SANe: The Seed Active Network For Mining Transcriptional Regulatory Programs of Seed Development

Chirag Gupta[1], Arjun Krishnan[2,a,b], Eva Collakova[3], Pawel Wolinski[4] and Andy Pereira[1,2,*]

[1] Crop, Soil, and Environmental Sciences, University of Arkansas, Fayetteville, Arkansas 72701

[2] Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, Virginia 24061

[3] Department of Plant Pathology, Physiology, and Weed Science, Virginia Tech, Blacksburg, Virginia 24061

[4] High Performance Computing Center, University of Arkansas, Fayetteville, Arkansas 72701

[a] Current address: Computational Mathematics, Science, and Engineering, Michigan State University, MI 48824

[b] Current address: Biochemistry and Molecular Biology, Michigan State University, MI 48824

*Corresponding author: Andy Pereira

apereira@uark.edu

Short Title: SANe: Seed Active Network

Abstract

Seed development is an evolutionarily important phase of the plant life cycle that governs the fate of next progeny. Distinct sub-regions within seeds have diverse roles in protecting and nourishing the embryo as it enlarges, and for the synthesis of storage reserves that serve as an important source of nutrients and energy for germination. Several studies have revealed that transcription factors (TFs) act in fine coordination to regulate target genes that ensure proper maintenance, metabolism, and development of the embryo. Here, we present genome-wide predictions of seed-specific regulatory interactions between TFs and their target genes in the model plant *Arabidopsis thaliana*. The network is based on a panel of high-resolution seed-specific gene expression datasets and takes the form of a module-regulatory network. TFs that are well studied in the literature were often found at the top of the predicted ranks for the module that corresponds to their validated function role. Furthermore, we brought together a dedicated web resource for the systematic analysis of transcriptional-level regulatory programs underlying the development of seeds (https://plantstress-pereira.uark.edu/SANe/). The platform will enable biologists to query a subset of modules, TFs of interest, as well as analyze new transcriptomes to find modules significantly perturbed in their experiment.

Introduction

The evolutionary success of plants lies in their ability to produce seeds and aid their dispersal, which ensures the progression of generations. Seeds are complex organized structures that help plants pause their life cycle under unfavorable conditions, and resume growth once environmental conditions become favorable. Like all angiosperms, in Arabidopsis, a double fertilization event marks the beginning of seed development that progresses into the development of the embryo, endosperm and seed coat over a period of 20-21 days after pollination. These morphologically distinct sub-compartments within a seed play diverse roles and function in concert during the entire phase of seed formation. During maturation, the synthesis of storage reserves takes place and developmental programs like desiccation tolerance and dormancy are initiated. These seed storage reserves are the fuel for seedling emergence during germination.

Several transcription factors (TFs) that regulate various aspects of seed development as well as germination have been revealed by genetic screens (Grossniklaus et al., 1998; Lotan et al., 1998; Ogas et al., 1999; Johnson et al., 2002; To et al., 2006). Among these TFs, three members of the B3 super family, namely, *LEAFY COTYLEDON 2 (LEC2)*, *ABSCISIC ACID INSENSITIVE 3 (ABI3)* and *FUSCA3 (FUS3)*, along with two members of the LEC1-type, *LEC1* and *LEC1-LIKE* that together form the 'LAFL' network (Jia et al., 2013), are the most prominent players of seed maturation. However, the existing LAFL network is still incomplete and represents only a subset of regulatory networks active during seed development. The functional roles of several other TFs that express in seed tissues remains largely unknown. Although genetic interactions, functional redundancy and cooperativity between TFs will be more accurately revealed by genetic perturbations, an underpinning of seed regulatory networks from a computational standpoint will provide tools for quick identification and prioritization of candidates for experimentation *in vivo*.

DNA microarrays have served as efficient experimental systems for simultaneously probing genome-wide transcriptional level activities of specific cellular states. In recent years, an upsurge in the availability of these high-throughput gene expression datasets motivated coexpression based approaches to be applied for an understanding of gene functions. An integrative analysis of expression datasets enables the estimation of similarity in patterns of gene expression across a diverse set of experimental conditions. Genes with similar expression

profiles are grouped into clusters of coexpressed genes. Functional (Castillo-Davis and Hartl, 2003) and genomic (Huttenhower et al., 2009) annotations of these gene clusters then aid in making functional predictions of uncharacterized genes within these clusters (Childs et al., 2011). There are several such coexpression databases across many model organisms that are now being actively used in gene function prediction and gene prioritization for experimental assays in plants (Obayashi and Kinoshita, 2011; Sato et al., 2012; Yim et al., 2013; Aoki et al., 2016).

Coexpression networks, however, lack information about regulatory interactions represented in the expression data. Genes encoding regulatory proteins (e.g., TFs) coordinately regulate the biological functions of multiple target genes by directly interacting with their promoters and activating or repressing their expression. Since TFs are themselves transcriptionally regulated, they can also be targets of other TFs, giving the network a hierarchical structure (Ma et al., 2004; Spitz and Furlong, 2012). Hence, a strongly coexpressed TF-gene pair might not necessarily mean a direct physical interaction, but can be observed as an indirect regulatory effect, even if they co-occur in a single functionally related cluster. Moreover, the affinity of a TF for a target gene can be highly tissue-specific or dependent on the metabolic state of the cell. Therefore, to deduce a regulatory network prioritizing TFs, the underlying expression data should have a unifying biological context (e.g. datasets for a specific tissue or condition) and coexpressed edges should be filtered for indirect interactions to minimize false positives. However, inferring accurate regulatory networks using solely gene expression data requires a large number of empirical data points for each space and time combination, for a robust statistical and biological inference. Nonetheless, for plant biologists, accumulated datasets in Arabidopsis are large enough to elucidate specificity of coexpression and predict key functional roles of TFs.

In recent years, several reverse engineering solutions have been brought forward that aim to model coexpression data in a way such that direct interactions involving known regulatory genes are given a priority (Basso et al., 2005; Faith et al., 2007; Huynh-Thu et al., 2010). These algorithms use a successive edge filtering step to recover potentially direct interactions between TFs and their targets. For example, the ARACNE algorithm assumes that in a triplet of connected nodes, the edge with lowest coexpression score is representative of an indirect interaction (Margolin et al., 2006). The GENIE method sets a feature selection problem for every

gene to find the best subset of regulators from all the remaining genes (Huynh-Thu et al., 2010). The CLR algorithm aims to identify direct transcriptional interactions by using a background correction scheme that suppresses noise arising due to high correlations between indirect interactions (Faith et al., 2007). These algorithms have all been successfully used for inferring plant gene regulatory networks (Yu et al., 2011; Chavez Montes et al., 2014; Vermeirssen et al., 2014).

In the work presented here, we focused on a comprehensive published gene expression dataset acquired from the seed development phases of Arabidopsis (Belmonte et al., 2013), and constructed a regulatory network highly predictive of seed-specific functions of TFs (Fig. 1). First, we harnessed the power of coexpression and graph clustering to partition genes into functionally related modules, and mapped the spatio-temporal activities of these modules. Simultaneously, for every identified TF in the Arabidopsis genome, we computed its partial coexpression score with every possible target gene and used these scores as a parameter for gene set enrichment analysis using coexpressed modules as gene sets. In this way, we could identify the modules that were statistically-most-likely targets of each TF. Using systematic reduction of data points and prior knowledge from the literature to interpret the associations, we observed that several TFs that are known to have an aberrant seed phenotype were predicted as the most significant regulators of modules for which their function has been experimentally validated. For example, a recently discovered association between the TF *AGL67* and desiccation tolerance (González-Morales et al., 2016), and *MYB107* and suberin (Lashbrooke et al., 2016) was correctly predicted in our network. These and several other correctly predicted associations (described later in the text) motivated us to create an online resource for the community. Our network, which we termed the 'Seed Active Network' or SANe, is hosted at https://plantstress-pereira.uark.edu/SANe/ to provide a network-based understanding of seed development.

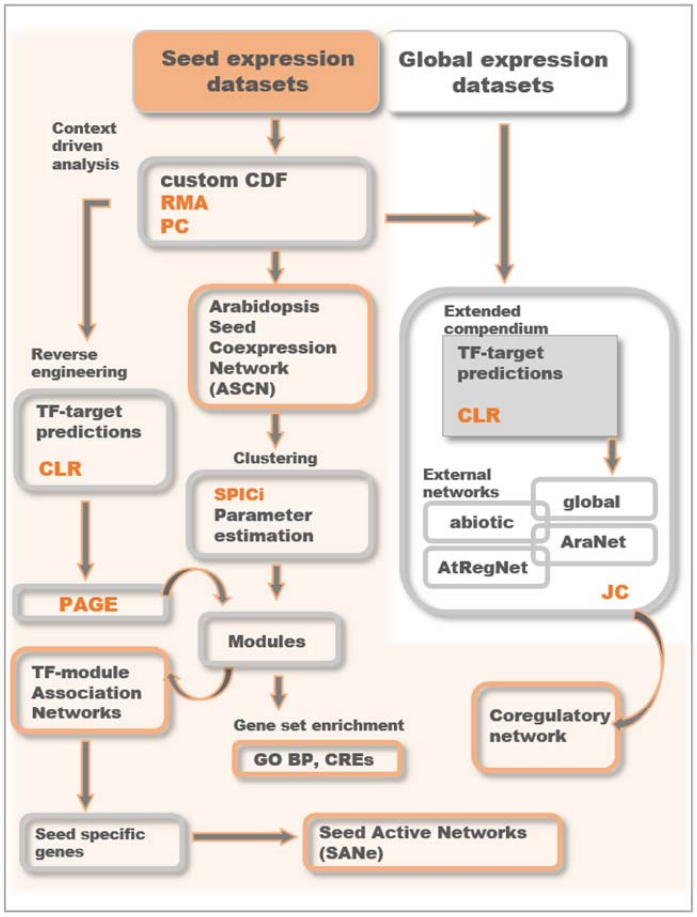

Figure 1: Pipeline for tissue-specific module regulatory network analysis. Two separate Arabidopsis gene expression compendiums (EC) were created: one from a seed-specific expression data series (GSE12404) and one from non-tissue-specific (global) 140 expression datasets. Datasets in both the EC were normalized individually using RMA algorithm in R. Z scores of Pearson's Correlations (PC) were calculated for all gene-pairs in both the EC. From seed EC, gene pairs with PC >0.73 (Z > 1.96) were connected to create the Arabidopsis Seed coexpression network (ASCN). ASCN was then clustered using SPICi at a range of clustering thresholds (T_d), and an optimal clustering parameter was chosen based on genome coverage and coherence of genes as a functional group. 1563 clusters obtained at T_d 0.80 were tested for enrichment of biological processes from the gene ontology and known plant *cis* regulatory elements for multiple databases. A list of 1921 TFs was supplied to the CLR (context likelihood of relatedness) algorithm to predict their targets in both the EC. In the seed EC, the PAGE algorithm was used to score the enrichment of CLR-weighted targets in the ASCN clusters, and a TF-module association network was created. The network was queried with a list of genes expressed predominantly in the seed as compared to other organs/tissues, and the Seed Active Network or SANe, was derived. Simultaneously, the seed-specific network was compared with the network created using the global EC and multiple other Arabidopsis regulatory networks downloaded from published studies.

Results

Seed coexpression network

To avoid implementing procedures of minimizing batch effects and the errors associated with microarray data integration (Chen et al., 2011; Nygaard et al., 2015), we chose Arabidopsis gene expression profiles from the data super series labeled GSE12404 in the gene expression omnibus (GEO) database (Barrett et al., 2007). This series is comprised of 87 samples derived from 6 discrete stages of seed development, and 5-6 different compartments within each stage, reflecting the most comprehensive source of Arabidopsis seed-specific gene expression profiles. With a sample size large enough for statistical inferences, these datasets were also devoid of the ambiguities introduced by the context under which the experiment was performed (intra-laboratory bias), one of the major problems in context-driven integrative analyses of gene expression data. We normalized and summarized this expression data into an integrated gene expression matrix using a custom CDF file of Arabidopsis microarray to reduce off-target hybridizations (Harb et al., 2010). Pearson's correlations (PC) scores between all gene-pairs in the gene expression matrix were then calculated and mapped to Z scores using Fisher's Z-transformation (Huttenhower et al., 2006). Gene pairs with significantly high correlation in expression (PC 0.753, Z-score >1.96) were connected and the rest filtered out. We named this core of raw coexpression data with ~7.6 million edges as the Arabidopsis seed coexpression network (ASCN).

