico analysis of the promoter sequences (1000 bp upstream from the transcription start site) of the hub genes using the Multiple Em for Motif Elicitation (MEME) tool identified four significant motifs ranging in length from 15 to 29 bp (Table 2). Motifs 1, 2 and 3 were detected across 229, 245 and 121 hub genes, respectively. Further analysis of the predicted motifs using the GOMo (Gene Ontology for Motifs) tool provided in the MEME suite indicated that motifs 1 and 3 may be involved in the DNA endoreduplication, polarity specification of axial/abaxial axis and hormone-mediated signaling pathways; motif 1 and 3 seem to function in association to cytokinin and gibberellic acid, respectively.

Validation of hub genes

Comparison of hub genes and DEGs showed that 31 hub genes are differentially expressed in SE (the expression values of differentially expressed hub genes are given in Additional file 3: Table S4). Further, expression analysis of these genes using *Arabidopsis* eFP browser demonstrated that two hub genes, *AT1G19540* and *AT5G44380* exhibit a seed-specific pattern of expression (Fig. 5).

Moreover, analysis of the expression profiles of hub genes in the *Arabidopsis* somatic embryo transcriptome dataset (E-MTAB-2465) published by Wickramasuriya and Dunwell (2015) revealed that 62 hub genes are differentially expressed in somatic embryonic tissues compared to leaf tissues ($|log_2\ FC| \geq 2.0$ and $p\text{-value} < 0.05$; Fig. 6). Of these, 15 genes were identified as DEGs in the present analysis. For instance, *CYSTEINE-RICH TRANSMEMBRANE MODULE 7* (*ATHCYSTM7*/AT2G33520), *HEPTAHELICAL TRANSMEMBRANE PROTEIN2* (AT4G30850), *INDOLE-3-ACETIC ACID INDUCIBLE 30* (*IAA30*/AT3G62100), *RPS9C*, *VASCULATURE COMPLEXITY AND CONNECTIVITY* (AT2G32280), *AT2G21820*, *AT2G38900* and *AT5G43770* showed marked expression in somatic embryonic tissues as compared to leaf tissues. Expression analysis using the *Arabidopsis* eFP browser further showed that *AT2G29300*, *AT2G21820*, *AT2G38900*, *AT5G43770*, *ATHCYSTM7* and *AT1G19540* exhibit a seed-specific pattern of gene expression.

As expected few hub genes highly expressed in leaf tissues were repressed in somatic embryos indicating the importance of gene regulation in SE (Fig. 6); for instance, *CELLULOSE SYNTHASE-LIKE B4* (AT2G32540), *CHOLINE/ETHANOLAMINE KINASE 3* (AT4G09760), *GLUTAMATE DECARBOXYLASE 2* (AT1G65960), *ISOPROPYLMALATE ISOMERASE 2* (AT2G43100), *PEROXIREDOXIN Q* (*PRXQ*/AT3G26060), *PHOTOSYNTHETIC NDH SUBCOMPLEX L 4* (*PnsL4/*AT4G39710), *PLASTID RIBOSOMAL PROTEIN S20* (AT3G15190), *STO* (AT1G06040), *SINAPOYLGLUCOSE 1* (*SNG1*/AT2G22990), *THYLAKOID RHODANESE-LIKE* (*TROL*/AT4G01050), *TONOPLAST INTRINSIC PROTEIN 2* (*TIP2*/AT3G26520), *AT3G50685*,

AT4G33666, *AT5G16010* and *AT5G54540* genes showed marked repression in somatic embryos compared to leaf tissues.

In summary, the present study identified a total of 78 hub genes as potential regulators of SE (Fig. 7), including genes showing marked over-expression as well as repression in SE. Of these, 41 genes have not been functionally annotated thus far. Analysis of the promoter sequences of these uncharacterized hub genes using the Plant *cis*-acting regulatory DNA elements (PLACE) database identified a total of 215 different plant *cis*-acting regulatory elements (CREs); ARR1AT, CAATBOX1, CACTFTPPCA1, DOFCOREZM, GATABOX, GT1CONSENSUS, POLLEN1LELAT52 and WRKY71OS observed in all 41 functionally uncharacterized potential hub genes. Moreover, several CREs related to embryogenesis were identified (Fig. 8). The functions of the predicted CREs are included in Table 3.

Distribution of embryogenesis-related genes across network modules

Further exploration of genes mapped to each network module found that 10 key regulators of SE including *LEC1*, *FUSCA3* (*FUS3*) and *ABSCISIC ACID INSENSITIVE 3* (*ABI3*) are present among the highly connected genes in the network (Table 4); SE-related marker genes, *LEC2*, *SERK1*, *WUS*, *BBM* and *WUSCHEL RELATED HOMEOBOX 2* (*WOX2*)) showed low variance in the present dataset and thus were not included in the GCN analysis. We also observed that the majority of previously published *EMB* genes [37] are localized to the blue and turquoise modules, which showed significant association with Stage I and Stage III, respectively (Fig. 9; see Additional file 4: Table S5).

In addition, we observed that 1,711 *Arabidopsis* TFs are distributed across all the gene modules except in light-green and royal-blue modules, with the highest number of TFs present in the turquoise module (the complete list of TFs included in the GCN is given in Additional file 5: Table S6). Notably, AP2/EREBP (APETALA2/Ethylene-responsive element binding proteins), bHLH (basic helix–loop–helix), bZIP, C2H2 (Cys2-His2), HB (Homeobox), NAC (NAM, ATAF, and CUC), MYB (MYB-domain), C3H and WRKY TF families were highly represented (Fig. 10a). Of these, members of AP2/EREBP, bHLH, C2H2, HB, NAC, MYB and WRKY TF families were involved in early SE (Fig. 10b). Interestingly, TFs that are targets of several microRNAs (miRNAs) were also recovered from the GCN (Additional file 5: Table S7).

Notably, several genes encoding epigenetic regulators were localized in network modules (Fig. 11). This included 14 genes involved in DNA modification, 51 genes involved in histone modification, 34 genes involved in chromatin remodeling, 15 genes encoding polycomb-group proteins and 55 genes associated with RNA silencing (see Additional file 6: Table S8). Each of these genes directly interacted with numerous modular genes forming a complex network.

Discussion

Plant embryogenesis is a meticulous developmental process that requires the regulation of multiple genes. A GCN will serve as a map of statistically significant gene interactions that helps in narrowing down the transcriptome to the potential gene interactions involved in biological processes. Recently,

Clercq and colleagues report an integrated gene regulatory network for *Arabidopsis* covering TFs and target genes [39]. In the present study, WGCNA was employed to explore potential clusters of highly co-regulated genes and hub genes associated with SE. Although WGCNA has been previously applied to construct a GCN for *Arabidopsis* zygotic embryogenesis (ZE) [33], to the best of our knowledge, this is the first report on the use of WGCNA to construct a GCN for *Arabidopsis* SE and to explore SE related network modules and hub genes. The findings of this study provide new insights into the molecular mechanism of SE in plants.

The GCN constructed for SE comprised of 26 network modules: black (674 genes), blue (2,973 genes), brown (2,437 genes), cyan (125 genes), dark-green (52 genes), dark-grey (39 genes), dark-orange (35 genes), dark-red (54 genes), dark-turquoise (52 genes), green (2,132 genes), green-yellow (189 genes), grey60 (79 genes), light-cyan (86 genes), light-green (59 genes), light-yellow (58 genes), magenta (338 genes), midnight-blue (117 genes), orange (35 genes), pink (357 genes), purple (271 genes), red (853 genes), royal-blue (56 genes), salmon (162 genes), tan (172 genes), turquoise (3,418 genes) and yellow (2,223 genes) modules. Among them 18 modules showed strong associations with different stages of SE; module-trait relationship analysis revealed that four, four, seven and four modules were significantly correlated with Stage I, Stage II, Stage III and Stage IV of SE, respectively. This suggests SE involves complex genetic networks.

Functional enrichment analysis using GO is one of the most widely used bioinformatics methods to classify genes into functionally related groups [40–42]. GO analysis of the co-expressed gene clusters (or network modules) showed that the initial stages of SE were mainly enriched with biological processes such as hormone-mediated signaling, biosynthesis pathways, ATP synthesis, DNA methylation and replication. Notably, genes involved in lipid transport, post-embryonic development, signal transduction and seed dormancy were enriched in later stages of SE; this indicates the developmental shift in the maturation phase with the accumulation of embryo-specific food reserves, a process that aids in withstanding dormancy and postembryonic development [2, 10, 43]. Furthermore, genes related to stress responses (e.g. oxidative and water stress), phytohormones (e.g. cytokinin, abscisic acid, gibberellin and jasmonic acid) and metabolic processes were enriched in all stages of somatic embryo development studied, from the initiation to maturation stage. These findings further confirmed the importance of cell-cell interactions [44], signaling [9, 13, 45] and transcriptional activation of stress responses [46, 47] during plant SE.

High-degree nodes or the genes with high network connectivity in GCN modules ('hub genes') may have important biological functions [36, 48–50]; often they may serve as biological markers. Several studies have successfully employed WGCNA to mine hub genes controlling biological processes [34, 51–54]. The present study reports 260 potential hub genes related to SE based on the degree of connectivity; These genes may play pivotal roles in the regulation of SE. Importantly, 13 TFs encoded by hub genes were identified in the co-expression network. They were *ARF9*, *NAC080, ZML2*, *bHLH39*, *KAN3*, *bZIP19*, *bZIP44*, *HDG4*, *FBH4*, *STO*, *CDC5*, *AT5G60820* and *AT4G01270*; functional roles of many of these genes in the regulation of SE are not reported. Previous studies have reported that ARF9 represses the expression of

its target genes such as TOPLESS (TPL) and TPL-related proteins [55, 56]. Wójcikowska and Gaj (2017) observed stable expression of *ARF9* during SE [57]. In addition, *KAN3*, a member of GARP TF family, has also exhibited an embryonic expression pattern.