Identification of clusters in coexpression data

Identification of communities, or clustering, is the most prominent step in network based interpretation of genomic data. In terms of gene expression data, clustering provides a useful way to group genes with similar expression profiles together. The need for gene grouping is based on the percept that expression similarity is indicative of similarity in function (Eisen et al., 1998). Therefore, clustering furthers an understanding of the function of a previously uncharacterized gene, based on known functions of other members of the same group. However, the choice of clustering method heavily influences the accuracy of functional predictions (Yeung et al., 2001). Clustering algorithms typically require either a predefined number of clusters, as in k-means clustering, or the process is semiautomatic (Langfelder and Horvath, 2008), and is sometimes computationally expensive.

In our network framework, we used an unbiased data-driven method to cluster genes within the ASCN. The density of a cluster, measured as the ratio of the number of observed

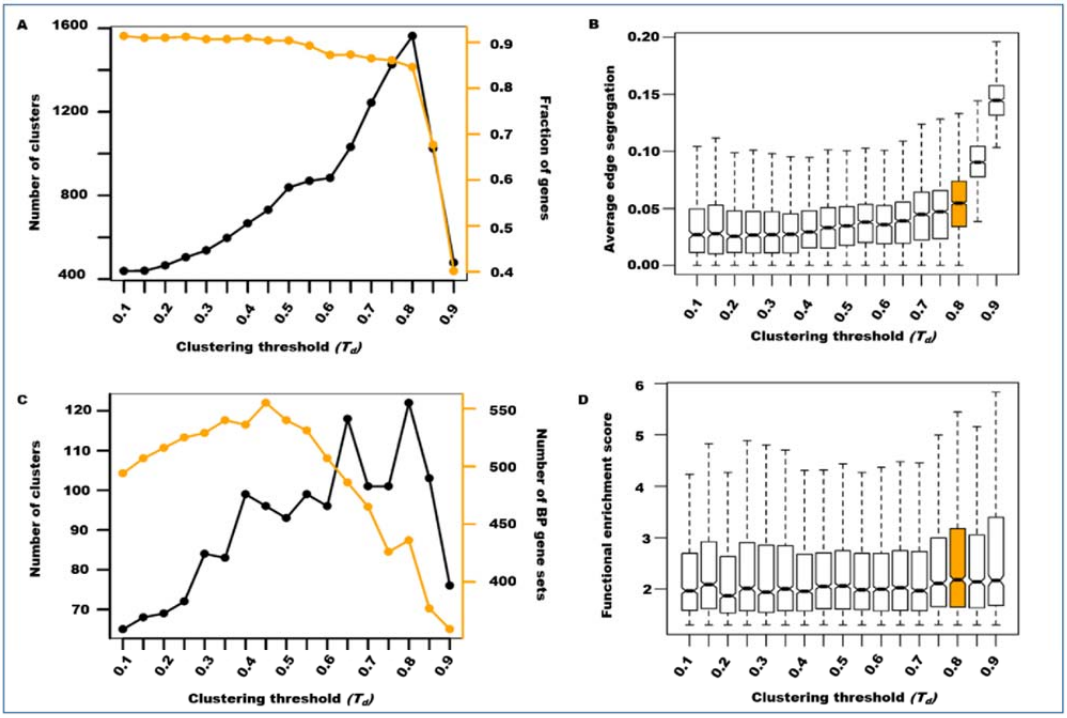

Figure 2: Evaluation of clustering threshold (T_d). Genes from the Arabidopsis seed coexpression network were clustered at a range of T_d values shown on the X axis of all the figures. Each T_d was examined by: A) A genome coverage plot measuring the number of clusters yielded and the fraction of original genes retained (orange line corresponding orange Y axis). B) Boxplots showing average edge segregation of all the clusters, indicating overall modularity of the network within each T_d. C) A plot showing number of clusters enriched with at least one BP term and the total number of BP terms retained (orange line corresponding orange Y axis) and D) Boxplots summarizing the enrichment scores [-1*log(FDR)] of the hypergeometric p-values obtained by BP-cluster overlap analysis.

171 edges in a cluster to the total number of expected edges, reflects cohesiveness among the
172 members of the same cluster. The SPICi algorithm evaluates density to group similar genes in a
173 biological network, while considering the confidence weight on each edge (Jiang and Singh,

2010). We sought to identify an optimum density threshold (T_d) that yields clusters at a granularity that delivers biological information, while preserving the inherent topological features of the network. A range of T_d values were evaluated for performance in loss or gain of information, with a goal of separating genes into as many clusters as possible, without losing many genes originally present on the microarray. At T_d 0.80, 84% of the ASCN genes formed 1563 clusters, after which a significant loss of information occurred, as indicated by a sharp fall in the fraction of total genes retained (Fig. 2A). At the same threshold of 0.80, the average modularity within clusters was also maximized (at a bearable cost of gene loss) (Fig 2B). Modularity measures how functionally separable the clusters are, in the sense that how well genes within a clusters interact with each other as compared to genes outside the cluster (Albert, 2005).

For a function-level analysis, it is also important that genes within each cluster are representative of common biological functions, as grouping genes would not yield any functional predictions if at least one putative function of the group is not known. To further establish confidence in T_d 0.80 as the best solution for partitioning, we evaluated each T_d for its ability to categorize known information about Arabidopsis biological pathways derived from the Gene Ontology (GO) annotated gene sets in the biological process (BP) category. A full set of annotation terms satisfying the parent-child relationships were used to find overlaps with clusters obtained at every T_d. The significance of overlap was tested under the hypergeometric distribution (see "Methods"). The functional coherence of the network, evaluated based on the total number of clusters with enriched BP terms, total number of distinct BP terms and the overall functional enrichment score, was also found to be best preserved at T_d 0.80 (Fig 2B and 2C).

Overall, the network lost its stability and collapsed at T_d values exceeding 0.80, as indicated by all measured clustering parameters (Fig. 2). Hence, 1563 dense clusters obtained at T_d 0.80 were used for further analysis. The total number of genes in these modules amounts to 17,949 (Supplemental Data S1).

Transcriptional regulators of seed modules

Modules of coexpressed genes in ASCN retained information about possible functional interactions between genes and their responses during different stages of seed development. This

greatly expanded upon the currently available functional annotations of Arabidopsis genes, as the genes that were lacking functional annotations now have at least one putative function assigned based on their module participation. The next task was to leverage on this information in the coexpression data and identify key TFs that statistically associate with each of the ASCN modules. There are 1921 unique locus IDs in the Plant Transcription Factor Database (Jin et al., 2014), the AGRIS database (Yilmaz et al., 2011) and the Database of Arabidopsis Transcription Factors (Guo et al., 2005), corresponding to TF genes in Arabidopsis. We used this comprehensive list to obtain transcriptional regulators for our analysis.

Simply associating genes as targets of TFs that they 'highly coexpress' with (first neighbors) is prone to the occurrence of false positives in a genome-scale analysis. This occurrence is mainly due to correlations arising from indirect regulation or coincidental coexpression of genes involved in different and unrelated processes that need to be active under the same circumstances. To minimize this effect, we calculated how likely a predicted TF-gene interaction was given the empirical background distribution of correlation scores of both the genes under consideration (Faith et al., 2007) (reported as a Z-score, see Methods) (Supplemental Data S2). Next, we sought to identify those modules that had higher enrichment of most probable targets for each TF. Instead of choosing an arbitrary cutoff for selecting targets, we used the entire set of predictions for each TF, weighted by Z-scores, and worked under the framework of Parametric Analysis of Gene set Enrichment (PAGE) (Kim and Volsky, 2005). The PAGE algorithm uses the normal distribution for statistical inference and states the degree of enrichment (here 'association') of a given gene set (here module) amongst the most highly scored predicted targets of a given TF. This analysis is essentially similar to that of a two-tail enrichment test with GO BP terms (treated as gene sets) (Ambavaram et al., 2011). Here, the difference was that gene sets from coexpression clusters observed in a specific tissue was used. To provide a normal distribution for association scoring, we used only those modules that had more than 10 genes, as suggested by the authors of the PAGE algorithm. Using this robust formulation, 1819 TFs were linked to 278 modules comprised of 10,526 genes (cluster 1 with 1621 genes was considered an outlier cluster because it contained disproportional number of genes as compared to other clusters). We labeled this network core as 'TF-Module Network' (TMN). TMN is represented as a matrix with TFs in rows and modules in columns, with each cell in the matrix representing a TF-module association score given by PAGE (Fig. 3A).

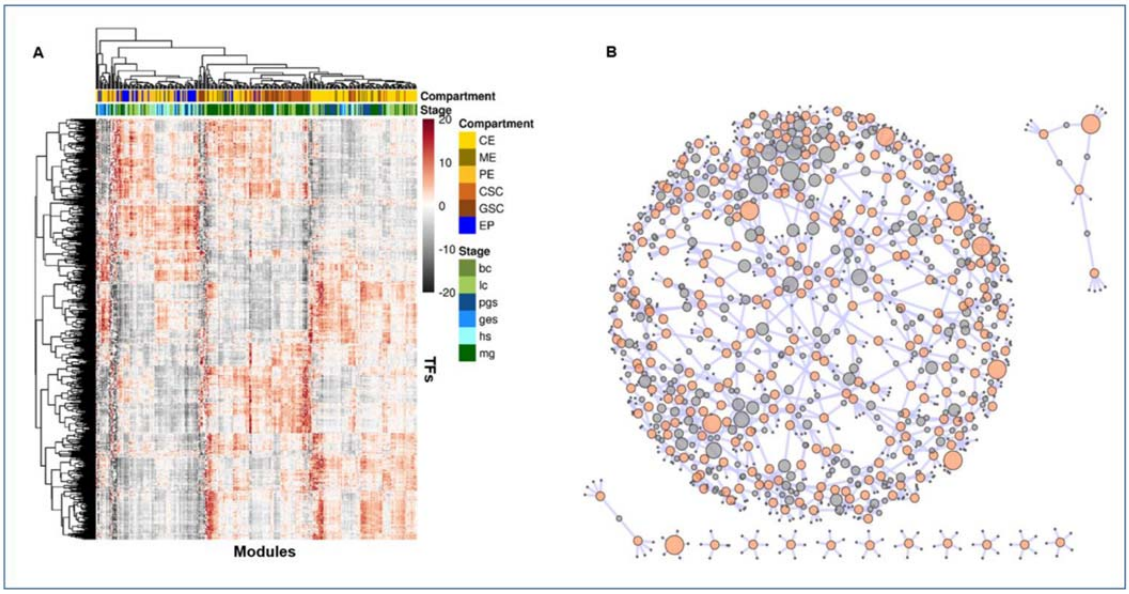

Figure 3: **TF-Module Association Network (TMN). A)** Heatmap representing association scores of 1819 TFs as regulators in the rows, and 10526 genes grouped into 278 coexpression modules represented along the columns. Each grid in the heatmap is color coded according to the level of enrichment of predicted targets of each TF regulator in the corresponding module. The red gradient indicates a positive score and grey indicates a negative score, estimated using the PAGE algorithm. The seed compartment in which the module has maximum expression is color-coded and represented on top of the heatmap (first row), where CE is chalazal endosperm, ME is micropylar endosperm, PE is peripheral endosperm, CSC and GSC is chalazal and general seed coat, respectively, and EP is embryo proper. The development stage in which the module has maximum expression is color-coded and represented on top of the heatmap (second row), where bc is bending cotyledon, lc is linear cotyledon, pgs is pre-globular stage, ges is globular embryo stage, hs is heart stage and mg is mature green stages. **B)** Predictions for each of the 278 modules were ranked and the top 5 predicted regulators for each module were visualized as a network graph. Each grey circle in the network plot is a TF and each orange circle is a module. The size of the grey circle is proportional to the out-going degree of the TF. Size of the orange circle was set to a constant, except for 9 bigger circles showing the modules described later in the main text. The network was visualized using Cytoscape version 3.3.0. Node names are hidden for ease in visualization. The cytoscape sessions file is provided as supplemental data S3, which can be loaded into Cytoscape for node names and further exploration of the network. The heatmap was drawn using gplots package in the R statistical computing environment.

The TMN provides a regulatory map of seed transcriptional activities, in the form of a bipartite graph, with TFs as one set of nodes and sets of genes reduced to their 'functions' as another set of nodes, and edges weighted by the degree of association between the corresponding

TF and the function. For visualization, we selected the top 5 predicted TF regulators for each module, ranked based on absolute association scores, and visualized TMN as a graph in Cytoscape (Fig. 3B; Supplemental Data S3). A total of 900 regulators were represented in top 5 predictions for each of the 278 modules. Most the modules were found indirectly connected due to combinatorial links between their predicted TF regulators, forming a dense network while 11 modules shared no common predicted TF regulators with other modules.