In addition, *ROOT UV-B SENSITIVE 6* (*RUS6*; AT5G49820), which encodes a DUF647 (DOMAIN OF UNKNOWN FUNCTION 647) containing protein, an ankyrin repeat-containing gene designated as *AT5G65860* and a gene that encodes hydroxyproline-*O*- glycosyltransferases (Hyp-*O*-GALT), *GALT4* (*AT1G27120*)) were also identified as hub genes in the co-expression network. The members of the *RUS* gene family play diverse roles in plant development [58]. Interestingly, knockout mutants of *RUS6* have shown a strong embryo-lethal phenotype. In *Arabidopsis*, ankyrin repeat-containing proteins have been classified into 16 groups [59], and of these, proteins with only ankyrin repeats have been associated with disease resistance, antioxidation, embryogenesis and development [60–62]. For instance, T-DNA mutants of the *EMB* 506 gene, which encodes a protein containing five ankyrin repeats, have shown defective embryo development at the globular-to-heart stage transition [62]. Moreover, Hyp-*O*-GALT enzymes are responsible for hydroxyproline glycosylation of arabinogalactan proteins, which are known to function in various aspects of plant growth and development including SE [63–65]. Although the hub genes identified in the present study are implicated to function in many plant developmental processes, the functions of many of the hub genes in SE remain to be elucidated. Hence, these genes could be potential targets for functional studies in the future.

Promoter analysis of the functionally uncharacterized hub genes using the PLACE database revealed the overrepresentation of two motifs in many of the promoter regions. These were EBOXBNNAPA (consensus sequence: CANNTG) and SEF4MOTIFGM7S (consensus sequence: [A/G]TTTTT[A/G]). Of these, EBOXBNNAPA ('E-box' motif) is a CRE found in the regulatory region of the napin gene, *napA* in *Brassica napus* [66]; this gene encodes a storage protein. Moreover, CANNTG provides the binding site for bHLH TFs [67]. bHLH is one of the most frequently represented gene families in DEGs in ZE [68] and SE, and is known to have diverse functions in plants [24] including cell proliferation [67]. The recognition sequence of SEF4MOTIFGM7S motif is known to interact with SEF3, a protein expressed in immature soybean seeds that acts as a transcriptional activator of the β-conglycinin α subunit gene [69]. Hence, the uncharacterized hub genes that showed considerable expression in embryonic tissues are more likely to play a significant role in plant embryo development.

Differential gene expression analysis of hub genes revealed that 78 genes could be considered as potential regulators of SE; of these 15 genes were differentially expressed in transcriptome datasets derived from two independent studies related to SE [24, 70]. One of the genes identified was *IAA30*, which is a member of one of the families of auxin signaling proteins (Aux/IAA; [71]). *iaa30* mutants have displayed significantly impaired SE efficiency, producing fewer somatic embryos per explant [68] and suggesting its role in the initiation phase of SE. Moreover, IAA30 is a target of two important SE marker genes, LEC2 and AGL15 [72, 73]. In addition, two hub genes, *AT1G19540* and *AT5G44380*, showed marked expression in seed development, suggesting their roles in embryogenesis.

To enhance our understanding of the regulatory mechanism of SE, the distribution of embryogenesis-related genes across the gene modules was examined. Horstman *et al*. (2017) report LEC1–LEC2–FUS3–BBM–ABI3 network to induce SE in *Arabidopsis* [74]. Moreover, Zheng *et al*. (2009) suggest a MADS-domain TF encoding gene, *AGL15* may associate with *LEC2*, *FUS3* and *ABI3* during SE [75]. However, a recent study has found that *AGL15* is not essential to promote SE [76]. In the present analysis, 10 key regulators of SE including *LEC1*, *ABI3*, *FUS3, AGL15* and three members of the AINTEGUMENTA-LIKE/PLETHORA (AIL/PLT) subfamily (*ANT*, *AIL5* and *AIL7*) were identified in the co-expression network. Consistent with previous literature, members of the AP2/EREBP, bHLH, bZIP, MYB, HB, WRKY, NAC, C3H and C2H2 TF families were over-represented in the GCN [68, 77]. In addition, members of the TF families (i.e., SPB (SQUAMOSA promoter binding protein-like), GRAS (GRAS-domain), trihelix, G2-like and CAMTA (CALMODULIN BINDING TRANSCRIPTION ACTIVATOR 3)) that are not or to a lesser extent reported to be involved in SE were identified. The members of GRAS, trihelix and CAMTA families are known to be involved in the regulation of stress responses [39, 78, 79].

Further, it is reported that miRNAs (e.g. miR156, miR159, miR162, miR164, miR166, miR167, miR169, miR168, miR171, miR319, miR393 and miR396) play an important role in SE [80–84]. Consistent with previous studies, several TFs targeted by miRNAs were recovered from the SE-related GCN. This included seven miR156/157 targeting genes of the SPB TF family, seven miR169 targeting genes of the CCAAT TF family, six miR396 targeting genes of the GRF TF family, five miR166/miR165 targeting genes of the HB TF family, five miR164 targeting genes of the NAC family and five miR159/miR319 targeting genes of the TCP TF family. These miRNA-targeted TF encoding genes may play a significant role in the regulation of SE responses.

Recent studies have uncovered critical roles of epigenetic modifications in the regulation of SE, in particular, DNA methylation/demethylation [85–87] and histone modifications [84, 88, 89]. Recently, an expression study on *Arabidopsis* embryos at single-cell resolution has provided evidence for distinct expression patterns for many epigenetic regulators across embryonic tissues [90]. Our co-expression network also revealed that many genes encoding epigenetic regulators such as *METHYLTRANSFERASE 1* (*MET1*), *CHROMOMETHYLASE 3* (*CMT3*), *DEMETER* (*DME*), *DEMETER-LIKE* (*DML1,-2*), histone acetyltransferases (*HISTONE ACETYLTRANSFERASE OF THE CBP FAMILY* (*HAC1,-4,-5,-12*), histone deacetylases (i.e. *HISTONE DEACETYLASE* (*HDA1,-2,-3,-5,-6,-8,-9,-14,-15,-17*), and histone demethylases (*JUMONJI DOMAIN-CONTAINING PROTEIN 16* (*JMJ14,-16,-22,-27,-29*) were co-expressed with key genes involved in the regulation of SE.

The present study showed that the WGCNA pipeline could be used to identify biologically relevant modules of SE. However, our analysis has some limitations. The main limitations were the small sample size used in the analysis and the lack of an independent dataset to replicate the findings. Langfelder and Horvth (2008) [32] recommend using at least 15 samples to construct robust networks. However, high-quality, clean data could also result in biologically meaningful networks even with < 15 samples. Therefore, further experiments are recommended to validate the hub genes discovered in the present study.

Conclusion

In this study, a GCN was successfully constructed for SE employing WGCNA. Gene modules and hub genes related to *Arabidopsis* somatic embryo development were successfully mined based on their statistical significance. The findings reported here provide a unique resource to advance the regulation of SE at the molecular level.

Methods

Data collection and gene filtering

In the present study, 'WGCNA' package in R software [32] was employed to identify significant gene modules and hub genes in *Arabidopsis* somatic embryo transcriptomes. The transcriptome data covering somatic embryo developmental stages of wild-type *Arabidopsis* were retrieved from the National Centre for Biotechnology Information (NCBI) GEO database (GEO accession: GSE48915) [70]. The dataset consisted of four developmental stages (zygotic embryos, proliferating tissues at 7 days of induction, proliferating tissues at 14 days of induction and mature somatic embryos) with two replicates for each stage (i.e., a total of eight samples). Subsequently, the genes with variance greater than the second quartile of variance were filtered to construct the GCN. In addition, genes with ($log_2 FC \geq 2.0 \mid log_2 FC \leq 2.0$) and $p\text{-}value < 0.05$ were considered as differentially expressed (FC: the ratio of expression between each pair of stages).

GCN construction

A gene co-expression similarity matrix was constructed between the expression profiles of the filtered genes using the Pearson correlation. The similarity matrix was then transformed into an adjacency matrix where each entry encodes the connection strength between each pair of genes ('nodes'). The adjacency matrix defines a measure of node dissimilarity from which the nodes (genes) are clustered into network modules. Consequently, the GCN was developed using the automatic one-step network construction and module detection method with the following parameters:

```
net = blockwiseModules(data, power = 15, corType = "pearson",   networkType
      = "signed", TOMType = "signed", minModuleSize = 30, mergeCutHeight =
      0.25)
```

The soft threshold value (power parameter) was decided by the scale-free topology fit index curve.

GCN visualization

The constructed modular networks were exported to Cytoscape (version 3.7.2) for visualization; gene correlations with *p-value <0.05* were filtered as significant gene correlations and visualized. The modular networks were analyzed by the 'Network Analyzer' tool in Cytoscape for a concise and informative representation of nodes and edges.

Validation of network modules

The robustness of the co-expression modules was assessed through module preservation and quality statistics, which were computed using the modulePreservation function in the WGCNA package [91]. The adjacency matrix of the network was taken as the reference and the dataset was selected as test data with 200 permutations (*nPermutations = 200*). The stability of the modules was tested through the statistics median rank and Zsummary.

Inferring module-stage relationships

Module-stage relationships of the GCN were evaluated through MEs. The correlation relationships between the MEs and different somatic embryo developmental stages were analyzed and visualized through a heatmap. Gene significance was calculated based on the *p-value* of the linear regression between the gene expression profile and the associated developmental stage.

Functional enrichment analysis

Functional enrichment analysis was performed to detect enriched biological processes in gene modules. GO terms enriched in each module were elucidated using the '*Singular Enrichment Analysis*' tool provided by agriGO v2.0 (http://systemsbiology.cau.edu.cn/agriGOv2/classification_analysis.php?category=Plant&&family=Brassicaceae) [92]. 'Arabidopsis genome locus (TAIR10)' was used as the reference and all other parameters were set as default for the analysis. Overrepresented GO terms in each network module were identified using the Hypergeometric test. To further explore the DEGs mapped to each gene module, the distribution of the following genes across modules was studied: SE-related marker genes [38], plant TFs (http://planttfdb.cbi.pku.edu.cn/index.php), *EMB* genes [37] and genes encoding epigenetic regulators [93, 94].

Identification and validation of hub genes

Genes in each module were arranged based on gene connectivity. The top 10 genes of each module were considered as hub genes. The transcriptome dataset published by Wickramasuriya and Dunwell in 2015 was retrieved from the ArrayExpress database (E-MTAB-2403) [24] to study the expression of hub genes during SE.