Modules active during seed development

Seed-specific genes were previously discovered as those that were present only in seed tissues, and not in other reproductive or vegetative parts of the plant (Le et al., 2010; Belmonte et al., 2013). We sought for those modules that harbored at least one such gene and identified a core set of 120 modules comprised of 7414 genes. We called these modules as 'active modules'. We reasoned that because these modules retained genes specific to seed development, their coexpression neighborhood – along with the top ranked regulators – will pave way to identification of transcriptional networks modulated specifically during seed development, or involved in important seed functions. Therefore, novel TFs that are already part of these modules, or emerge as the top regulators will automatically become the primary candidates for testing seed phenotypes, largely reducing the search space. Also, the strategy of probing TMN with a list of genes already prioritized had less chances of observing false positives from a gamut of predicted regulatory programs, while making the process of interpreting the regulation patterns easier. We labelled this core of 120 active modules along with their scored TF regulators as the 'Seed Active Network' (SANe) (Supplemental Data S4).

We simultaneously mapped the expression patterns of each module spatially and temporally (seed compartment wise and development stage wise), by averaging the expression of module genes in each seed-compartment irrespective of the development stage or within each development stage irrespective of the seed compartment. After interfacing the expression patterns of each module with BPs and known *cis* regulatory elements (CREs) (Supplemental Data S5 and S6; see "Methods") and predicted sets of top regulators, a few modules that had high expression in different seed compartments (embryo, endosperm and seed-coat regions) were visually examined using heatmaps (Fig. 4). These modules expand a wide variety of cellular processes, including flavonoid metabolism during seed coat formation, lipid storage and

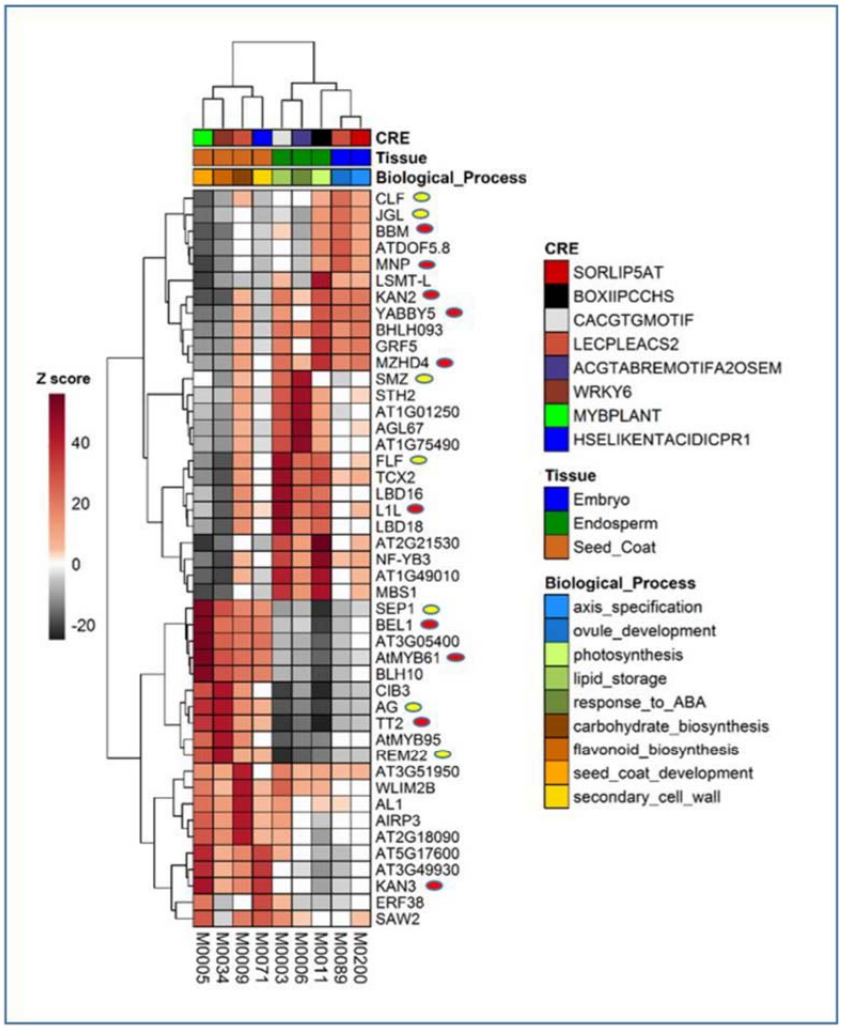

Figure 4: Heatmap representation of a subset of the SANe. Modules with high relative expression in embryo, endosperm and seed coat regions were extracted from SANe. Modules are shown in columns and for each module, the top 5 predicted TF regulators are shown in rows. Each grid in the heatmap is colored according to the association score estimated using the PAGE algorithm. Positive and negative scores are shaded in red or black gradient, respectively, as indicated by the color key. Literature identified TFs with validated seed-specific phenotypes or phenotypes observed in other reproductive stages/tissues are marked with a red ellipse or a yellow ellipse, respectively. CRE, cell-type and functional annotations for each module are shown above the heatmap (top three rows; colored boxes). Modules annotated for embryo, endosperm and seed coat are indicated in blue, green and brown boxes, respectively, in the middle row. CRE and functional annotation for each module is color- coded uniquely in the top and bottom rows, respectively.

268 photosynthesis during endosperm development and auxin transport and tissue development from
269 early to late stages of embryogenesis. Visualization of a few modules revealed that there is a
270 high intra-module connectivity between modules that participate in the same developmental

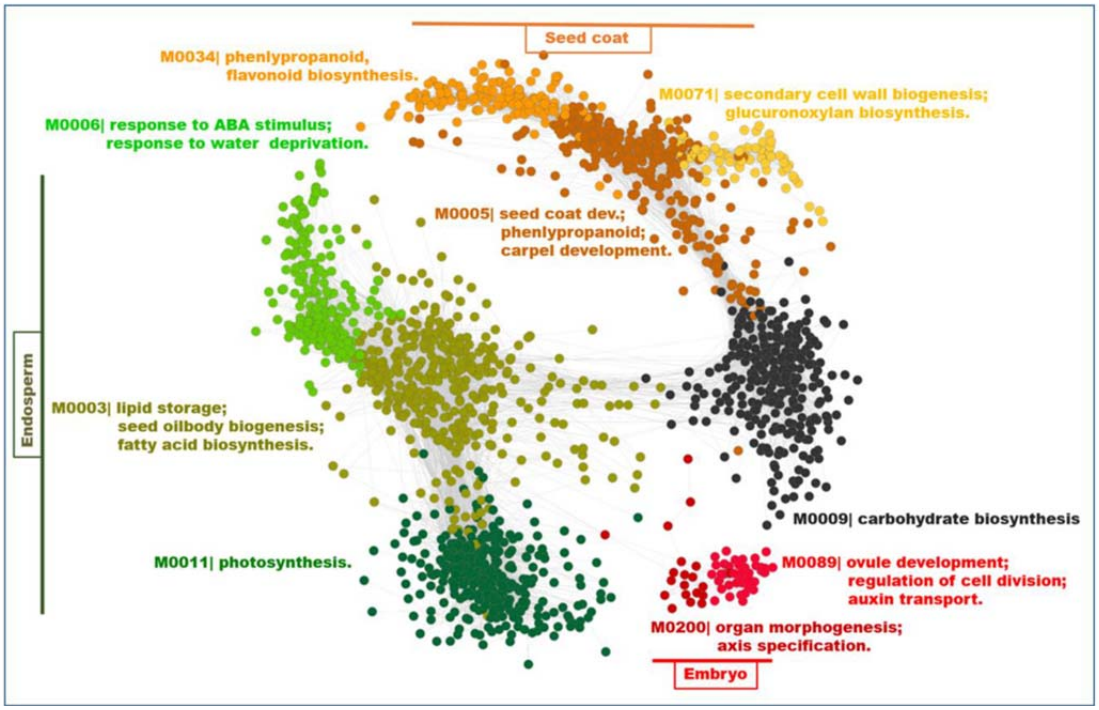

Figure 5: Visualization of seed modules. A graphical representation of seed coexpression modules. Each circle represents a gene. Each module is color coded uniquely. Modules are grouped according to the seed compartments (indicated by horizontal or vertical lines and text boxes), and labelled with the BP term most highly over-represented within each module. Genes are left unlabeled to facilitate visualization. The network was drawn in Cytoscape version 3.3.0. .

271 program in a tissue-specific manner, albeit with different biological goals (Fig. 5). A few such
272 modules are described below.

273 **Modules for early embryo development**

Three modules designated as M0089, M0200 and M0277 comprised 54, 31 and 33 genes, respectively, expressed at relatively high levels in the embryonic tissue when compared to other seed compartments (Fig. 6A). These genes are significantly enriched with BP terms like "organ development", "tissue development", "axis specification" and "auxin transport". This is consistent with processes related to embryo development, involving morphogenesis-related and other cellular processes that govern gene activity related to cell division and expansion, maintenance of meristems and cell fate determination (Wendrich and Weijers, 2013).

M0089 harbors genes related to reproductive tissue development and cell division. ATDOF5.8 (AT5G66940) was predicted as the top regulator of M0089. The *ATDOF5.8* gene is most highly expressed in embryo and meristem cells (Supplemental Fig. S1A) based on the Genevisible tool in GENEVESTIGATOR (Zimmermann et al., 2004). It has been shown that ATFOD5.8 is an abiotic stress-related TF that acts upstream of ANAC069/NTM2 (AT4G01550) (He et al., 2015). Interestingly, the *NTM2* gene resides at a locus adjacent to another NAC domain TF, NTM1 (AT4G01540), a regulator of cell division in vegetative tissues (Kim et al., 2006). Kim et al. did not detect *NTM2* expression in leaves by RT-PCR. However, they indicated that because both *NTM* genes have similar structural organization, encoding proteins with a few differences in the protein chain, NTM2 could be involved in similar processes in other tissues. Our predictions suggest that *NTM2* could be in the ATDOF5.8 regulon associated with modulating cell division activity in the seed. This leads to a new testable hypothesis pertaining to regulation of cell division during embryogenesis. Among other known regulators, BABY BOOM (BBM, AT5G17430) was predicted as one of the top ranked TF (rank 4) of M0089. BBM is an AP2 TF that regulates the embryonic phase of development (Boutilier et al., 2002).

296 YAB5 (AT2G26580) and ATMYB62 (AT1G68320) were predicted the top ranked
297 regulators of M0200 and M0277, respectively. While the agreement of YAB5 as a determinant
298 of abaxial leaf polarity (Husbands et al., 2015) and enrichment of M0200 with GO BP term “axis

299 specification" (GO:0009798) justifies this association, the association of ATMYB62 with
300 M0277 indicates a hormonal interaction likely representing a transition between the growth
301 stages. *ATMYB62* encodes a regulator of gibberellic acid biosynthesis (Devaiah et al., 2009) and

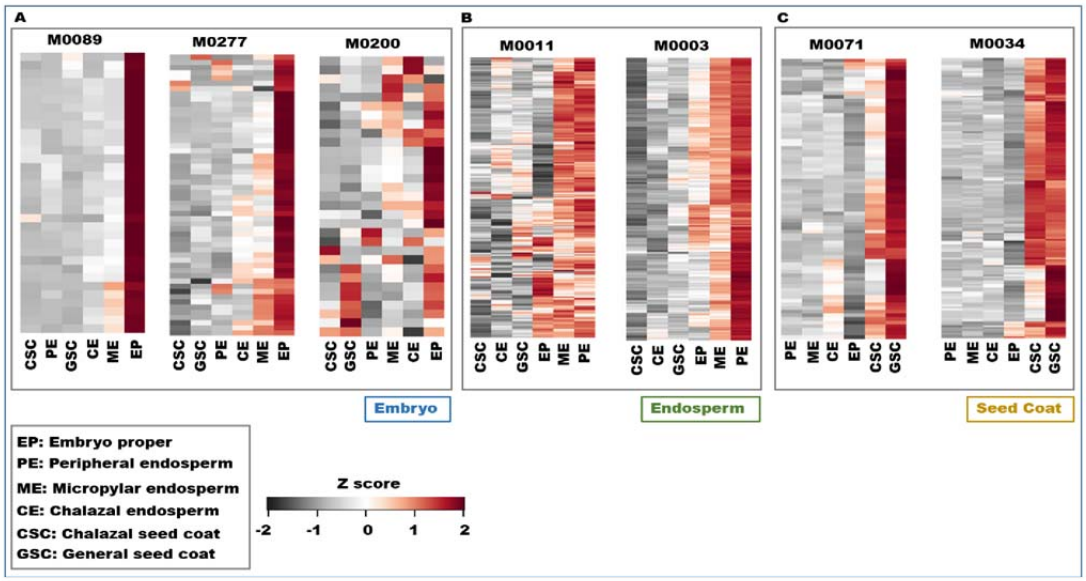

Figure 6: Expression profiling of gene modules. Expression patterns of modules in embryo, endosperm and seed coat regions represented as heatmaps in A), B), and C), respectively. Seed compartments are represented as columns and genes as rows. Gene names are hidden for ease in visualization. Expression values of genes in each module were averaged across samples from the same tissue-type/seed-compartment (embryo, endosperm and seed coat). Average expression values were scaled and represented as a Z score in the heatmaps. Red indicates higher expression of a gene in a particular compartment and black gradient indicates lower expression relative to other compartments.

302 is expressed specifically during seed development (Belmonte et al., 2013). M0277 is enriched
303 with "auxin transport" genes (GO:0009926). The *ATMYB62* gene is preferentially expressed in
304 the abscission zone and other reproductive tissues (Supplemental Fig. S1B).

Modules for Endosperm Development

The endosperm has a profound influence on seed development by supplying nutrients to the growing embryo (Portereiko et al., 2006; Chen et al., 2015). The importance of endosperm cellularization for embryo vitality has been shown through mutants deficient in endosperm-specific fertilization events (Kohler et al., 2003). The overall seed size depends on endosperm development and is controlled through the relative dosage of accumulated paternal and maternal alleles (Luo et al., 2005).

We found that genes in modules M0003 and M0011 had maximal expression levels in endosperm tissues (Fig. 6B). M0003 is significantly enriched with genes involved in lipid storage (GO:0019915) and fatty acid biosynthesis (GO:0006633). LEC1-LIKE (L1L, AT5G47670) emerged as the top regulator of this module. LIL is related to LEAFY COTYLEDON 1 (LEC1) and functions during early seed filling as a positive regulator of seed storage compound accumulation (Kwong et al., 2003). Interestingly, L1L is also part of this module indicating that, apart from being a master regulator, its activity is also modulated during the late seed filling stages as observed previously (Kwong et al., 2003), which correlates with the overall expression pattern of genes within this module (Supplemental Fig. S2). The presence of 44 other TFs in this module, including FUS3 and ABI3, key regulators of seed maturation (Keith et al., 1994; Luerßen et al., 1998; Yamamoto et al., 2009), points to the importance of this module in nutrient supply to the developing embryo. LDB18 (AT2G45420) is a LOB-domain containing protein of unknown function predicted as the second ranked regulator of this module. GENEVESTIGATOR analysis showed that both *L1L* and *LDB18* are most highly expressed in the micropylar endosperm (Supplemental Fig. S3).

M0011 is comprised of 357 genes, including 7 TFs and is characterized by containing genes with high expression levels in the micropylar endosperm (ME) and the peripheral endosperm (PE). GO enrichment analysis showed the highest scores for photosynthesis (GO:0015979) of genes in this module. Close examination of these genes revealed that virtually all aspects associated with chloroplast formation and function were represented, including chloroplast biogenesis and membrane component synthesis, chlorophyll biosynthesis, plastidic gene expression, photosynthetic light harvesting and electron transport chain, ATP production,

redox regulation and oxidative stress responses, Calvin cycle and photosynthetic metabolism, metabolite transport, and retrograde signaling. Interestingly, genes encoding photorespiratory enzymes (glycine decarboxylase, glyoxylate reductase, and hydroxypyruvate reductase) were also present in M0011. Developing oilseeds are known to keep extremely high levels of CO_2 that would suppress photorespiration (Goffman et al., 2004), and the implications of expression of these genes on photosynthetic metabolism are not clear.

The presence of mostly photosynthetic genes in M0011 seems also unusual, but the results are consistent with findings of (Belmonte et al., 2013), showing that specific types of endosperm cells are photosynthetic, as they contain differentiated chloroplasts and express photosynthesis-related genes. Fully differentiated embryos at the seed-filling stages and the chlorophyll-containing inner integument ii2 of the seed coat are parts of oilseeds that are also capable of photosynthesis (Belmonte et al., 2013; Sreenivasulu and Wobus, 2013). Although seeds obtain the majority of nutrients maternally, Arabidopsis embryos remain green during seed filling and maintain a functional photosynthesis apparatus similar to that in leaves (Allorent et al., 2015). As part of photoheterotrophic metabolism, photosynthesis provides at least 50% of reductant in oilseed embryos and CO_2 is re-fixed through the RuBisCo bypass that helps to increase carbon-use efficiency in developing oilseeds (Ruuska et al., 2004; Schwender et al., 2004; Goffman et al., 2005; Fait et al., 2006). The roles for photosynthesis in ME and PE remain to be investigated and include (i) providing carbon and energy for storage compound accumulation in the endosperm and the embryo and (ii) increasing the availability of oxygen to the endosperm and differentiating, yet-to-be photosynthetic, embryos in a high-CO_2 environment.

CRE analysis revealed the highest number of motifs enriched in the promoters of genes in M0011, suggesting extensive coordination between different regulators. Light-related motifs BOXIIPCCHS (ACGTGGC), IRO2OS (CACGTGG3), IBOXCORENT (GATAAGR) and the ABA-responsive element ACGTABREMOTIFA2OSEM are the most over-represented motifs in this module. The highest ranked regulator of M0011 is a SMAD/FHA domain-containing protein (AT2G21530) that is most highly expressed in the cotyledons (Supplemental Fig. S4A). The known seed-specific regulator of oil synthesis and accumulation WRI1 (AT3G54320) was identified as the sixth ranked regulator of this module and is suggested to be predominantly

365 expressed in the embryo and endosperm (Supplemental Fig. S4B). *WRI1* encodes an AP2/ERF-
366 binding protein and *wri1* seeds have about 80% reduction in oil content relative to the wild type
367 seeds (Ruuska et al., 2002). Genetic and molecular analysis revealed that WRI1 functions
368 downstream of LEC1 (Baud et al., 2007). Along with WRI1 itself, six other TFs are part of this
369 module, including AT2G21530, a zinc finger (C2H2) protein (AT3G02970), NF-YB3
370 (AT4G14540), PLT3 (AT5G10510), GIF1 (AT5G28640) and PLT7 (AT5G65510).

371 **Modules for Seed Coat development**

372 The seed coat has important functions in protecting the embryo from pathogen attack and
373 mechanical stress. The seed coat encases the dormant seed until germination and maintains the
374 dehydrated state by being impermeable to water. M0034 is comprised of 149 genes with the
375 highest expression in general, and specifically in chalazal seed coat relative to other tissues (Fig.
376 6C). This module is enriched with genes annotated under the GO BP terms "phenylpropanoid
377 biosynthetic process" (GO:0009699) and "flavonoid biosynthesis process" (GO: 0009813). The
378 AP2/B3-like TF AT3G46770 is highly expressed in seed coat (Supplemental Fig. S5A) and
379 predicted as the top regulator in this module. B3 domain TFs are well known for functioning
380 during seed development and transition into dormancy in *Arabidopsis* (Suzuki and McCarty,
381 2008) and, to some extent, their functions are conserved in cereals (Grimault et al., 2015). The
382 seed-coat-specific expression of AT3G46770 is a compelling incentive for testing AT3G46770
383 mutants for seed-related phenotypes, which to the best of our knowledge, has never been
384 considered. There were 21 other TFs belonging to this module, of which six are part of the MYB
385 family. TRANSPARENT TESTA 2 (TT2), a MYB family regulator of flavonoid synthesis (Nesi
386 et al., 2001), was ranked fourth in our predictions for this module.

387 M0071 is composed of 77 genes encoding, surprisingly, only 3 TFs, ERF38
388 (AT2G35700), BEL1-LIKE HOMEODOMAIN 1 (BLH1, AT2G35940) and a C2H2 super
389 family protein (AT3G49930). This module is enriched with genes involved in "xylan metabolic
390 process" (GO:0045491), "cell wall biogenesis" (GO:0009834), and "carbohydrate biosynthetic
391 process" (GO:0016051). KANADI3/KAN3 (AT4G17695) was predicted as the top regulator of
392 this module. KANADI group of functionally redundant TFs (KAN1, 2, and 3) has been shown to
393 play roles in modulating auxin signaling during embryogenesis and organ polarity (Eshed et al.,
394 2004; McAbee et al., 2006; Izhaki and Bowman, 2007). In the case of another KANADI TF,

KAN4, encoded by the *ABERRANT TESTA SHAPE* gene, the lack of the KAN4 protein resulted in congenital integument fusion (McAbee et al., 2006). It is reasonable to hypothesize that KAN3 could be acting in a redundant manner with KAN4 to regulate seed coat formation during late stages of maturation, as the expression pattern of *KAN3* is higher in seed coat than in other organs or cell types (Supplemental Fig. S5B).

Module M0006 is related to seed desiccation tolerance

M0006 is comprised of 220 genes expressed predominantly during the mature green stage (Fig. 7A), and enriched with genes involved in "response to abscisic acid stimulus" (GO:0009737), "response to water" (GO:0009415) and terms related to embryonic development (GO:0009793), altogether suggesting an involvement of these genes in acquisition of desiccation tolerance (DT). We predicted AGL67 (AT1G77950) as a major regulator of this module, among 23 other TFs that are part of this module (Fig. 7B). AGL67 has been recently confirmed as a major TF involved in acquisition of DT (González-Morales et al., 2016), validating our prediction. Additionally, the authors of this study analyzed the mutants of 16 genes (TFs and non-TFs) that had reduced germination percentage, of which 12 are in our network and 7 of these are a part of M0006. These 7 genes include PIRL8 (AT4G26050), ERF23 (AT1G01250), OBAP1A (AT1G05510), DREB2D (AT1G75490), AT1G77950 (AGL67), AT2G19320 and MSRB6 (AT4G04840).

Characteristics of seed-specific networks

The primary objective of this network analysis pipeline was to capture gene regulation information in a tissue-specific manner. To examine the effect of this approach and to identify the distinguishing characteristics of the seed regulatory network that differed from a global network (non-tissue specific regulatory network), we extended the seed expression compendium to incorporate an additional set of 140 datasets related to profiling gene expression from various organs of the Arabidopsis plant, including vegetative and seedling growth stages. Using the same reverse engineering approach as described above, we scored each TF-target pair on this extended expression compendium (EEC). Next, to delineate the distinguishing properties of seed networks, we compared the level of co-regulation induced by TFs, measured as similarity in the

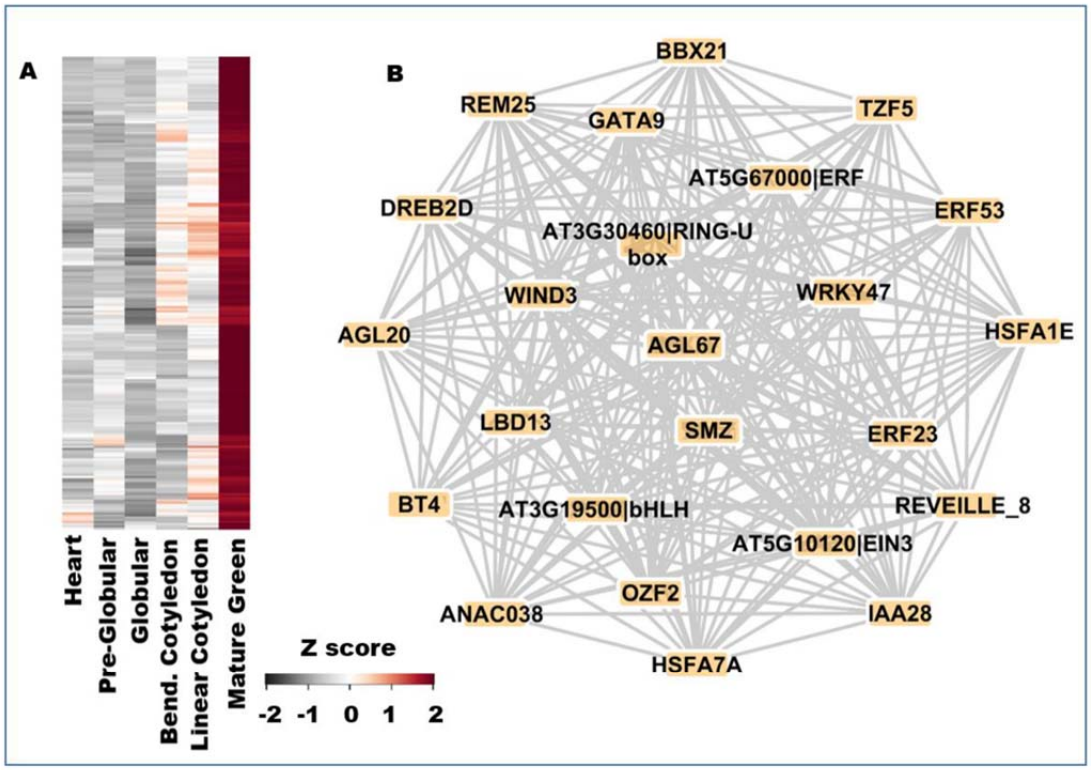

Figure 7: Module M0006. A) Expression patterns of genes in module M0006. Seed development stages are represented as columns and genes as rows. Gene names are hidden for ease in visualization. B) Coexpression links between TFs in M0006. Nodes are labeled according to their corresponding gene symbols if present in TAIR, else labeled with their corresponding locus ID and the family the protein belongs to.