In silico analysis of hub genes

The promoter sequences of hub genes (1000 bp upstream from the transcription start site) were retrieved from 'The Arabidopsis Information Resource' (TAIR) database and analysed using the MEME tool in the MEME Suite 5.3.3 [95]. The following parameters were used in the analysis: number of motifs: 10, motif site distribution: zero or once per occurrence (ZOOPS), minimum width: 6, maximum width: 50 and background model: zero-order model of sequences. Further, the biological significance of the predicted MEME motifs was investigated using the GOMo version 5.3.3 [96] provided in the MEME Suite. Additionally, the retrieved promoter sequences were searched against the PLACE database to identify overrepresented *cis*-acting regulatory elements [97].

Abbreviations

ABI3
ABSCISIC ACID INSENSITIVE 3
AGL
AGAMOUS-LIKE
AIL
AINTEGUMENTA-LIKE
BBM
BABY BOOM
bHLH
Basic helix-loop-helix
bZIP
BASIC LEUCINE-ZIPPER
C2H2
Cys2-His2
CRE
Cis-acting regulatory element
DEG
Differentially expressed gene
EMB
EMBRYO-DEFECTIVE

FC
Fold Change
FUS3
FUSCA3
GCN
Gene co-expression network
GEO
Gene Expression Omnibus
GO
Gene Ontology
HB
HOMEOBOX
IAA30
INDOLE-3-ACETIC ACID INDUCIBLE 30
JMJ
JUMONJI DOMAIN-CONTAINING PROTEIN
KAN3
KANADI 3
LEC
LEAFY COTYLEDON
ME
Module eigengene
MEME
Multiple Em for Motif Elicitation
miRNA
microRNA
PLACE
Plant cis-acting regulatory DNA elements
r
Pearson correlation coefficient
RUS
ROOT UV-B SENSITIVE
SE
Somatic embryogenesis
STO
SALT TOLERANCE
TF
Transcription factor
WGCNA
Weighted Gene Correlation Network Analysis

Declarations

Competing interests

The authors declare that they have no competing interests.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The data sets supporting the results of this article are included within the article, in supplementary material submitted with the article.

Funding

Not applicable.

Author contributions

KKD participated in the design of the study, performed GCN construction and analysis, and drafted the manuscript. JMD helped to interpret data and draft the manuscript. AMW conceived the study and participated in the design of the study, and helped to draft the manuscript. All authors have read and approved the final manuscript.

Corresponding author

Correspondence to A. M. Wickramasuriya (anushka@pts.cmb.ac.lk)

Acknowledgements

Not applicable.

References

1. Steward FC, Mapes MO, Mears K. Growth and organized development of cultured cells. II. Organization in cultures grown from freely suspended cells. Am J Bot. 1958;45:705–8.
2. Zimmerman JL. Somatic embryogenesis: a model for early development in higher plants. Plant Cell. 1993;5:1411–23.
3. Capron A, Chatfield S, Provart N, Berleth T. Embryogenesis: pattern formation from a single cell. The Arabidopsis Book. 2009;7:e0126.
4. de Vries SC, Weijers D. Plant embryogenesis. Curr Biol. 2017;27:R870–3.
5. Etienne H. Somatic embryogenesis protocol: Coffee (*Coffea arabica* L. and *C. canephora* P.). In: Jain SM, Gupta PK, editors. Protocol for somatic embryogenesis in woody plants. Berlin/Heidelberg: Springer-Verlag; 2005. p. 167–79.
6. Steinmacher DA, Clement CR, Guerra MP. Somatic embryogenesis from immature peach palm inflorescence explants: towards development of an efficient protocol. Plant Cell Tiss Organ Cult. 2007;89:15–22.
7. Manrique-Trujillo S, Díaz D, Reaño R, Ghislain M, Kreuze J. Sweetpotato plant regeneration via an improved somatic embryogenesis protocol. Sci Hortic. 2013;161:95–100.
8. Vinoth S, Gurusaravanan P, Jayabalan N. Optimization of somatic embryogenesis protocol in *Lycopersicon esculentum* L. using plant growth regulators and seaweed extracts. J Appl Phycol. 2014;26:1527–37.
9. Méndez-Hernández HA, Ledezma-Rodríguez M, Avilez-Montalvo RN, Juárez-Gómez YL, Skeete A, Avilez-Montalvo J, et al. Signaling overview of plant somatic embryogenesis. Front Plant Sci. 2019;10:77.
10. Dantu PK, Tomar UK. Somatic embryogenesis. In: Tripathi G, editor. Cellular and biochemical science. New Delhi: IK International House Pvt Ltd; 2010. p. 892–908.
11. El-Esawi MA. Nonzygotic embryogenesis for plant development. In: Anis M, Ahmad N, editors. Plant tissue culture: propagation, conservation and crop improvement. Singapore: Springer Singapore; 2016. p. 583–98.
12. Zeng F, Zhang X, Cheng L, Hu L, Zhu L, Cao J, et al. A draft gene regulatory network for cellular totipotency reprogramming during plant somatic embryogenesis. Genomics. 2007;90:620–8.
13. Smertenko A, Bozhkov P. The life and death signalling underlying cell fate determination during somatic embryogenesis. In: Nick P, Opatrny Z, editors. Applied plant cell biology. Berlin, Heidelberg: Springer Berlin Heidelberg; 2014. p. 131–78.
14. Cetz-Chel JE, Loyola-Vargas VM. Transcriptome profile of somatic embryogenesis. In: Loyola-Vargas VM, Ochoa-Alejo N, editors. Somatic embryogenesis: fundamental aspects and applications. Cham: Springer International Publishing; 2016. p. 39–52.

15. Hecht V, Vielle-Calzada J-P, Hartog MV, Schmidt EDL, Boutilier K, Grossniklaus U, et al. The Arabidopsis *Somatic Embryogenesis Receptor Kinase 1* gene is expressed in developing ovules and embryos and enhances embryogenic competence in culture. Plant Physiol. 2001;127:803–16.
16. Yang X, Zhang X. Regulation of somatic embryogenesis in higher plants. CRC Crit Rev Plant Sci. 2010;29:36–57.
17. Gaj MD, Zhang S, Harada JJ, Lemaux PG. Leafy cotyledon genes are essential for induction of somatic embryogenesis of *Arabidopsis*. Planta. 2005;222:977–88.
18. Ikeda M, Takahashi M, Fujiwara S, Mitsuda N, Ohme-Takagi M. Improving the efficiency of adventitious shoot induction and somatic embryogenesis via modification of *WUSCHEL* and *LEAFY COTYLEDON 1*. Plants. 2020;9:1434.
19. Lotan T, Ohto M, Yee KM, West MAL, Lo R, Kwong RW, et al. *Arabidopsis* LEAFY COTYLEDON1 is sufficient to induce embryo development in vegetative cells. Cell. 1998;93:1195–205.
20. Stone SL, Kwong LW, Yee KM, Pelletier J, Lepiniec L, Fischer RL, et al. *LEAFY COTYLEDON2* encodes a B3 domain transcription factor that induces embryo development. Proc Natl Acad Sci USA. 2001;98:11806–11.
21. Stone SL, Braybrook SA, Paula SL, Kwong LW, Meuser J, Pelletier J, et al. *Arabidopsis* LEAFY COTYLEDON1 induces maturation traits and auxin activity: implications for somatic embryogenesis. Proc Natl Acad Sci USA. 2008;105:3151–6.
22. Boutilier K, Offringa R, Sharma VK, Kieft H, Ouellet T, Zhang L, et al. Ectopic expression of BABY BOOM triggers a conversion from vegetative to embryonic growth. Plant Cell. 2002;14:1737–49.
23. Zuo J, Niu Q-W, Frugis G, Chua N-H. The *WUSCHEL* gene promotes vegetative-to-embryonic transition in *Arabidopsis*. Plant J. 2002;30:349–59.
24. Wickramasuriya AM, Dunwell JM. Global scale transcriptome analysis of *Arabidopsis* embryogenesis *in vitro*. BMC Genomics. 2015;16:301.
25. Indoliya Y, Tiwari P, Chauhan AS, Goel R, Shri M, Bag SK, et al. Decoding regulatory landscape of somatic embryogenesis reveals differential regulatory networks between *japonica* and *indica* rice subspecies. Sci Rep. 2016;6:23050.
26. Singla B, Tyagi AK, Khurana JP, Khurana P. Analysis of expression profile of selected genes expressed during auxin-induced somatic embryogenesis in leaf base system of wheat (*Triticum aestivum*) and their possible interactions. Plant Mol Biol. 2007;65:677–92.
27. Zeng F, Zhang X, Zhu L, Tu L, Guo X, Nie Y. Isolation and characterization of genes associated to cotton somatic embryogenesis by suppression subtractive hybridization and macroarray. Plant Mol Biol. 2006;60:167–83.
28. Salvo SAGD, Hirsch CN, Buell CR, Kaeppler SM, Kaeppler HF. Whole transcriptome profiling of maize during early somatic embryogenesis reveals altered expression of stress factors and embryogenesis-related genes. PLoS One. 2014;9:e111407.
29. Rajesh MK, Fayas TP, Naganeeswaran S, Rachana KE, Bhavyashree U, Sajini KK, et al. De novo assembly and characterization of global transcriptome of coconut palm (*Cocos nucifera* L.)

embryogenic calli using Illumina paired-end sequencing. Protoplasma. 2016;253:913–28.

30. van Dam S, Võsa U, van der Graaf A, Franke L, de Magalhães JP. Gene co-expression analysis for functional classification and gene–disease predictions. Brief Bioinform. 2017;19:575–92.
31. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol. 2005;4.
32. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559.
33. Gao P, Xiang D, Quilichini TD, Venglat P, Pandey PK, Wang E, et al. Gene expression atlas of embryo development in *Arabidopsis*. Plant Reprod. 2019;32:93–104.
34. Shaik R, Ramakrishna W. Genes and co-expression modules common to drought and bacterial stress responses in *Arabidopsis* and rice. PLoS One. 2013;8:e77261.
35. Tai Y, Liu C, Yu S, Yang H, Sun J, Guo C, et al. Gene co-expression network analysis reveals coordinated regulation of three characteristic secondary biosynthetic pathways in tea plant (*Camellia sinensis*). BMC Genomics. 2018;19:616.
36. Zhu M, Xie H, Wei X, Dossa K, Yu Y, Hui S, et al. WGCNA analysis of salt-responsive core transcriptome identifies novel hub genes in rice. Genes. 2019;10:719.
37. Meinke DW. Genome-wide identification of *EMBRYO-DEFECTIVE* (*EMB*) genes required for growth and development in *Arabidopsis*. New Phytol. 2020;226:306–25.
38. Magnani E, Jiménez-Gómez JM, Soubigou-Taconnat L, Lepiniec L, Fiume E. Profiling the onset of somatic embryogenesis in *Arabidopsis*. BMC Genomics. 2017;18:998.
39. De Clercq I, Van de Velde J, Luo X, Liu L, Storme V, Van Bel M, et al. Integrative inference of transcriptional networks in *Arabidopsis* yields novel ROS signalling regulators. Nat Plants. 2021;7:500–13.
40. Consortium TGO. Creating the Gene Ontology resource: design and implementation. Genome Res. 2001;11:1425–33.
41. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet. 2000;25:25–9.
42. Rue-Albrecht K, McGettigan PA, Hernández B, Nalpas NC, Magee DA, Parnell AC, et al. GOexpress: an R/Bioconductor package for the identification and visualisation of robust gene ontology signatures through supervised learning of gene expression data. BMC Bioinformatics. 2016;17:126.
43. West MAL, Harada JJ. Embryogenesis in higher plants: an overview. Plant Cell. 1993;5:1361–9.
44. Williams EG, Maheswaran G. Somatic embryogenesis: factors influencing coordinated behaviour of cells as an embryogenic group. Ann Bot. 1986;57:443–62.
45. Smertenko A, Bozhkov PV. Somatic embryogenesis: life and death processes during apical–basal patterning. J Exp Bot. 2014;65:1343–60.
46. Zavattieri MA, Frederico AM, Lima M, Sabino R, Arnholdt-Schmitt B. Induction of somatic embryogenesis as an example of stress-related plant reactions. Electro Journal of Biotech.

2010;13:1–9.

47. Jin F, Hu L, Yuan D, Xu J, Gao W, He L, et al. Comparative transcriptome analysis between somatic embryos (SEs) and zygotic embryos in cotton: evidence for stress response functions in SE development. Plant Biotechnol J. 2014;12:161–73.
48. Qiu J, Du Z, Wang Y, Zhou Y, Zhang Y, Xie Y, et al. Weighted gene co-expression network analysis reveals modules and hub genes associated with the development of breast cancer. Medicine. 2019;98:e14345.
49. Liu Y, Gu H-Y, Zhu J, Niu Y-M, Zhang C, Guo G-L. Identification of hub genes and key pathways associated with bipolar disorder based on weighted gene co-expression network analysis. Front Physiol. 2019;10:1081.
50. Zhu Z, Jin Z, Deng Y, Wei L, Yuan X, Zhang M, et al. Co-expression network analysis identifies four hub genes associated with prognosis in soft tissue sarcoma. Front Genet. 2019;10:37.
51. Du J, Wang S, He C, Zhou B, Ruan Y-L, Shou H. Identification of regulatory networks and hub genes controlling soybean seed set and size using RNA sequencing analysis. J Exp Bot. 2017;68:1955–72.
52. Zhang X, Feng H, Li Z, Li D, Liu S, Huang H, et al. Application of weighted gene co-expression network analysis to identify key modules and hub genes in oral squamous cell carcinoma tumorigenesis. Onco Targets Ther. 2018;11:6001–21.
53. Wang Q, Zeng X, Song Q, Sun Y, Feng Y, Lai Y. Identification of key genes and modules in response to cadmium stress in different rice varieties and stem nodes by weighted gene co-expression network analysis. Sci Rep. 2020;10:9525.
54. Zhang F, Wang L, Bai P, Wei K, Zhang Y, Ruan L, et al. Identification of regulatory networks and hub genes controlling nitrogen uptake in tea plants [*Camellia sinensis* (L.) O. Kuntze]. J Agric Food Chem. 2020;68:2445–56.
55. Causier B, Ashworth M, Guo W, Davies B. The TOPLESS interactome: a framework for gene repression in Arabidopsis. Plant Physiol. 2012;158:423–38.
56. Gulzar B, Mujib A, Malik MQ, Sayeed R, Mamgain J, Ejaz B. Genes, proteins and other networks regulating somatic embryogenesis in plants. J Genet Eng Biotechnol. 2020;18:31.
57. Wójcikowska B, Gaj MD. Expression profiling of *AUXIN RESPONSE FACTOR* genes during somatic embryogenesis induction in Arabidopsis. Plant Cell Rep. 2017;36:843–58.
58. Perry N, Leasure CD, Tong H, Duarte EM, He Z-H. RUS6, a DUF647-containing protein, is essential for early embryonic development in *Arabidopsis thaliana*. BMC Plant Biol. 2021;21:232.
59. Becerra C, Jahrmann T, Puigdomènech P, Vicient CM. Ankyrin repeat-containing proteins in *Arabidopsis*: characterization of a novel and abundant group of genes coding ankyrin-transmembrane proteins. Gene. 2004;340:111–21.
60. Yan J, Wang J, Zhang H. An ankyrin repeat-containing protein plays a role in both disease resistance and antioxidation metabolism: an *Arabidopsis* protein in disease resistance. Plant J. 2002;29:193–202.

61. Zhang H, Scheirer DC, Fowle WH, Goodman HM. Expression of antisense or sense RNA of an ankyrin repeat-containing gene blocks chloroplast differentiation in arabidopsis. Plant Cell. 1992;4:1575–88.
62. Albert S, Despres B, Guilleminot J, Bechtold N, Pelletier G, Delseny M, et al. The *EMB*506 gene encodes a novel ankyrin repeat containing protein that is essential for the normal development of Arabidopsis embryos. Plant J. 1999;17:169–79.
63. Poon S, Heath RL, Clarke AE. A chimeric arabinogalactan protein promotes somatic embryogenesis in cotton cell culture. Plant Physiol. 2012;160:684–95.
64. Basu D, Tian L, Wang W, Bobbs S, Herock H, Travers A, et al. A small multigene hydroxyproline-*O*-galactosyltransferase family functions in arabinogalactan-protein glycosylation, growth and development in Arabidopsis. BMC Plant Biol. 2015;15:295.
65. Duchow S, Dahlke RI, Geske T, Blaschek W, Classen B. Arabinogalactan-proteins stimulate somatic embryogenesis and plant propagation of *Pelargonium sidoides*. Carbohydr Polym. 2016;152:149–55.
66. Stålberg K, Ellerström M, Ezcurra I, Ablov S, Rask L. Disruption of an overlapping E-box/ABRE motif abolished high transcription of the *napA* storage-protein promoter in transgenic *Brassica napus* seeds. Planta. 1996;199:515–19.
67. Heim MA. The basic helix-loop-helix transcription factor family in plants: a genome-wide study of protein structure and functional diversity. Mol Biol Evol. 2003;20:735–47.
68. Gliwicka M, Nowak K, Balazadeh S, Mueller-Roeber B, Gaj MD. Extensive modulation of the transcription factor transcriptome during somatic embryogenesis in *Arabidopsis thaliana*. PLoS One. 2013;8:e69261.
69. Allen RD, Bernier F, Lessard PA, Beachy RN. Nuclear factors interact with a soybean beta-conglycinin enhancer. Plant Cell. 1989;1:623–31.
70. Becker MG, Chan A, Mao X, Girard IJ, Lee S, Elhiti M, et al. Vitamin C deficiency improves somatic embryo development through distinct gene regulatory networks in *Arabidopsis*. J Exp Bot. 2014;65:5903–18.
71. Liscum E, Reed JW. Genetics of Aux/IAA and ARF action in plant growth and development. Plant Mol Biol. 2002;49:387–400.
72. Braybrook SA, Stone SL, Park S, Bui AQ, Le BH, Fischer RL, et al. Genes directly regulated by LEAFY COTYLEDON2 provide insight into the control of embryo maturation and somatic embryogenesis. Proc Natl Acad Sci USA. 2006;103:3468–73.
73. Wójcik AM, Wójcikowska B, Gaj MD. Current perspectives on the auxin-mediated genetic network that controls the induction of somatic embryogenesis in plants. Int J Mol Sci. 2020;21:1333.
74. Horstman A, Li M, Heidmann I, Weemen M, Chen B, Muino JM, et al. The BABY BOOM transcription factor activates the LEC1-ABI3-FUS3-LEC2 network to induce somatic embryogenesis. Plant Physiol. 2017;175:848–57.
75. Zheng Y, Ren N, Wang H, Stromberg AJ, Perry SE. Global identification of targets of the *Arabidopsis* MADS domain protein AGAMOUS-Like15. Plant Cell. 2009;21:2563–77.

76. Joshi S, Keller C, Perry SE. The EAR motif in the *Arabidopsis* MADS transcription factor AGAMOUS-Like 15 is not necessary to promote somatic embryogenesis. Plants. 2021;10:758.
77. Nowak K, Gaj MD. Transcription factors in the regulation of somatic embryogenesis. In: Loyola-Vargas VM, Ochoa-Alejo N, editors. Somatic embryogenesis: fundamental aspects and applications. Cham: Springer International Publishing; 2016. p. 53–79.
78. Pant P, Iqbal Z, Pandey BK, Sawant SV. Genome-wide comparative and evolutionary analysis of Calmodulin-binding Transcription Activator (CAMTA) family in *Gossypium* species. Sci Rep. 2018;8:5573.
79. Kaplan-Levy RN, Brewer PB, Quon T, Smyth DR. The trihelix family of transcription factors – light, stress and development. Trends Plant Sci. 2012;17:163–71.
80. Siddiqui ZH, Abbas ZK, Ansari MW, Khan MN. The role of miRNA in somatic embryogenesis. Genomics. 2019;111:1026–33.
81. Alves A, Cordeiro D, Correia S, Miguel C. Small non-coding RNAs at the crossroads of regulatory pathways controlling somatic embryogenesis in seed plants. Plants. 2021;10:504.
82. Wójcik AM, Gaj MD. miR393 contributes to the embryogenic transition induced *in vitro* in Arabidopsis via the modification of the tissue sensitivity to auxin treatment. Planta. 2016;244:231–43.
83. Szyrajew K, Bielewicz D, Dolata J, Wójcik AM, Nowak K, Szczygieł-Sommer A, et al. MicroRNAs are intensively regulated during induction of somatic embryogenesis in Arabidopsis. Front Plant Sci. 2017;8.
84. Nowak K, Morończyk J, Wójcik A, Gaj MD. AGL15 controls the embryogenic reprogramming of somatic cells in *Arabidopsis* through the histone acetylation-mediated repression of the miRNA biogenesis genes. Int J Mol Sci. 2020;21:6733.
85. Chen X, Xu X, Shen X, Li H, Zhu C, Chen R, et al. Genome-wide investigation of DNA methylation dynamics reveals a critical role of DNA demethylation during the early somatic embryogenesis of *Dimocarpus longan* Lour. Tree Physiol. 2020;40:1807–26.
86. Grzybkowska D, Nowak K, Gaj MD. Hypermethylation of auxin-responsive motifs in the promoters of the transcription factor genes accompanies the somatic embryogenesis induction in Arabidopsis. Int J Mol Sci. 2020;21:6849.
87. Ji L, Mathioni SM, Johnson S, Tucker D, Bewick AJ, Do Kim K, et al. Genome-wide reinforcement of DNA methylation occurs during somatic embryogenesis in soybean. Plant Cell. 2019;31:2315–31.
88. Rodríguez-Sanz H, Moreno-Romero J, Solís M-T, Köhler C, Risueño MC, Testillano PS. Changes in histone methylation and acetylation during microspore reprogramming to embryogenesis occur concomitantly with Bn*HKMT* and Bn*HAT* expression and are associated with cell totipotency, proliferation, and differentiation in *Brassica napus*. Cytogenet Genome Res. 2014;143:209–18.
89. Wójcikowska B, Botor M, Morończyk J, Wójcik AM, Nodzyński T, Karcz J, et al. Trichostatin A triggers an embryogenic transition in Arabidopsis explants via an auxin-related pathway. Front Plant Sci. 2018;9:1353.