424 predicted targets of each TF-pair, using Jaccard's coefficient (JC), in both the seed-specific
425 network and the global regulatory network created using EEC. As expected, a larger number of
426 TFs have very few common targets, and this number is high for fewer TFs in both the networks

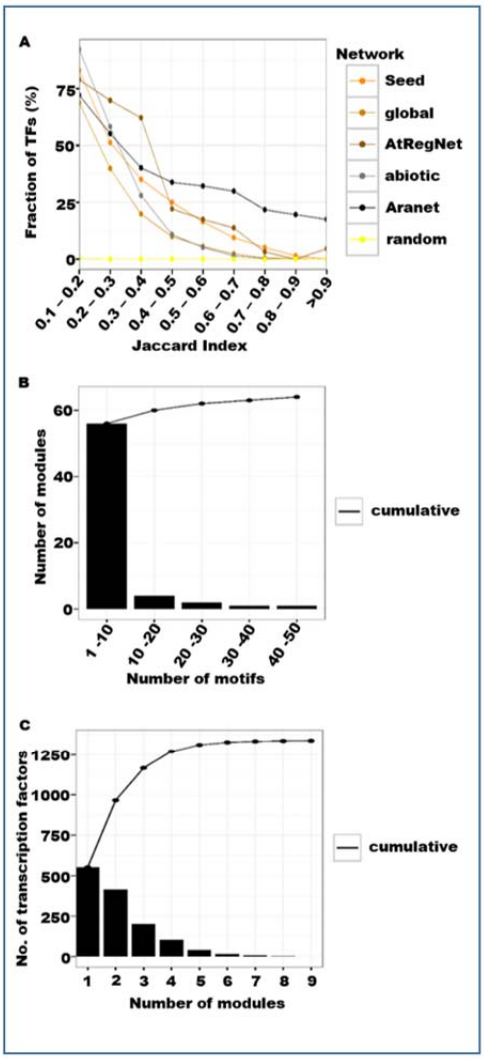

Figure 8: Characteristics of seed networks. A) Comparison of the fraction of TFs possibly coregulating the same sets of genes, evaluated using the Jaccard's Index (JI) of overlap between the predicted targets of each TF-pair, for 5 different regulatory networks and a random network. **B)** A histogram showing bins of number of motifs significantly over-represented in the promoters of genes within each module in TF-MAN. **C)** Distribution of TF-module edges in SANe follows a scale-free topology, with a large number of regulators associated with fewer modules, and a few regulators (e.g., NAP57, KAN3) associated with a large number of modules.

427 (Fig. 8A). A larger number of TFs have similar targets in the seed network at any given JC bin,
428 as compared to the global network.

Although false positives and false negatives are part of any network based predictions, we suspected that the trends observed in comparison of the seed-specific and the global network could be trivial if there were correlated errors arising from the same network prediction pipeline for both networks. To overcome this uncertainty, we downloaded and analyzed the recently published Arabidopsis oxidative stress gene regulatory network predicted from a compendium of microarrays conditioned on abiotic stress (Vermeirssen et al., 2014). This abiotic-stress specific network is essentially a consensus network of an ensemble of reverse engineering algorithms, and performed remarkably well in validations (Vermeirssen et al., 2014). We then computed the overlaps in the predicted targets of TFs in this network (as done for networks in this study) and observed that it follows a trend very similar to that of the global network (Fig. 8A), indicating that there was no major bias introduced by our approach.

To extend the comparisons, we performed the same operation to the *Arabidopsis thaliana* Regulatory Network (AtRegNet) and the AraNet (Lee et al., 2010). AtRegNet harbors about 17,000 direct edges validated for TFs and their target genes. AraNet is a co-functional network derived by integrating 24 -omics datasets from multiple organisms in a machine-learning framework. Both networks showed a similar gradual decrease in fraction of TFs with similar targets with higher JC values (Fig. 8A), similar to trends observed in networks with a 'functional context' above. However, we used these networks for comparison only as a rough guide as the AraNet was not designed to prioritize regulatory interactions and holds only approximately 60,000 such edges, and the AtRegNet harbors very few TFs when compared to those in our list. We assumed that both these limitations would make the analysis suffer from the extreme loss of transcriptional signal. However, the robustness of gene relationships predicted in the AraNet was clearly evident as more than 20% of the original TFs in the network presumably interacted even in the highest JC bin, larger than any other networks compared. Overall, the number of TFs observed at any given JC bin in all networks was significantly larger than in a random network. All TF-pairs with JC > 0.70 (arbitrarily chosen stringency) from the seed-specific network were connected and visualized as a graph in Cytoscape (Shannon et al., 2003) revealing many connections supported by multiple networks (Supplemental Fig. S6)

About 59% of all genes (23% of all modules) in TMN have at least one known plant CRE enriched in their coexpression neighborhood, with a few modules harboring a large number of

different CREs (e.g., Photosynthesis module described earlier) (Fig. 8B). Approximately 45% of total edges in ASCN have an absolute PC score more than 0.9, indicating a highly cohesive network structured for a subtle developmental program.

For evaluation of 'hubs', we selected top 10 TF predictions for each active module in SANe (based on ranked association scores), and counted the number of modules associated with each TF. We observed that 41% of these TFs (552 out of 1339), likely regulate expression of genes in only one module each, while a single TF, NAP57 (AT3G57150), was predicted to be associated with the maximum number of modules (9 out of 120) (Fig. 8C). The *NAP57* gene encodes the Arabidopsis dyskerin homolog involved in maintaining telomerase activity (Kannan et al., 2008). As expected, 5 out of 9 modules containing genes whose expression is predicted to be regulated by NAP57 are enriched in GO BP terms such as "DNA metabolic process", "ribonucleoprotein complex biogenesis", "RNA processing" and "ribosome biogenesis". This association was true even on the level of individual targets predictions for majority of the other seed-hubs, in both, the seed and global networks (Table 1), indicating that these TFs are responsible for perpetual regulation of important basic processes like biogenesis of cell components, maintenance of cell shape and structure, nucleic acid metabolism etc. A weak but significant enrichment was found between WRKY13 (AT4G39410), a biotic and abiotic stress regulator (Qiu et al., 2007; Xiao et al., 2013), and the GO term 'immune system response' only in the seed network.

The SANe webserver

The data generated in this study are represented on a web-based interactive platform available at https://plantstress-pereira.uark.edu/SANe/. The platform allows users to investigate seed development in three different modes (Fig. 9):1) Select modules with high expression in compartment – or stage-specific manner, 2) Using the 'cluster enrichment tool' to upload a differential expression profile (e.g. transcriptome of a TF mutant) and identify clusters that significantly perturb in their experiment and 3) enter the locus ID of a TF of interest to identify clusters that are likely regulated by that TF, enabling the user to gain a insight on its functional role prior to an *in vivo* validation. Furthermore, the webserver allows users to visualize the expression of resulting modules/clusters as publication-ready downloadable heatmaps, as well as plot gene connection graphs using Cytoscape (Lopes et al., 2010).

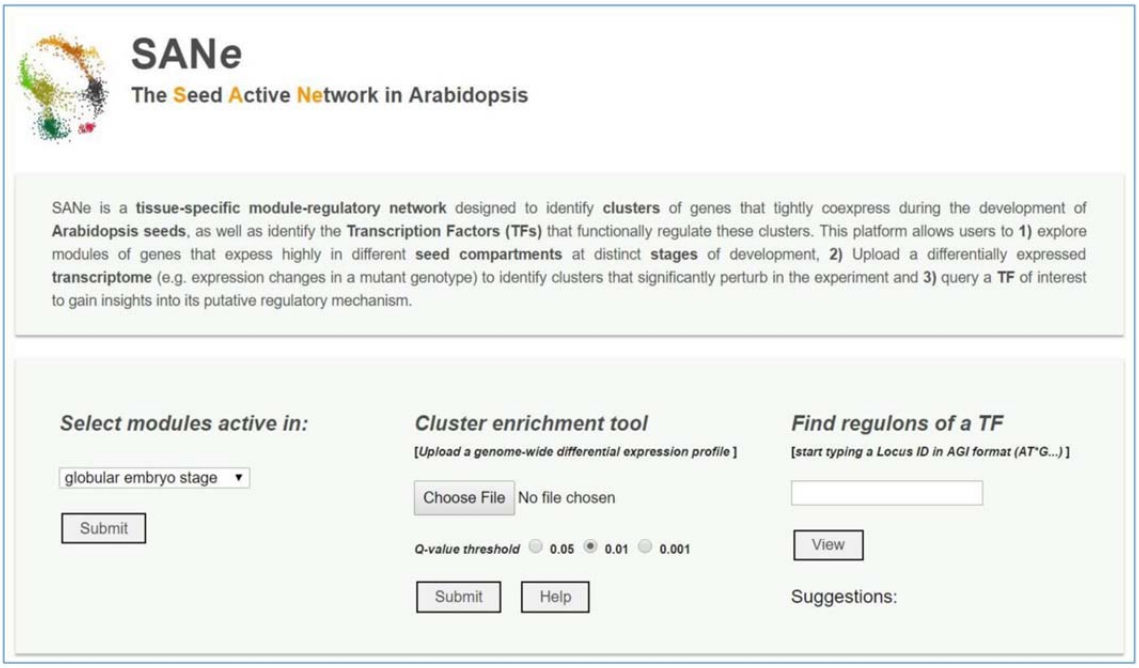

Figure 9: Screenshot of the SANe webserver.

The SANe web platform (https://plantstress-pereira.uark.edu/SANe/) allows users to identify modules active in distinct seed compartments in different stages of development, upload a new transcriptome in the cluster enrichment tool that uses the parametric analysis to identify enriched modules, and find the regulons of a TF of interest.

489

490

Discussion

Plant seeds are complex structures and seed formation is perhaps the most important developmental phase of a plant life cycle, as it determines the fate of the next progeny. Distinct cell types and organs within a seed gradually develop during a period of 20-21 days after pollination in Arabidopsis. In addition, each organ is subjected to its own developmental program and has different, but equally important functions, from feeding and providing optimal growth conditions to protecting the embryo to ensure species propagation. These processes are tightly regulated by synergistically acting TFs (To et al., 2006).

We devised a new methodology that relies on existing statistical methods that are widely accepted, for the discovery of a modular regulatory network. Using a seed-tissue specific expression dataset, this method facilitated identification of modules of co-regulated genes, the corresponding development phases in which the modules express most, CREs that drive the biological functions encoded by the genes within modules, and TF regulators that likely govern the expression of the genes in the modules. Our method is limited to making functional predictions for TFs in a tissue-specific manner, and might not accurately predict individual targets of a given TF. This limitation is partly due to the use of a single data-type; a heterogenous approach should be undertaken (e.g. high-throughput DNA binding essays in conjunction with expression data) for studies aiming at specific individual targets. Nevertheless, the statistically significant functional associations predicted here are of superior quality, as seen in evidence from the literature, and can serve as the first step in selecting TFs for targeted downstream experiments. The network inference pipeline presented here can be used to enhance any coexpression based study.

Previous studies have reported a few seed-specific genes, including TFs (Le et al., 2010; Belmonte et al., 2013). We prioritized these genes in our network to derive an active subnetwork, referred to as Seed Active Network (SANe). We described selected modules containing genes with high expression in specific seed components, including embryo, endosperm and seed coat. We observed that, in most of the cases, the top predicted regulators of these modules are already known in the literature for their involvement in seed development, self-validating our approach. Several additional regulators are known to modulate other processes, including flower development, indicating conserved regulons of pre-fertilization events. Our results suggest that

associating regulators to gene sets with a shared function, as opposed to individual genes, provides biologically plausible predictions that are worth for validating *in planta* phenotypes using reverse genetics. As a community resource, our network is accessible through an online platform supported with query driven tools to enable a network based discovery of seed regulatory mechanisms.

It appears that during seed development, photosynthesis and storage compound synthesis is tightly coordinated by several regulators acting coordinately. This was evident from CRE enrichment analysis, as two complementary methods detected the module annotated for photosynthesis and related processes (M0011) harboring genes with the largest number of known plant motifs in their promoters when compared to the rest of the modules. Coordinate regulation of photosynthetic carbon metabolism has been shown previously (Bailey et al., 2007; Ambavaram et al., 2014). Our analysis reveals that much of the processes related to embryo development are conserved throughout the plant life cycle such as cell division and differentiation, as observed by similar roles of regulatory genes in developing embryos and roots. However, plants have developed intrinsic mechanisms that can modulate gene activity in specialized cells, perhaps as duplicated genes with similar functional roles. Such a phenomenon was evident in the case of two TF genes, *NTM1* and *NTM2* that are in close proximity to each other and possibly have similar biological roles in distinct parts of a plant.

The data generated by our work has the potential to further our knowledge of fundamental processes that regulate diverse specific aspects of seed development in Arabidopsis and can be extrapolated to related agriculturally important crops due to conservation of these basic processes (Magallón and Sanderson, 2002; Comparot-Moss and Denyer, 2009; Vriet et al., 2010). Based on our results, a cell- and developmental stage-specific network inference provides superior quality of predictions in the context of known information. Our network analysis pipeline can be further used to systematically increase this information-base for a variety of plant organs (e.g., parts from a post-germination stage network). Comparisons of different stage/tissue specific networks will throw light on the changing molecular mechanisms of a cell and reveal differentially modulated transcriptional networks during different growth stages.

Materials and Methods

Gene expression quantification

Affymetrix ATH1 Arabidopsis gene expression data was downloaded from GEO, and 6 datasets were selected from the super series labeled GSE12404 for seed expression compendium. In addition, 140 other datasets were used in the EEC (Supplemental Data S7). All datasets were individually processed in R Bioconductor using a custom CDF file for Arabidopsis (Harb et al., 2010). The re-annotated CDF assigns probe-sets to specific genes and increases the accuracy in expression quantification**.** Using Robust Multi-array average algorithm (RMA) (Irizarry et al., 2003), probe level expression values were background corrected, normalized and summarized into gene level expression values. Values from replicate arrays were then averaged and assembled in an integrated expression matrix of genes as rows and samples as columns, with each cell in the matrix representing log transformed expression value of genes in the corresponding samples. This procedure resulted in two expression matrices: a seed-specific expression matrix and a global expression matrix.

Coexpression network and cluster identification

Pearson's Correlation (PC) were calculated for each gene pair using expression values in both gene expression matrices. PCs were Fisher Z transformed and standardized to a N(0,1) distribution, where a Z-score of a gene-pair represents the number of standard deviations the score lies away from the mean (Huttenhower et al., 2006). The following procedure was applied only to the seed network. Gene pairs with Z scores above 1.96 (PC 0.75) were retained and connected to create a coexpression network with 21,267 genes connected with approximately 7.6 million edges. SPICi, a fast clustering algorithm (Jiang and Singh, 2010), was used to cluster the network at a range of T_d values ranging from 0.1 to 0.90, keeping a minimum cluster size of 3. Each T_d value was evaluated on three criteria: i) total number of clusters yielded and the fraction of original genes retained in those clusters ii) average modularity following the (Newman and Girvan, 2004) algorithm and iii) functional coherence of clusters based on GO BP term annotations. At T_d 0.80, expression values of each gene within each of 1563 clusters were averaged across the same parts of the seed and in different developmental stages, resulting in two

578 expression profiles for each module. Expression values were scaled and plotted as heatmaps in R
579 using the gplots package (https://CRAN.R-project.org/package=gplots).

580 **Functional annotations of coexpression clusters**

581 The TAIR gene association file was downloaded from the plant GSEA website
582 (http://structuralbiology.cau.edu.cn/PlantGSEA/download.php) (Yi et al., 2013). The .gmt files
583 were filtered to remove generic terms that annotate more than 500 genes, and the remaining list
584 of terms in the BP category were used for testing overlaps with clusters. The significance of
585 overlap of a target gene set (e.g. a cluster) with BP terms was calculated using a cumulative
586 hypergeometric test. The p-values obtained were adjusted for false discovery rate and converted
587 to qvalues using the Benjamini-Hochberg method (Benjamini and Hochberg, 1995). Enrichment
588 scores were reported as (-1) * log (qvalue).

589

590

591 **Analysis of known CREs**

592 We used a pattern-based method to search for CREs over-represented in the promoters of co-
593 regulated genes. First, all known plant motifs were identified from PLACE (Higo et al., 1999)
594 and AGRIS databases (Palaniswamy et al., 2006). Subsequently, 1000-bp upstream promoter
595 regions of all Arabidopsis genes were downloaded from TAIR and scanned for occurrence of
596 these motifs using DNA-pattern matching tool (Medina-Rivera et al., 2015), yielding a list of
597 403 motifs present at least once in the promoters of ~17000 genes. A few of these motifs,
598 perhaps involved in functions common to all the promoters, are ubiquitously present in almost all
599 the genes. To detect a reliable presence-absence signal in the context of our analysis, we
600 removed motifs that were found in more than 50% of all the genes considered in the network.
601 Thus, a list of 341 unique motifs were used for enrichment (overlap) analysis using a
602 hypergeometric test as described above.

603 **Module Regulatory Network analysis**

604 A list of 1921 Arabidopsis TF regulators was curated from the Plant Transcription Factor
605 Database, the AGRIS database and the Database of Arabidopsis Transcription Factors (Guo et

al., 2005; Yilmaz et al., 2011; Jin et al., 2014). For every TF-gene pair, a Z score representing specific correlation score was calculated using the CLR algorithm (Faith et al., 2007). The Parametric Analysis of Geneset Enrichment (PAGE) algorithm (Kim and Volsky, 2005) was used to evaluate enrichment of CLR scored targets of each TF within each module. P-values were calculated form Z scores of enrichment and corrected for FDR using the Benjamini and Hochberg procedure (Benjamini and Hochberg, 1995).

Global regulatory network and comparison of different networks

A global regulatory network was constructed the same way as the seed-specific network, except that EEC of 140 datasets was used. The Arabidopsis abiotic stress regulatory network was obtained from (Vermeirssen et al., 2014). Information on interactions reported in AtRegNet and AraNet was downloaded from http://arabidopsis.med.ohio-state.edu/downloads.html and http://www.functionalnet.org/aranet/download.html, respectively. Regulatory interactions (edges with at least one node as a regulator from our list) were identified from AraNet. For all three externally downloaded networks described above, and the global and seed-specific networks from this study, Jaccard coefficient (JC) of overlap in the predicted targets of each regulator pair was calculated using a perl script. JC scores were binned and the fraction of regulators retained from the original individual network within each bin was plotted in R. The random network was created by preserving the node degree and randomly reshuffling all the edges of the seed network.

Network data was parsed using the Sleipnir library (Huttenhower et al., 2008), Network Analysis Tools (NeAT) (Brohee et al., 2008) and scripts written in R and perl.

Author Contributions

C.G. and A.K. conceived the computational procedure. C.G. designed the network, conducted statistical analysis and drafted the manuscript. A.K. provided data. E.C. interpreted the results and contributed text. A.P. designed the experiments and coordinated research. C.G. created the webserver with contributions from P.W. All authors contributed to writing the manuscript.

Acknowledgements

This work was supported by the NSF grant award MCB-1052145 "A Systems Biology Approach to Cellular Regulation of Seed Filling"

Figure Legends

Figure 1: Pipeline for tissue-specific module regulatory network analysis. Two separate Arabidopsis gene expression compendiums (EC) were created: one from a seed-specific expression data series (GSE12404) and one from non-tissue-specific (global) 140 expression datasets. Datasets in both the EC were normalized individually using RMA algorithm in R. Z scores of Pearson's Correlations (PC) were calculated for all gene-pairs in both the EC. From seed EC, gene pairs with PC >0.73 (Z > 1.96) were connected to create the Arabidopsis Seed coexpression network (ASCN). ASCN was then clustered using SPICi at a range of clustering thresholds *(T_d),* and an optimal clustering parameter was chosen based on genome coverage and coherence of genes as a functional group. 1563 clusters obtained at T_d 0.80 were tested for enrichment of biological processes from the gene ontology and known plant *cis* regulatory elements for multiple databases. A list of 1921 TFs was supplied to the CLR (context likelihood of relatedness) algorithm to predict their targets in both the EC. In the seed EC, the PAGE algorithm was used to score the enrichment of CLR-weighted targets in the ASCN clusters, and a TF-module association network was created. The network was queried with a list of genes expressed predominantly in the seed as compared to other organs/tissues, and the Seed Active Network or SANe, was derived. Simultaneously, the seed-specific network was compared with the network created using the global EC and multiple other Arabidopsis regulatory networks downloaded from published studies.

Figure 2: Evaluation of clustering threshold (T_d). Genes from the Arabidopsis seed coexpression network were clustered at a range of T_d values shown on the X axis of all the figures. Each T_d was examined by: A) A genome coverage plot measuring the number of clusters yielded and the fraction of original genes retained (orange line corresponding orange Y axis). B) Boxplots showing average edge segregation of all the clusters, indicating overall modularity of

the network within each T_d. C) A plot showing number of clusters enriched with at least one BP term and the total number of BP terms retained (orange line corresponding orange Y axis) and D) Boxplots summarizing the enrichment scores [-1*log(FDR)] of the hypergeometric p-values obtained by BP-cluster overlap analysis.

Figure 3: TF-Module Association Network (TMN). A) Heatmap representing association scores of 1819 TFs as regulators in the rows, and 10526 genes grouped into 278 coexpression modules represented along the columns. Each grid in the heatmap is color coded according to the level of enrichment of predicted targets of each TF regulator in the corresponding module. The red gradient indicates a positive score and grey indicates a negative score, estimated using the PAGE algorithm. The seed compartment in which the module has maximum expression is color-coded and represented on top of the heatmap (first row), where CE is chalazal endosperm, ME is micropylar endosperm, PE is peripheral endosperm, CSC and GSC is chalazal and general seed coat, respectively, and EP is embryo proper. The development stage in which the module has maximum expression is color-coded and represented on top of the heatmap (second row), where bc is bending cotyledon, lc is linear cotyledon, pgs is pre-globular stage, ges is globular embryo stage, hs is heart stage and mg is mature green stages. B) Predictions for each of the 278 modules were ranked and the top 5 predicted regulators for each module were visualized as a network graph. Each grey circle in the network plot is a TF and each orange circle is a module. The size of the grey circle is proportional to the out-going degree of the TF. Size of the orange circle was set to a constant, except for 9 bigger circles showing the modules described later in the main text. The network was visualized using Cytoscape version 3.3.0. Node names are hidden for ease in visualization. The cytoscape sessions file is provided as supplemental data S3, which can be loaded into Cytoscape for node names and further exploration of the network. The heatmap was drawn using gplots package in the R statistical computing environment.

Figure 4: Heatmap representation of a subset of the SANe. Modules with high relative expression in embryo, endosperm and seed coat regions were extracted from SANe. Modules are shown in columns and for each module, the top 5 predicted TF regulators are shown in rows. Each grid in the heatmap is colored according to the association score estimated using the PAGE algorithm. Positive and negative scores are shaded in red or black gradient, respectively, as

693 indicated by the color key. Literature identified TFs with validated seed-specific phenotypes or
694 phenotypes observed in other reproductive stages/tissues are marked with a red ellipse or a
695 yellow ellipse, respectively. CRE, cell-type and functional annotations for each module are
696 shown above the heatmap (top three rows; colored boxes). Modules annotated for embryo,
697 endosperm and seed coat are indicated in blue, green and brown boxes, respectively, in the
698 middle row. CRE and functional annotation for each module is color- coded uniquely in the top
699 and bottom rows, respectively.

700 **Figure 5: Visualization of seed modules.** A graphical representation of seed coexpression
701 modules. Each circle represents a gene. Each module is color coded uniquely. Modules are
702 grouped according to the seed compartments (indicated by horizontal or vertical lines and text
703 boxes), and labelled with the BP term most highly over-represented within each module. Genes
704 are left unlabeled to facilitate visualization. The network was drawn in Cytoscape version 3.3.0. .