90. Kao P, Schon MA, Mosiolek M, Nodine MD. Gene expression variation in *Arabidopsis* embryos at single-nucleus resolution.Development. 2021;148:dev199589.
91. Langfelder P, Luo R, Oldham MC, Horvath S. Is my network module preserved and reproducible? PLoS Comput Biol. 2011;7:e1001057.
92. Tian T, Liu Y, Yan H, You Q, Yi X, Du Z, et al. agriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update. Nucleic Acids Res. 2017;45:W122–9.
93. Xiang D, Venglat P, Tibiche C, Yang H, Risseeuw E, Cao Y, et al. Genome-wide analysis reveals gene expression and metabolic network dynamics during embryo development in *Arabidopsis*. Plant Physiol. 2011;156:346–56.
94. Pikaard CS, Mittelsten Scheid O. Epigenetic regulation in plants. Cold Spring Harb Perspect Biol. 2014;6:a019315.
95. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 2009;37 suppl_2:W202-208.
96. Buske FA, Bodén M, Bauer DC, Bailey TL. Assigning roles to DNA regulatory motifs using comparative genomics. Bioinformatics. 2010;26:860–6.
97. Higo K, Ugawa Y, Iwamoto M, Korenaga T. Plant cis-acting regulatory DNA elements (PLACE) database: 1999. Nucleic Acids Res. 1999;27:297–300.

Tables

Tables 1-4 are in the supplementary files section.

Figures

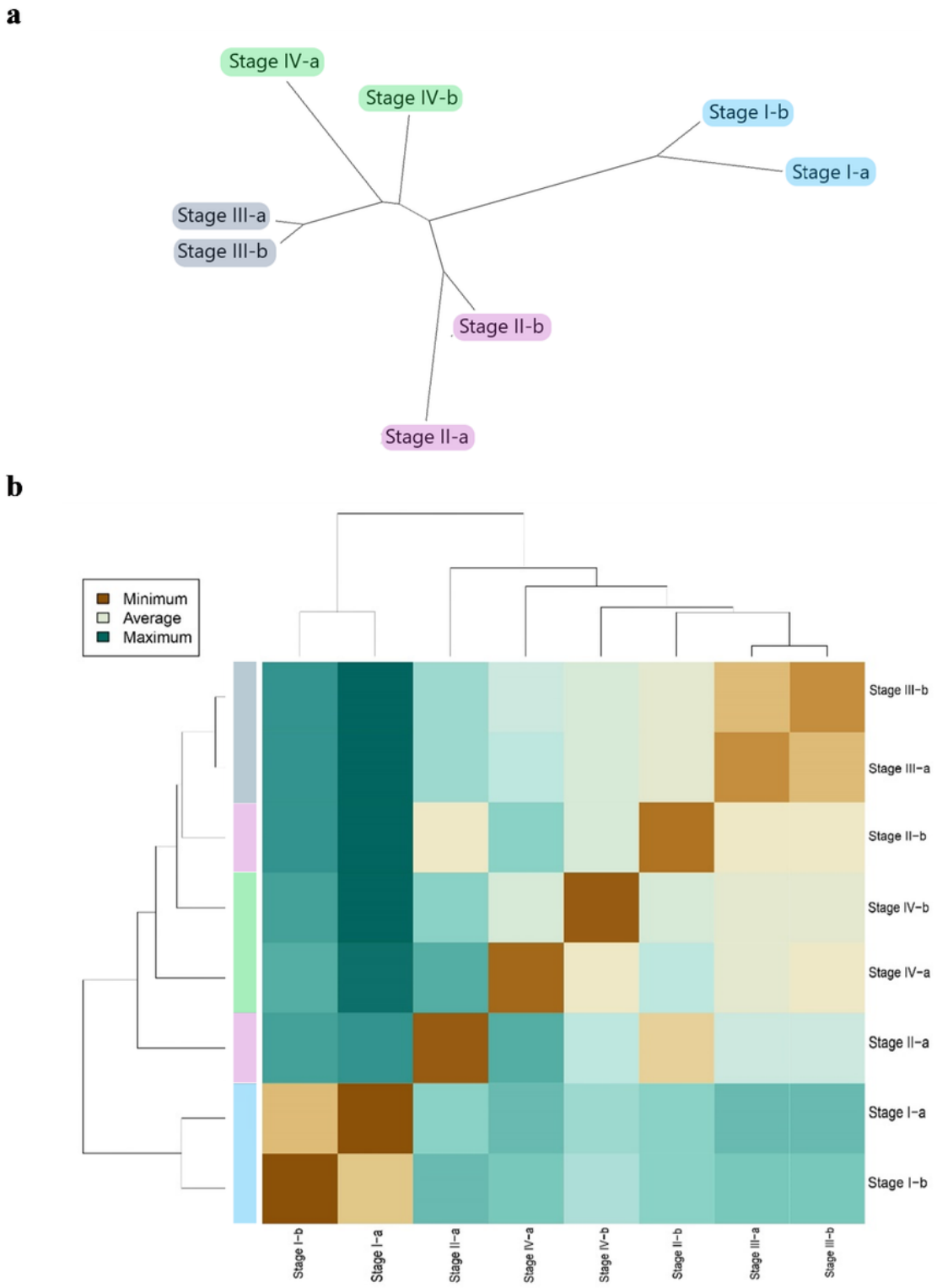

Figure 1

Hierarchical clustering of somatic embryo transcriptomes based on their Euclidean distance using average linkage clustering (replicates of each stage are labeled as ‘a’ and ‘b’). **a.** Unrooted hierarchical clustering dendrogram (the length between nodes corresponds to the distance between samples); **b.** Hierarchical clustering heatmap visualizing the correlations between the samples by colour as indicated in the legend. The degree of correlation is indicated by minimum, average and maximum.

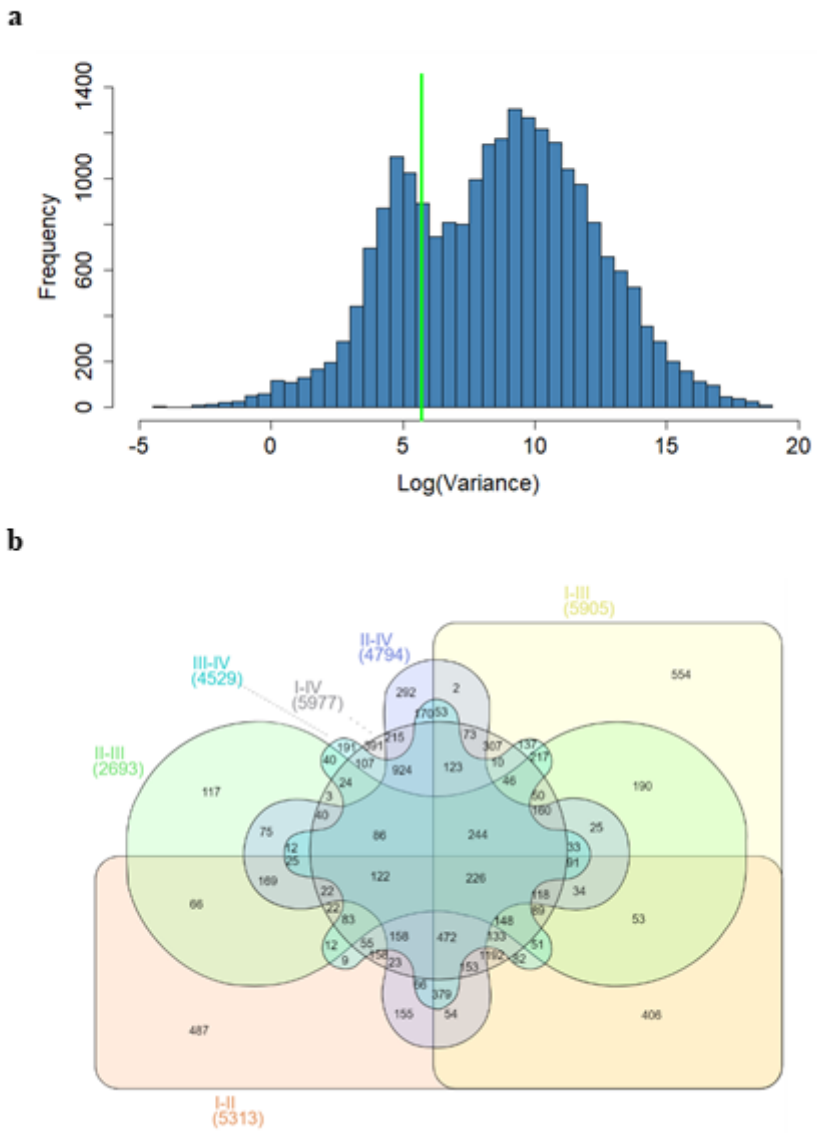

Figure 2

Filtering of genes **a.** Histogram for the variance of gene expression. The green line indicates the filter threshold (i.e. variance greater than 0.25 quantile); **b.** Distribution of DEGs between different somatic embryo developmental stages.