705 **Figure 6: Expression profiling of gene modules.** Expression patterns of modules in embryo,
706 endosperm and seed coat regions represented as heatmaps in A), B), and C), respectively. Seed
707 compartments are represented as columns and genes as rows. Gene names are hidden for ease in
708 visualization. Expression values of genes in each module were averaged across samples from the
709 same tissue-type/seed-compartment (embryo, endosperm and seed coat). Average expression
710 values were scaled and represented as a Z score in the heatmaps. Red indicates higher expression
711 of a gene in a particular compartment and black gradient indicates lower expression relative to
712 other compartments.

713 **Figure 7: Module M0006.** A) Expression patterns of genes in module M0006. Seed
714 development stages are represented as columns and genes as rows. Gene names are hidden for
715 ease in visualization. B) Coexpression links between TFs in M0006. Nodes are labeled according
716 to their corresponding gene symbols if present in TAIR, else labeled with their corresponding
717 locus ID and the family the protein belongs to.

718 **Figure 8: Characteristics of seed networks.** A) Comparison of the fraction of TFs possibly co-
719 regulating the same sets of genes, evaluated using the Jaccard's Index (JI) of overlap between the
720 predicted targets of each TF-pair, for 5 different regulatory networks and a random network. B)
721 A histogram showing bins of number of motifs significantly over-represented in the promoters of
722 genes within each module in TF-MAN. C) Distribution of TF-module edges in SANe follows a

723 scale-free topology, with a large number of regulators associated with fewer modules, and a few
724 regulators (e.g., NAP57, KAN3) associated with a large number of modules.

725

726

Tables

Table 1: **Regulatory hubs of seed development**. 23 regulators (transcription factors) that were found associated with the largest number of coexpressed modules in SANe were selected and listed in descending order acoording to the number of modules they regulate. Targets of these regulators in the seed and the global network, with absolute Z score > 3 were selected and tested for overlaps with BP terms in the GO database. The score columns represent (-1) * log (q-value) values from a cumulative hypergeometric test of enrichment. Only the most highly scored gene sets are reported in the table.

Network	**Seed**		**Global**	
Transcription Factor	**Biological Process**	**Enrichment Score**	**Biological Process**	**Enrichment Score**
NAP57 AT3G57150	ribonucleoprotein complex biogenesis	57.05	ribosome biogenesis	72.07
HDT3 (AT5G03740)	ribonucleoprotein complex biogenesis	45.50	ribonucleoprotein complex biogenesis	71.39
AT4G37130	ribonucleoprotein complex biogenesis	40.00	RNA metabolism	35.36
EMB2746 (AT5G63420)	ribonucleoprotein complex biogenesis	56.44	RNA metabolism	29.30
C3H (AT5G60820)	ribonucleoprotein complex biogenesis	12.97	vesicle-mediated transport	10.38
JMJ22 (AT5G06550)	ribonucleoprotein complex biogenesis	36.76	ribosome biogenesis	75.24
WRKY13 (AT4G39410)	immune system process	3.04	N.D	NA
TFIIIA (AT1G72050)	ribosome biogenesis	49.90	RNA metabolism	49.30
VOZ1 (AT1G28520)	ribosome biogenesis	13.30	cellular biopolymer catabolism	4.77
NFD1 (AT4G30930)	ribonucleoprotein complex biogenesis	71.39	ribosome biogenesis	75.24

KAN3 (AT4G17695)	jasmonic acid biosynthesis	4.64	response to salicylic acid stimulus	2.95
HDT1 (AT3G44750)	ribonucleoprotein complex biogenesis	41.25	ribosome biogenesis	72.44
HAT3.1 (AT3G19510)	RNA metabolism	7.88	RNA metabolism	20.48
FZF (AT2G24500)	RNA metabolism	23.67	ribosome biogenesis	68.91
IAA8 (AT2G22670)	polysaccharide metabolism	5.33	transmembrane receptor protein tyrosine kinase signaling pathway	10.47
SMAD/FHA (AT2G21530)	photosynthesis	54.25	photosynthesis	68.43
AT1G78280	cellular biopolymer metabolism	5.88	ribosome biogenesis	8.46
ZFP4 (AT1G66140)	N.D.	N.A.	ion transport	4.40
TRB1 (AT1G49950)	maintenance of root meristem identity	2.33	protein modification	6.33
SEUSS (AT1G43850)	microtubule-based process	2.71	negative regulation of gene expression	8.25
NAC017 (AT1G34190)	vesicle-mediated transport	4.70	vesicle-mediated transport	11.55
ATU2AF35A (AT1G27650)	RNA metabolism	20.60	RNA metabolism	15.07
AT1G17520	proteolysis	3.39	RNA metabolism	12.81

734

Parsed Citations

Albert R (2005) Scale-free networks in cell biology. Journal of Cell Science 118: 4947
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Allorent G, Osorio S, Ly Vu J, Falconet D, Jouhet J, Kuntz M, Fernie AR, Lerbs-Mache S, Macherel D, Courtois F, Finazzi G (2015) Adjustments of embryonic photosynthetic activity modulate seed fitness in Arabidopsis thaliana. New Phytologist 205: 707-719
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Ambavaram MM, Krishnan A, Trijatmiko KR, Pereira A (2011) Coordinated activation of cellulose and repression of lignin biosynthesis pathways in rice. In Plant Physiol, Vol 155, United States, pp 916-931
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Ambavaram MMR, Basu S, Krishnan A, Ramegowda V, Batlang U, Rahman L, Baisakh N, Pereira A (2014) Coordinated regulation of photosynthesis in rice increases yield and tolerance to environmental stress. Nat Commun 5
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Aoki Y, Okamura Y, Tadaka S, Kinoshita K, Obayashi T (2016) ATTED-II in 2016: A Plant Coexpression Database Towards Lineage-Specific Coexpression. Plant Cell Physiol 57: e5
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Bailey KJ, Gray JE, Walker RP, Leegood RC (2007) Coordinate Regulation of Phosphoenolpyruvate Carboxylase and Phosphoenolpyruvate Carboxykinase by Light and CO2 during C4 Photosynthesis. Plant Physiology 144: 479-486
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Edgar R (2007) NCBI GEO: mining tens of millions of expression profiles-database and tools update. Nucl Acids Res 35
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A (2005) Reverse engineering of regulatory networks in human B cells. Nat Genet 37: 382-390
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Baud S, Mendoza MS, To A, Harscoet E, Lepiniec L, Dubreucq B (2007) WRINKLED1 specifies the regulatory action of LEAFY COTYLEDON2 towards fatty acid metabolism during seed maturation in Arabidopsis. Plant J 50: 825-838
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Belmonte MF, Kirkbride RC, Stone SL, Pelletier JM, Bui AQ, Yeung EC, Hashimoto M, Fei J, Harada CM, Munoz MD, Le BH, Drews GN, Brady SM, Goldberg RB, Harada JJ (2013) Comprehensive developmental profiles of gene activity in regions and subregions of the Arabidopsis seed. Proceedings of the National Academy of Sciences 110: E435-E444
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological) 57: 289-300
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Boutilier K, Offringa R, Sharma VK, Kieft H, Ouellet T, Zhang L, Hattori J, Liu C-M, van Lammeren AAM, Miki BLA, Custers JBM, van Lookeren Campagne MM (2002) Ectopic Expression of BABY BOOM Triggers a Conversion from Vegetative to Embryonic Growth. The Plant Cell 14: 1737-1749
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Brohee S, Faust K, Lima-Mendez G, Sand O, Janky R, Vanderstocken G, Deville Y, van Helden J (2008) NeAT: a toolbox for the analysis of biological networks, clusters, classes and pathways. Nucleic Acids Res 36: W444-451
Pubmed: Author and Title
CrossRef: Author and Title

Google Scholar: Author Only Title Only Author and Title

Castillo-Davis CI, Hartl DL (2003) GeneMerge--post-genomic analysis, data mining, and hypothesis testing. Bioinformatics 19: 891-892
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Chavez Montes RA, Coello G, Gonzalez-Aguilera KL, Marsch-Martinez N, de Folter S, Alvarez-Buylla ER (2014) ARACNe-based inference, using curated microarray data, of Arabidopsis thaliana root transcriptional regulatory networks. In BMC Plant Biol, Vol 14, England, p 97
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Chen C, Grennan K, Badner J, Zhang D, Gershon E, Jin L, Liu C (2011) Removing Batch Effects in Analysis of Expression Microarray Data: An Evaluation of Six Batch Adjustment Methods. PLoS ONE 6: e17238
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Chen L-Q, Lin IW, Qu X-Q, Sosso D, McFarlane HE, Londoño A, Samuels AL, Frommer WB (2015) A Cascade of Sequentially Expressed Sucrose Transporters in the Seed Coat and Endosperm Provides Nutrition for the Arabidopsis Embryo. The Plant Cell 27: 607-619
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Childs KL, Davidson RM, Buell CR (2011) Gene Coexpression Network Analysis as a Source of Functional Annotation for Rice Genes. PLoS ONE 6: e22196
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Comparot-Moss S, Denyer K (2009) The evolution of the starch biosynthetic pathway in cereals and other grasses. Journal of Experimental Botany 60: 2481-2492
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Devaiah BN, Madhuvanthi R, Karthikeyan AS, Raghothama KG (2009) Phosphate Starvation Responses and Gibberellic Acid Biosynthesis Are Regulated by the MYB62 Transcription Factor in Arabidopsis. Molecular Plant 2: 43-58
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences 95: 14863-14868
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Eshed Y, Izhaki A, Baum SF, Floyd SK, Bowman JL (2004) Asymmetric leaf development and blade expansion in Arabidopsis are mediated by KANADI and YABBY activities. Development 131: 2997-3006
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Fait A, Angelovici R, Less H, Ohad I, Urbanczyk-Wochniak E, Fernie AR, Galili G (2006) Arabidopsis Seed Development and Germination Is Associated with Temporally Distinct Metabolic Switches. Plant Physiology 142: 839-854
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS (2007) Large-Scale Mapping and Validation of Escherichia coli Transcriptional Regulation from a Compendium of Expression Profiles. PLoS Biol 5: e8
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Escherichia coli Transcriptional Regulation from a Compendium of Expression Profiles.&num=10&btnG=Search+Scholar&as_epq=&as_oq=&as_eq=&as_occt=any&as_publication=&as_yhi=&as_allsubj=all&hl=en&lr=&c2coff=1" target="_blank">Title Only Escherichia coli Transcriptional Regulation from a Compendium of Expression Profiles.&num=10&btnG=Search+Scholar&as_occt=any&as_sauthors=Faith&as_ylo=2007&as_allsubj=all&hl=en&c2coff=1" target="_blank">Author and Title

Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS (2007) Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS Biol 5
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Goffman FD, Alonso AP, Schwender J, Shachar-Hill Y, Ohlrogge JB (2005) Light enables a very high efficiency of carbon storage in

developing embryos of rapeseed. Plant Physiol 138: 2269-2279
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Goffman FD, Ruckle M, Ohlrogge J, Shachar-Hill Y (2004) Carbon dioxide concentrations are very high in developing oilseeds. Plant Physiol Biochem 42: 703-708
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

González-Morales SI, Chávez-Montes RA, Hayano-Kanashiro C, Alejo-Jacuinde G, Rico-Cambron TY, de Folter S, Herrera-Estrella L (2016) Regulatory network analysis reveals novel regulators of seed desiccation tolerance in Arabidopsis thaliana. Proceedings of the National Academy of Sciences 113: E5232-E5241
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Grimault A, Gendrot G, Chaignon S, Gilard F, Tcherkez G, Thévenin J, Dubreucq B, Depège-Fargeix N, Rogowsky PM (2015) Role of B3 domain transcription factors of the AFL family in maize kernel filling. Plant Science 236: 116-125
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Grossniklaus U, Vielle-Calzada JP, Hoeppner MA, Gagliano WB (1998) Maternal control of embryogenesis by MEDEA, a polycomb group gene in Arabidopsis. Science 280: 446-450
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Guo A, He K, Liu D, Bai S, Gu X, Wei L, Luo J (2005) DATF: a database of Arabidopsis transcription factors. Bioinformatics 21: 2568-2569
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Harb A, Krishnan A, Ambavaram MMR, Pereira A (2010) Molecular and Physiological Analysis of Drought Stress in Arabidopsis Reveals Early Responses Leading to Acclimation in Plant Growth. Plant Physiology 154: 1254-1271
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

He L, Su C, Wang Y, Wei Z (2015) ATDOF5.8 protein is the upstream regulator of ANAC069 and is responsive to abiotic stress. Biochimie 110: 17-24
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Higo K, Ugawa Y, Iwamoto M, Korenaga T (1999) Plant cis-acting regulatory DNA elements (PLACE) database: 1999. Nucleic Acids Res 27: 297-300
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Husbands AY, Benkovics AH, Nogueira FTS, Lodha M, Timmermans MCP (2015) The ASYMMETRIC LEAVES Complex Employs Multiple Modes of Regulation to Affect Adaxial-Abaxial Patterning and Leaf Complexity[OPEN]. The Plant Cell
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Huttenhower C, Hibbs M, Myers C, Troyanskaya OG (2006) A scalable method for integration and functional analysis of multiple microarray datasets. Bioinformatics 22: 2890-2897
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Huttenhower C, Mutungu KT, Indik N, Yang W, Schroeder M, Forman JJ, Troyanskaya OG, Coller HA (2009) Detailing regulatory networks through large scale data integration. Bioinformatics 25: 3267-3274
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Huttenhower C, Schroeder M, Chikina MD, Troyanskaya OG (2008) The Sleipnir library for computational functional genomics. Bioinformatics 24: 1559-1561
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P (2010) Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. PLoS ONE 5: e12776

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4: 249-264
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Izhaki A, Bowman JL (2007) KANADI and Class III HD-Zip Gene Families Regulate Embryo Patterning and Modulate Auxin Flow during Embryogenesis in Arabidopsis. The Plant Cell 19: 495-508
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Jia H, McCarty DR, Suzuki M (2013) Distinct Roles of LAFL Network Genes in Promoting the Embryonic Seedling Fate in the Absence of VAL Repression. Plant Physiology 163: 1293-1305
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Jiang P, Singh M (2010) SPICi: a fast clustering algorithm for large biological networks. Bioinformatics 26: 1105-1111
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Jin J, Zhang H, Kong L, Gao G, Luo J (2014) PlantTFDB 3.0: a portal for the functional and evolutionary study of plant transcription factors. Nucleic Acids Research 42: D1182-D1187
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Johnson CS, Kolevski B, Smyth DR (2002) TRANSPARENT TESTA GLABRA2, a Trichome and Seed Coat Development Gene of Arabidopsis, Encodes a WRKY Transcription Factor. The Plant Cell 14: 1359-1375
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Kannan K, Nelson AD, Shippen DE (2008) Dyskerin is a component of the Arabidopsis telomerase RNP required for telomere maintenance. Mol Cell Biol 28: 2332-2341
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Keith K, Kraml M, Dengler NG, McCourt P (1994) fusca3: A Heterochronic Mutation Affecting Late Embryo Development in Arabidopsis. The Plant Cell 6: 589-600
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Kim S-Y, Volsky DJ (2005) PAGE: Parametric Analysis of Gene Set Enrichment. BMC Bioinformatics 6: 144-144
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Kim Y-S, Kim S-G, Park J-E, Park H-Y, Lim M-H, Chua N-H, Park C-M (2006) A Membrane-Bound NAC Transcription Factor Regulates Cell Division in Arabidopsis. The Plant Cell 18: 3132-3144
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Kohler C, Hennig L, Bouveret R, Gheyselinck J, Grossniklaus U, Gruissem W (2003) Arabidopsis MSI1 is a component of the MEA/FIE Polycomb group complex and required for seed development. Embo j 22: 4804-4814
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Kwong RW, Bui AQ, Lee H, Kwong LW, Fischer RL, Goldberg RB, Harada JJ (2003) LEAFY COTYLEDON1-LIKE Defines a Class of Regulators Essential for Embryo Development. The Plant Cell 15: 5-18
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Langfelder P, Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9: 559
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Lashbrooke JG, Cohen H, Levy-Samocha D, Tzfadia O, Panizel I, Zeisler V, Massalha H, Stern A, Trainotti L, Schreiber L, Costa F, Aharoni A (2016) MYB107 and MYB9 Homologs Regulate Suberin Deposition in Angiosperms. The Plant Cell

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Le BH, Cheng C, Bui AQ, Wagmaister JA, Henry KF, Pelletier J, Kwong L, Belmonte M, Kirkbride R, Horvath S, Drews GN, Fischer RL, Okamuro JK, Harada JJ, Goldberg RB (2010) Global analysis of gene activity during Arabidopsis seed development and identification of seed-specific transcription factors. Proceedings of the National Academy of Sciences 107: 8063-8070
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Lee I, Ambaru B, Thakkar P, Marcotte EM, Rhee SY (2010) Rational association of genes with traits using a genome-scale gene network for Arabidopsis thaliana. Nat Biotechnol 28: 149-156
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Lopes CT, Franz M, Kazi F, Donaldson SL, Morris Q, Bader GD (2010) Cytoscape Web: an interactive web-based network browser. Bioinformatics 26
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Lotan T, Ohto M, Yee KM, West MA, Lo R, Kwong RW, Yamagishi K, Fischer RL, Goldberg RB, Harada JJ (1998) Arabidopsis LEAFY COTYLEDON1 is sufficient to induce embryo development in vegetative cells. Cell 93: 1195-1205
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Luerßen H, Kirik V, Herrmann P, Miséra S (1998) FUSCA3encodes a protein with a conserved VP1/ABI3-like B3 domain which is of functional importance for the regulation of seed maturation inArabidopsis thaliana. 15: 755-764
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Luo M, Dennis ES, Berger F, Peacock WJ, Chaudhury A (2005) MINISEED3 (MINI3), a WRKY family gene, and HAIKU2 (IKU2), a leucine-rich repeat (LRR) KINASE gene, are regulators of seed size in Arabidopsis. Proceedings of the National Academy of Sciences of the United States of America 102: 17531-17536
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Ma H-W, Buer J, Zeng A-P (2004) Hierarchical structure and modules in the Escherichia coli transcriptional regulatory network revealed by a new top-down approach. BMC Bioinformatics 5: 199
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Magallón S, Sanderson MJ (2002) Relationships among seed plants inferred from highly conserved genes: sorting conflicting phylogenetic signals among ancient lineages. American Journal of Botany 89: 1991-2006
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Favera RD, Califano A (2006) ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. BMC Bioinformatics 7: S7-S7
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

McAbee JM, Hill TA, Skinner DJ, Izhaki A, Hauser BA, Meister RJ, Venugopala Reddy G, Meyerowitz EM, Bowman JL, Gasser CS (2006) ABERRANT TESTA SHAPE encodes a KANADI family member, linking polarity determination to separation and growth of Arabidopsis ovule integuments. 46: 522-531
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Medina-Rivera A, Defrance M, Sand O, Herrmann C, Castro-Mondragon JA, Delerce J, Jaeger S, Blanchet C, Vincens P, Caron C, Staines DM, Contreras-Moreira B, Artufel M, Charbonnier-Khamvongsa L, Hernandez C, Thieffry D, Thomas-Chollier M, van Helden J (2015) RSAT 2015: Regulatory Sequence Analysis Tools. Nucleic Acids Research
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Nesi N, Jond C, Debeaujon I, Caboche M, Lepiniec L (2001) The Arabidopsis TT2 Gene Encodes an R2R3 MYB Domain Protein That Acts as a Key Determinant for Proanthocyanidin Accumulation in Developing Seed. The Plant Cell 13: 2099-2114
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. Physical Review E 69: 026113
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Nygaard V, Rødland EA, Hovig E (2015) Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. Biostatistics
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Obayashi T, Kinoshita K (2011) COXPRESdb: a database to compare gene coexpression in seven model animals. Nucleic Acids Research 39: D1016-D1022
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Ogas J, Kaufmann S, Henderson J, Somerville C (1999) PICKLE is a CHD3 chromatin-remodeling factor that regulates the transition from embryonic to vegetative development in Arabidopsis. Proceedings of the National Academy of Sciences 96: 13839-13844
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Palaniswamy SK, James S, Sun H, Lamb RS, Davuluri RV, Grotewold E (2006) AGRIS and AtRegNet. A Platform to Link cis-Regulatory Elements and Transcription Factors into Regulatory Networks. Plant Physiology 140: 818-829
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Portereiko MF, Lloyd A, Steffen JG, Punwani JA, Otsuga D, Drews GN (2006) AGL80 Is Required for Central Cell and Endosperm Development in Arabidopsis. The Plant Cell 18: 1862-1872
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Qiu D, Xiao J, Ding X, Xiong M, Cai M, Cao Y, Li X, Xu C, Wang S (2007) OsWRKY13 Mediates Rice Disease Resistance by Regulating Defense-Related Genes in Salicylate- and Jasmonate-Dependent Signaling. Molecular Plant-Microbe Interactions 20: 492-499
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Ruuska SA, Girke T, Benning C, Ohlrogge JB (2002) Contrapuntal Networks of Gene Expression during Arabidopsis Seed Filling. The Plant Cell 14: 1191-1206
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Ruuska SA, Schwender J, Ohlrogge JB (2004) The Capacity of Green Oilseeds to Utilize Photosynthesis to Drive Biosynthetic Processes. Plant Physiology 136: 2700-2709
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Sato Y, Namiki N, Takehisa H, Kamatsuki K, Minami H, Ikawa H, Ohyanagi H, Sugimoto K, Itoh J-I, Antonio BA, Nagamura Y (2012) RiceFREND: a platform for retrieving coexpressed gene networks in rice. Nucleic Acids Research
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Schwender J, Goffman F, Ohlrogge JB, Shachar-Hill Y (2004) Rubisco without the Calvin cycle improves the carbon efficiency of developing green seeds. Nature 432: 779-782
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res 13
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Spitz F, Furlong EE (2012) Transcription factors: from enhancer binding to developmental control. Nat Rev Genet 13: 613-626
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Sreenivasulu N, Wobus U (2013) Seed-development programs: a systems biology-based comparison between dicots and monocots. Annu Rev Plant Biol 64: 189-217

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Suzuki M, McCarty DR (2008) Functional symmetry of the B3 network controlling seed development. Current Opinion in Plant Biology 11: 548-553

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

To A, Valon C, Savino G, Guilleminot J, Devic M, Giraudat J, Parcy F (2006) A Network of Local and Redundant Gene Regulation Governs Arabidopsis Seed Maturation. The Plant Cell 18: 1642-1651

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Vermeirssen V, De Clercq I, Van Parys T, Van Breusegem F, Van de Peer Y (2014) Arabidopsis Ensemble Reverse-Engineered Gene Regulatory Network Discloses Interconnected Transcription Factors in Oxidative Stress. The Plant Cell 26: 4656-4679

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Vriet C, Welham T, Brachmann A, Pike M, Pike J, Perry J, Parniske M, Sato S, Tabata S, Smith AM, Wang TL (2010) A Suite of Lotus japonicus Starch Mutants Reveals Both Conserved and Novel Features of Starch Metabolism. Plant Physiology 154: 643-655

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Wendrich JR, Weijers D (2013) The Arabidopsis embryo as a miniature morphogenesis model. New Phytologist 199: 14-25

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Xiao J, Cheng H, Li X, Xiao J, Xu C, Wang S (2013) Rice WRKY13 Regulates Cross Talk between Abiotic and Biotic Stress Signaling Pathways by Selective Binding to Different cis-Elements. Plant Physiology 163: 1868-1882

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Yamamoto A, Kagaya Y, Toyoshima R, Kagaya M, Takeda S, Hattori T (2009) Arabidopsis NF-YB subunits LEC1 and LEC1-LIKE activate transcription by interacting with seed-specific ABRE-binding factors. 58: 843-856

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Yeung KY, Haynor DR, Ruzzo WL (2001) Validating clustering for gene expression data. Bioinformatics 17: 309-318

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Yi X, Du Z, Su Z (2013) PlantGSEA: a gene set enrichment analysis toolkit for plant community. Nucleic Acids Res 41: W98-103

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Yilmaz A, Mejia-Guerra MK, Kurz K, Liang X, Welch L, Grotewold E (2011) AGRIS: the Arabidopsis Gene Regulatory Information Server, an update. Nucleic Acids Res 39: D1118-1122

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Yim WC, Yu Y, Song K, Jang CS, Lee B-M (2013) PLANEX: the plant co-expression database. BMC Plant Biology 13: 83

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Yu X, Li L, Zola J, Aluru M, Ye H, Foudree A, Guo H, Anderson S, Aluru S, Liu P, Rodermel S, Yin Y (2011) A brassinosteroid transcriptional network revealed by genome-wide identification of BESI target genes in Arabidopsis thaliana. Plant J 65: 634-646

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Zimmermann P, Hirsch-Hoffmann M, Hennig L, Gruissem W (2004) GENEVESTIGATOR. Arabidopsis Microarray Database and Analysis Toolbox. Plant Physiology 136: 2621-2632

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title