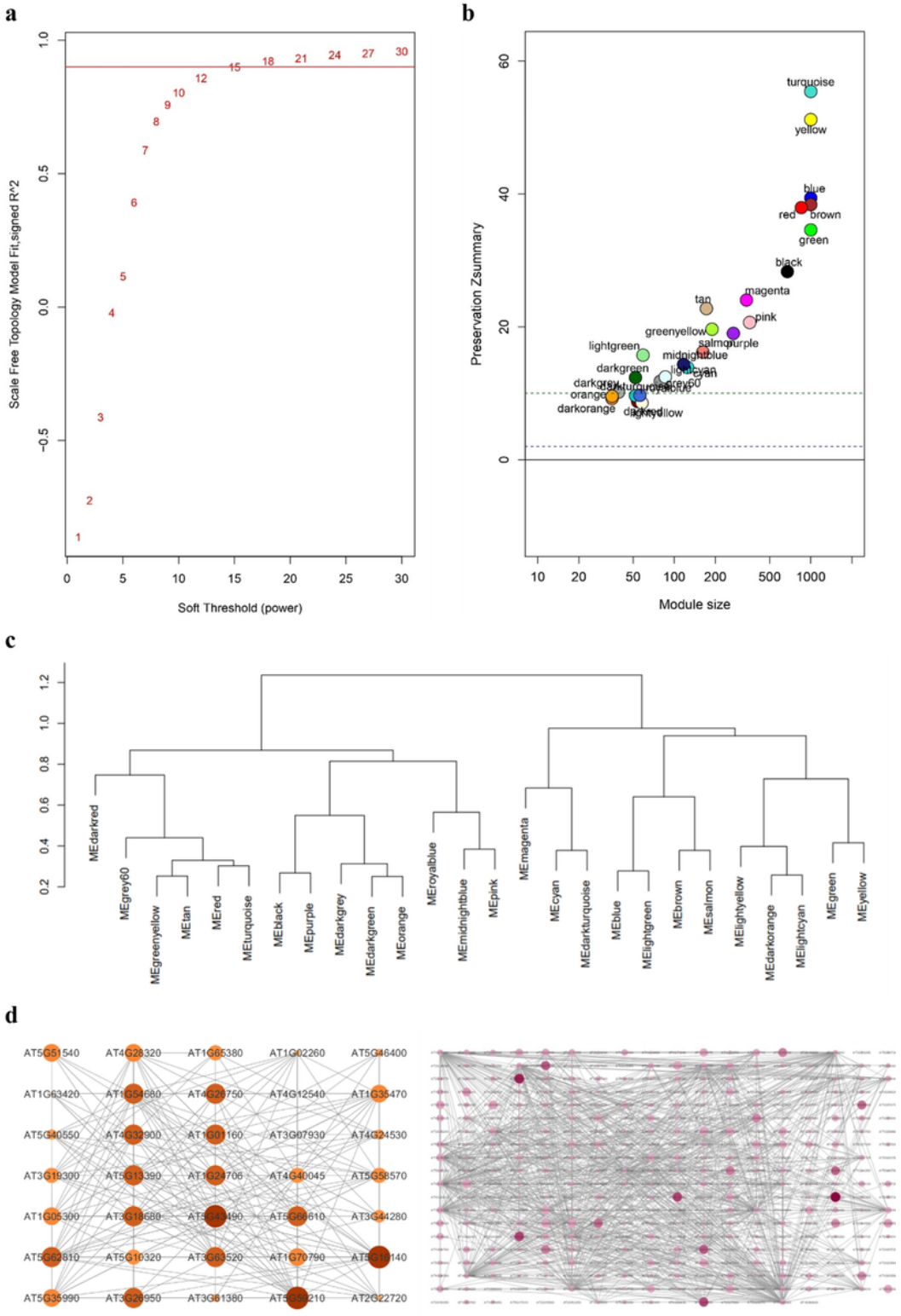

Figure 3

Construction of the draft GCN for SE. **a.** Network topology for different soft-thresholding powers; **b.** Module preservation statistics; **c.** Hierarchical clustering dendrogram of MEs; **d.** Visualization of network module connections of the intramodular genes in the orange (left) and pink (right) modules. The size of the circle is proportional to the degree of connectivity of the node. Darker shades indicate high gene connectivity.

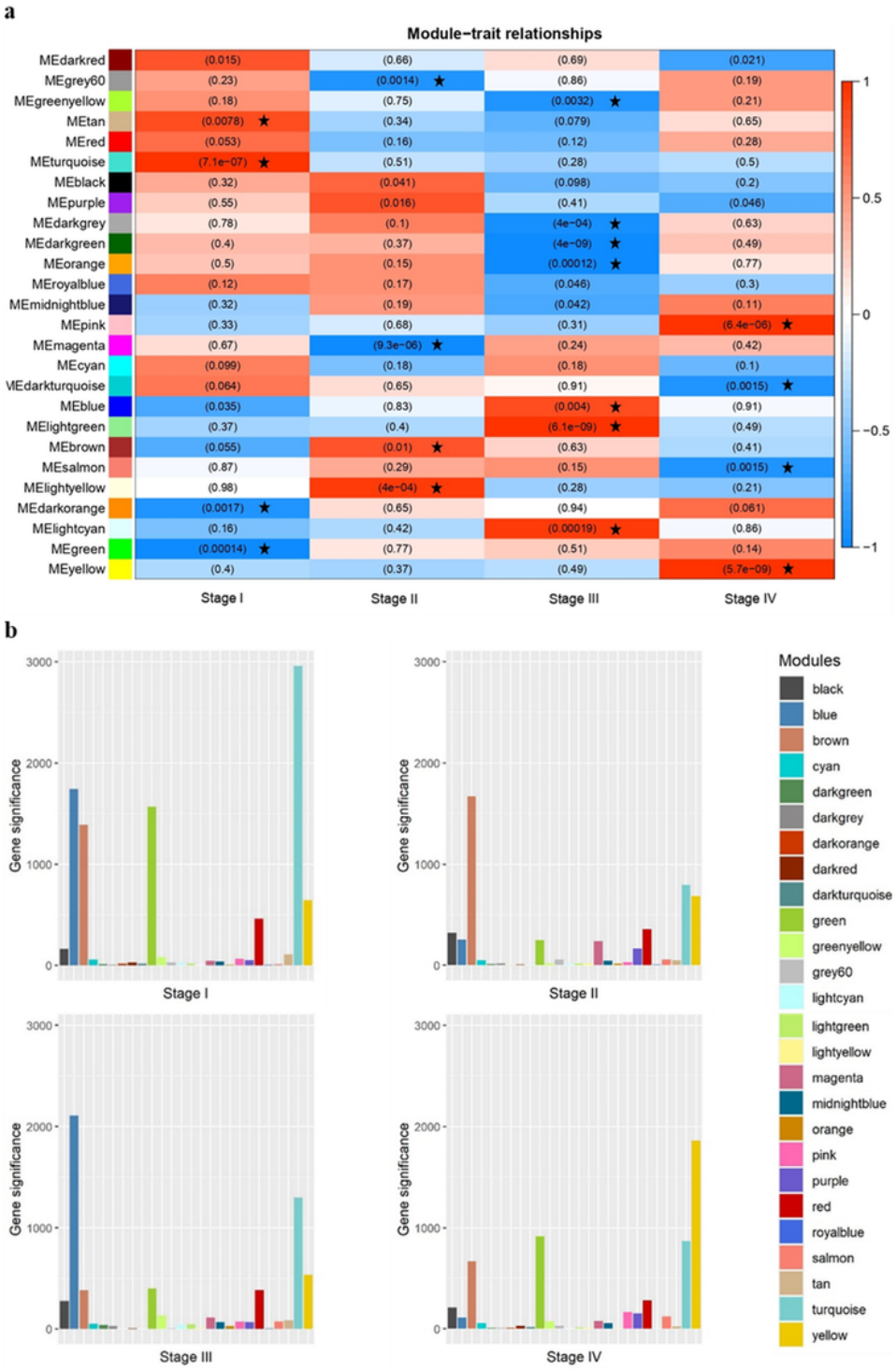

Figure 4

Stage-specific gene modules detected by WGCNA. **a.** Module-trait relationship heatmap. Each row corresponds to a module and each column corresponds to a stage. The degree of correlation is illustrated with the colour legend. The numbers in the table correspond to the *p*-value. Modules that are significantly associated with each somatic embryo development stage ($|r| > 0.8$ and $p\text{-value} \leq 0.01$) are indicated by

an asterisk; **b.** Gene significance values of co-expression modules related to different somatic embryo developmental stages.

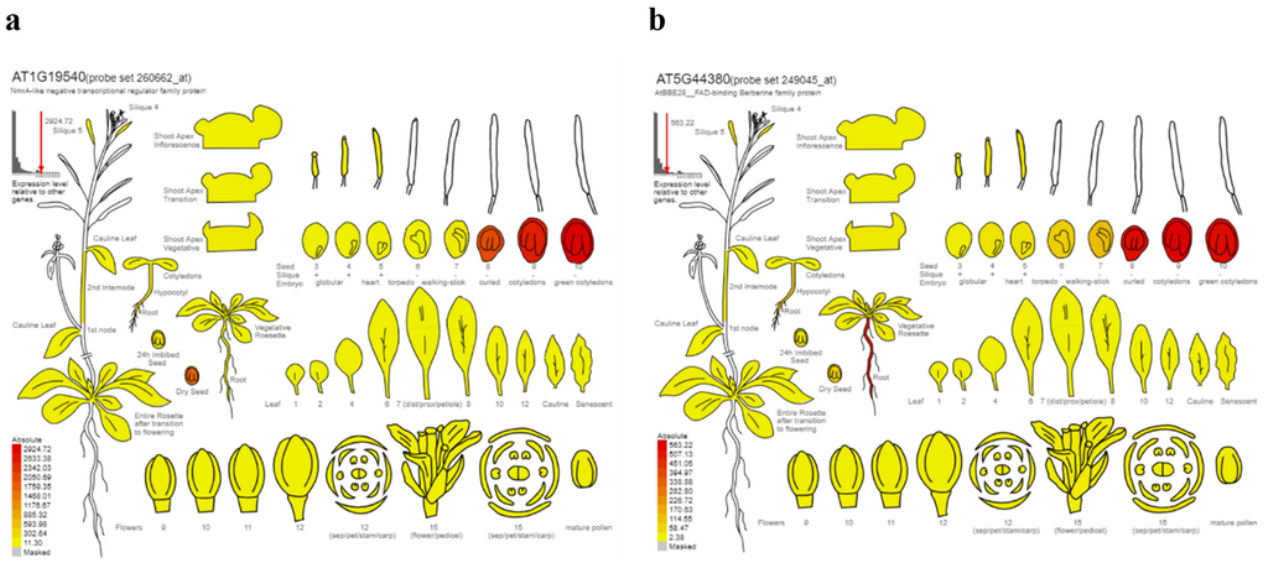

Figure 5

Expression patterns of two hub genes, *AT1G19540* and *AT5G44380* when viewed through the *Arabidopsis* eFP browser. The normalized expression value for each gene is colour-coded as indicated by the legend.

Figure 6

Transcript abundance extracted from the somatic embryo transcriptome dataset, E-MTAB-2465 for the hub genes. The hub genes significantly up-regulated ($log_2\ FC \geq 2.0$) and down-regulated ($log_2\ FC \leq -2.0$) in somatic embryonic tissues compared to leaf tissues are indicated with yellow asterisks and red diamonds, respectively.

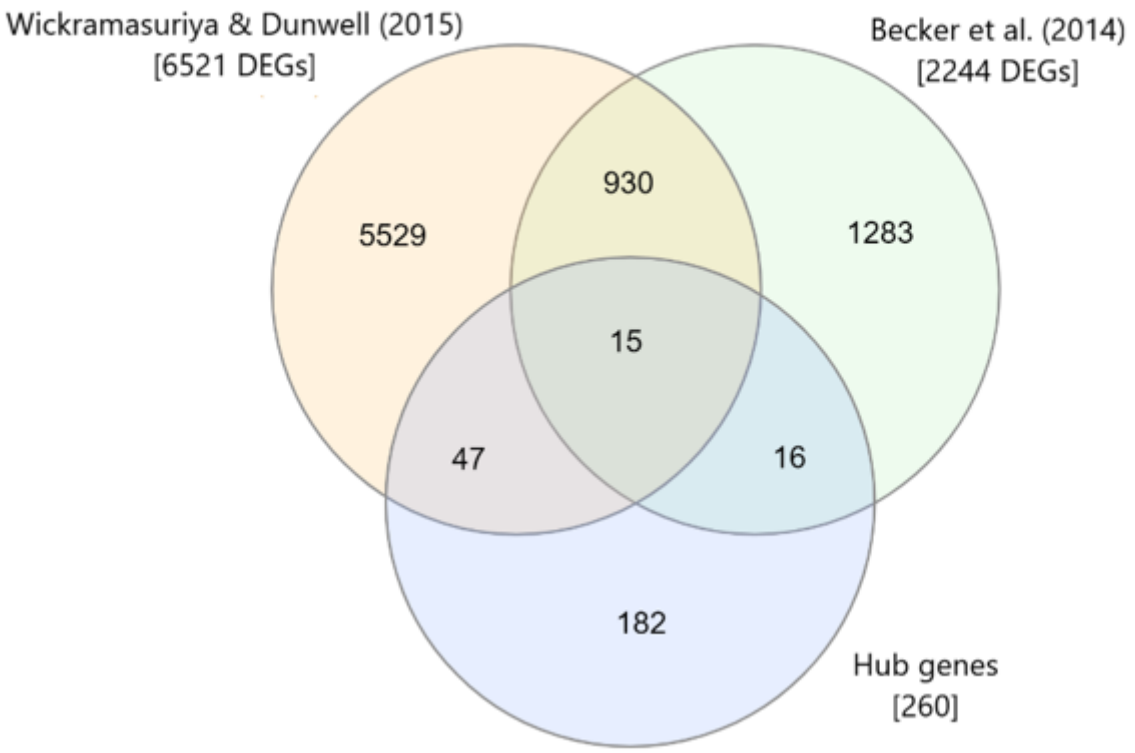

Figure 7

Venn diagram indicating the intersection of hub genes and DEGs ($|log_2 FC| \geq 2.0$ and $p\text{-}value < 0.05$) obtained from E-MTAB-2465 (Wickramasuriya and Dunwell, 2015) and GSE48915 (Becker *et al.*, 2014).

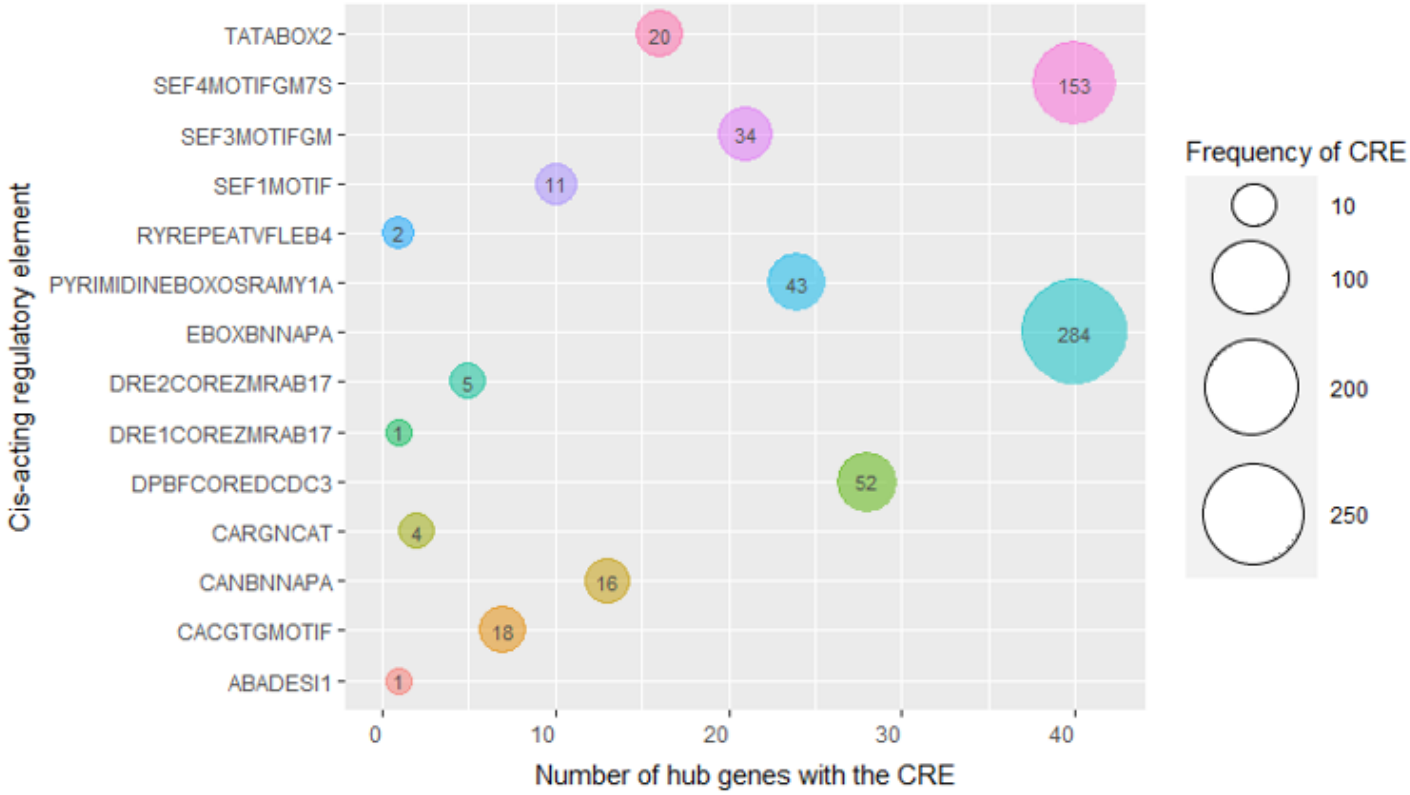

Figure 8

The distribution of several important plant CREs present in the promoter regions of functionally uncharacterized hub genes. The number of hub genes that contain the relevant CRE in their promoter region is indicated by the *x*-axis. The size of the circle depicts the occurrence of the CREs within the promoter regions of the hub genes (as indicated within the circle).

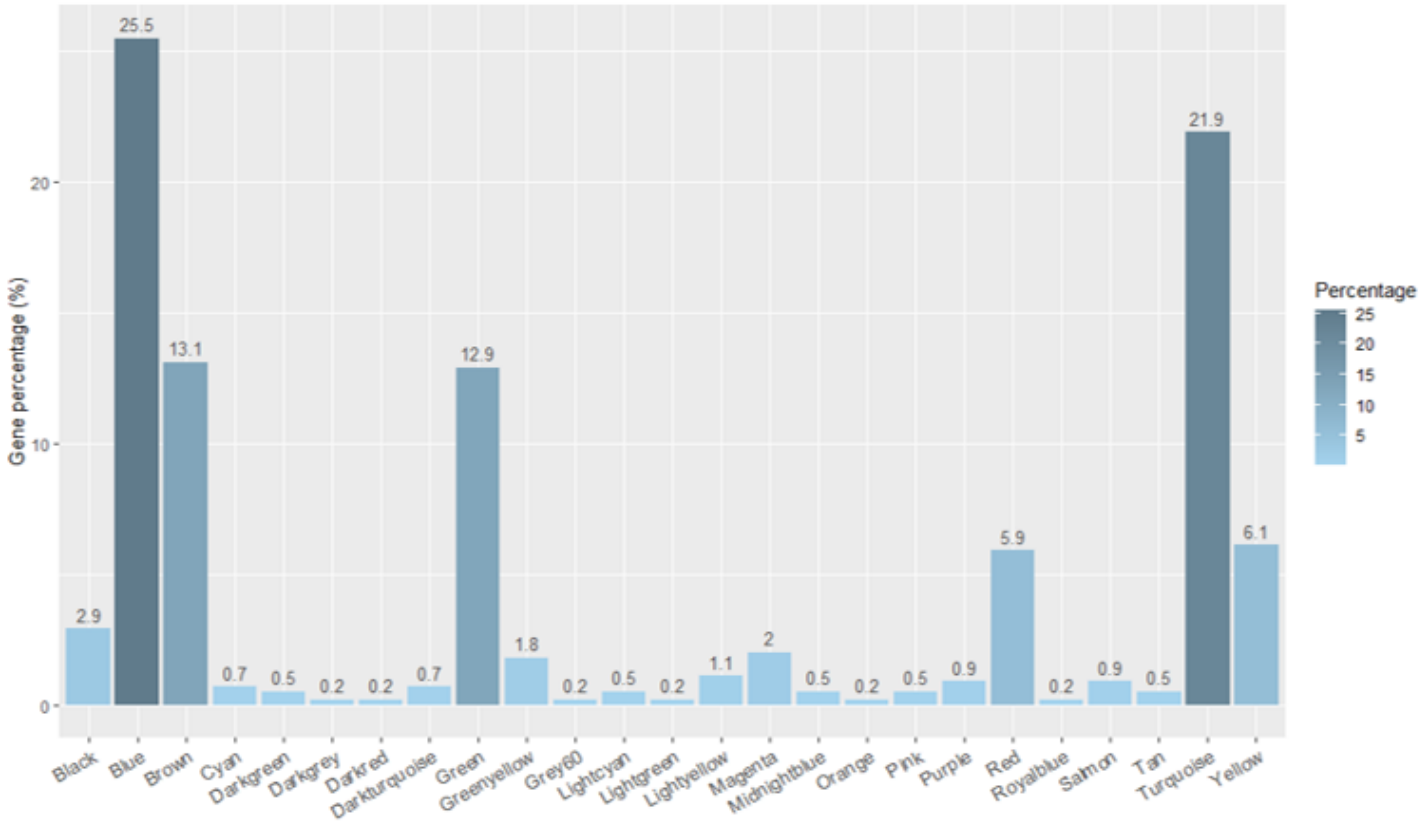

Figure 9

Distribution of *EMB* genes across gene modules. The coloured bars represent the ratio between the number of *EMB* genes in each module and the total number of *EMB* genes in the network.

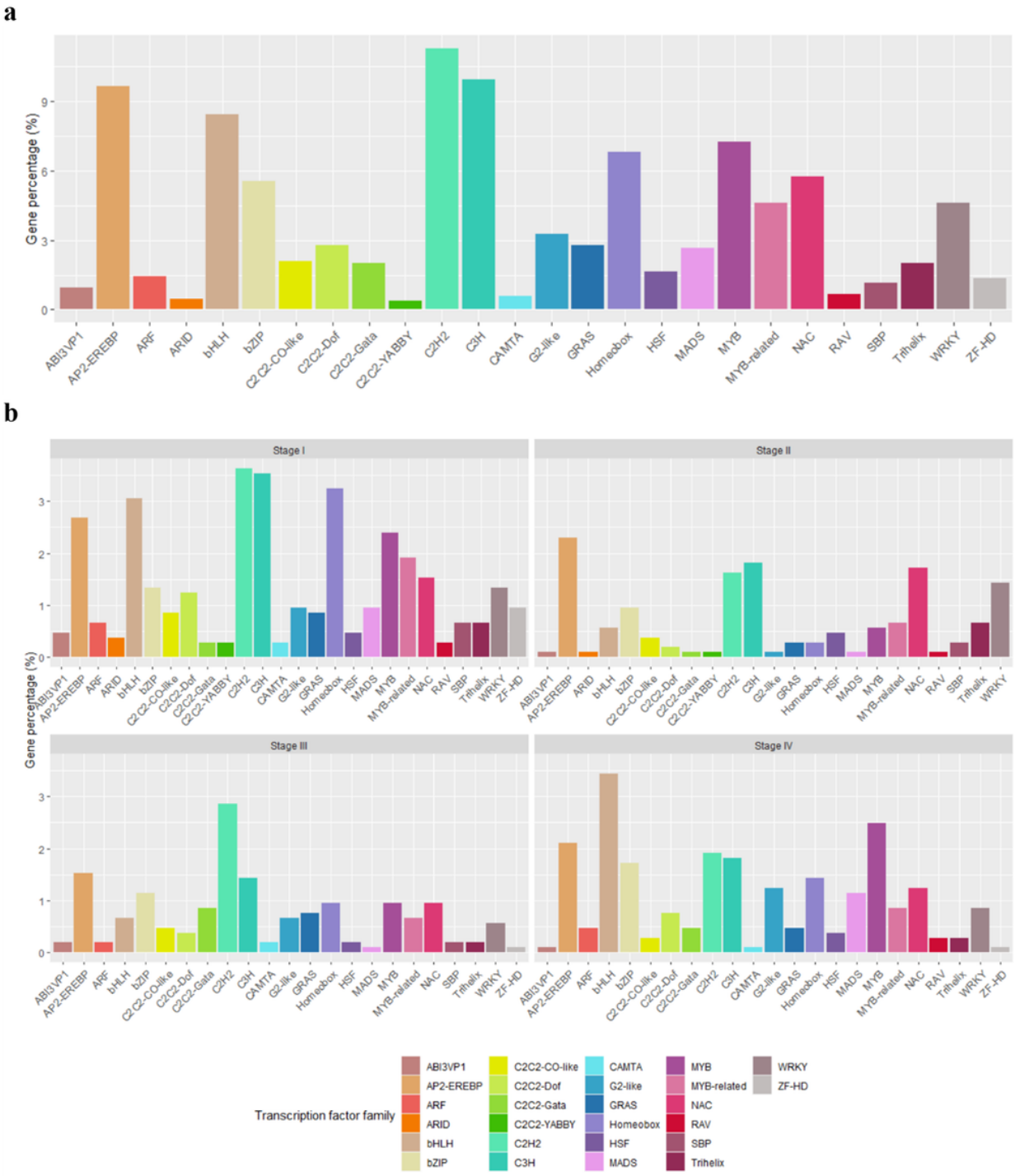

Figure 10

Distribution of TFs in SE. **a.** Overall distribution of TFs. The percentage is calculated as the ratio of TFs belonging to each family with respect to the total number of TFs in the network. **b.** Distribution of TFs across different somatic embryo developmental stages. The percentage is calculated for each stage as the ratio of TFs present in each family with respect to the total number of TFs in the network.

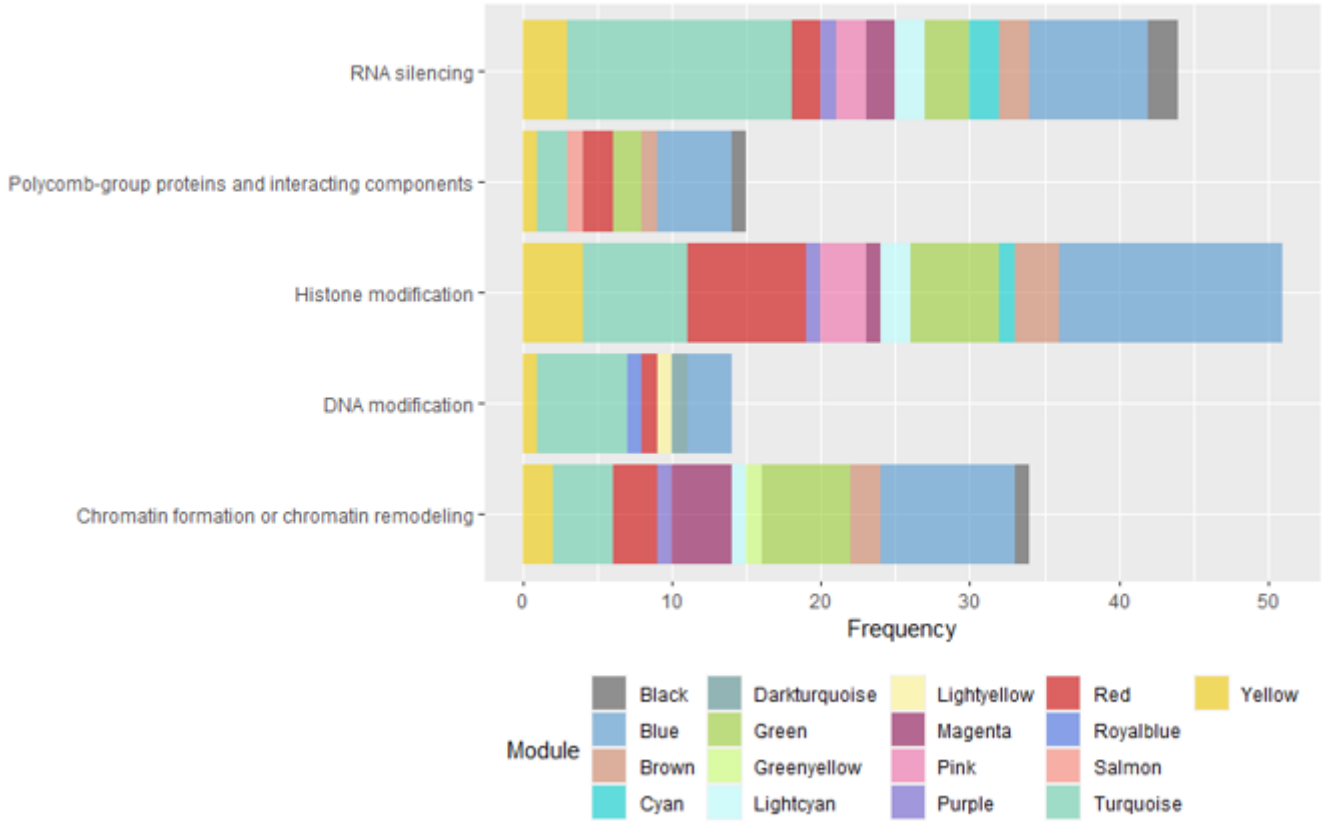

Figure 11

Distribution of genes encoding epigenetic regulators across the network modules

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1TableS1S2.xlsx
- Additionalfile2TableS3.xlsx
- Additionalfile3TableS4.xlsx
- Additionalfile4TableS5.xlsx
- Additionalfile5TableS6S7.xlsx
- Additionalfile6TableS8.xlsx
- ListofAdditionaldatafiles.docx
- Tables.docx