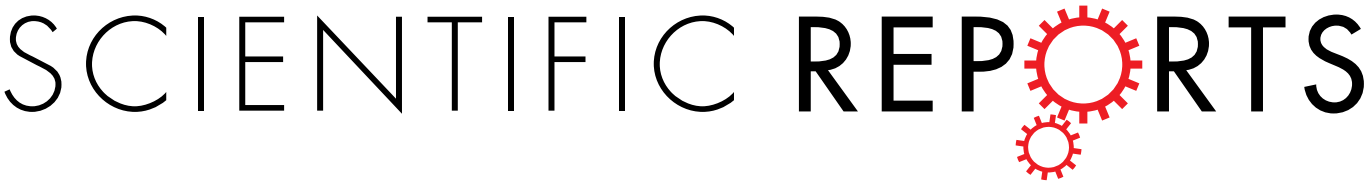


OPEN

Pan- and core- network analysis of co-expression genes in a model plant

Received: 09 August 2016
Accepted: 14 November 2016
Published: 16 December 2016

Fei He[1] & Sergei Maslov[1,2,3,4]

Genome-wide gene expression experiments have been performed using the model plant Arabidopsis during the last decade. Some studies involved construction of coexpression networks, a popular technique used to identify groups of co-regulated genes, to infer unknown gene functions. One approach is to construct a single coexpression network by combining multiple expression datasets generated in different labs. We advocate a complementary approach in which we construct a large collection of 134 coexpression networks based on expression datasets reported in individual publications. To this end we reanalyzed public expression data. To describe this collection of networks we introduced concepts of 'pan-network' and 'core-network' representing union and intersection between a sizeable fractions of individual networks, respectively. We showed that these two types of networks are different both in terms of their topology and biological function of interacting genes. For example, the modules of the pan-network are enriched in regulatory and signaling functions, while the modules of the core-network tend to include components of large macromolecular complexes such as ribosomes and photosynthetic machinery. Our analysis is aimed to help the plant research community to better explore the information contained within the existing vast collection of gene expression data in Arabidopsis.

Coexpression networks represent all pairwise relationships of genes that have similar profiles in a given set of expression samples. Expression levels of such connected genes are either directly or indirectly co-regulated by the same regulatory elements. Since genes under the same regulatory control tend to be functionally associated, the most common use of coexpression networks is to infer unknown and to validate known gene functional roles and regulatory interactions between genes[1]. A coexpression network consists of all significantly correlated pairs of genes with correlation coefficients above a certain threshold. Once a coexpression network is constructed, the next step usually involves identification of densely interconnected modules, which are often enriched with genes involved in a specific biological function[2]. Coexpression network analysis has been actively used in plant functional genomics. For example, new genes involved in flavonoid biosynthetic process[3], starch metabolism[4], aliphatic glucosinolate biosynthesis[5], lignin biosynthesis[6,7] and photorespiration[8] have been identified with the help of coexpression networks. Compared with animal (especially human) data, functional gene annotation in plants is less comprehensive even in a well-studied model organism such as Arabidopsis. Thus it is especially valuable to leverage the use of the existing plant transcriptomics data in order to improve identification of new and validation of existing gene and protein functions[9–12].

A common approach to building coexpression networks is to infer correlation relationships from a combination of multiple expression datasets produced by different labs[1,2,13–15]. For example, Mao *et al.* combined all the datasets from the AtGenExpress project (~1000 samples) to construct a coexpression network[13]. The larger sample size improves the statistical significance of relationships between genes. The inevitable experimental noise within microarray data may give rise to false positive interactions in which pairs of genes have high degree of coexpression in only one dataset but very low coexpression in other datasets[16]. The traditional approach relies on increasing the number of samples to infer more reliable correlation relationships[17]. On the other hand,

[1]Biology Department, Brookhaven National Laboratory, Upton, NY 11973, USA. [2]Department of Bioengineering, Department of Bioengineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA. [3]Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA. [4]National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA. Correspondence and requests for materials should be addressed to F.H. (email: plane83@gmail.com) or S.M. (email: maslov@illinois.edu)

indiscriminately combining multiple samples may not be universally good. The combined samples need to be biologically comparable[18]. Furthermore, batch effects may give rise to false positive and spurious correlations between genes when microarray data from different labs are combined[19,20].

Another problem with combined co-expression networks is that it may miss rare gene interactions formed under specific conditions such as a particular disease[21]. Increasing amount of evidence indicates that different gene networks operate in different biological contexts[22,23]. Thus, it is increasingly important to compare and contrast coexpression networks generated from individual datasets[24–26]. Experimental results suggest that more than one third of genetic interactions are condition-specific[27]. Several studies have also demonstrated that coexpression of genes varies in different conditions. Southworth *et al.* studied the difference between coexpression networks of young and old mouse brains and found genes involved in memory have more network connections in the young than in the old animals[28]. By leveraging the concept of differential rewiring, Hudson *et al.* captured the phenotypic differences between two breeds of cows[29]. Compared with normal tissue, many coexpression relationships were lost in cancer tissue[30].

A published study of changes in gene expression usually has its own experimental design created in order to answer a specific biological question or several related questions, such as, to understand the mechanisms of plant heat shock response[31] or biological function of a plant hormone[32]. It might not be universally good to detect coexpression based on a combined dataset of microarray samples from different labs. Here we construct and analyze a comprehensive collection of 134 coexpression networks each based on expression samples from an individual published study, thus preserving context-specific network structure. We assume the network generated from each Gene Expression Omnibus (GEO) series represent a particular regulatory response specific to the experimental design and biological query of that study. Therefore, expression datasets from individual studies are ideal for the detection of condition-specific networks. We found most coexpression edges can only be detected in a few networks, while only a small fraction of edges can be repeatedly detected in many networks. Those two types of edges are enriched for distinctly different biological functions. Our analysis may provide guide for future coexpression network analysis in plants.

Results

A large collection of Arabidopsis coexpression networks. Previous work has combined expression datasets from many labs to build coexpression networks for animals or plants[1,13,14]. Although this strategy has been intensively used by plant scientists to assist gene characterization[33], it preferentially captures the relationships between genes that are conserved across most contexts. In contrast to this, in this study we built coexpression networks based on individual expression datasets from the model plant Arabidopsis in order to capture network aspects that appeared in a specific experimental setup[21].

Thousands of gene expression profiling datasets are available for Arabidopsis in public repositories such as GEO. We focused on the Affymetrix GPL198 platform, since it is the most widely used platform and its annotation is continuously updated. Many of those datasets are not suitable for network inference simply because the number of samples are not enough for a robust inference of correlation. Similar to a previous study performed in humans, we limited our analysis to datasets with at least 20 samples; 134 such datasets were acquired from GEO. Each of them was normalized in the same manner, and genes with very low mRNA abundance or genes without any significant changes were removed (see Methods). The top 0.1% most coexpressed gene pairs within each dataset were then used to build individual networks[34]. We used GSE series number to identify individual published studies, thus each of our 134 networks is labelled with the GSE number of the corresponding GEO experiment/series (Fig. 1).

The number of nodes in most of the series-based networks was between 500 and 5,000, while the number of edges was between 50,000 and 100,000 (Fig. 2a,b). The Pearson Correlation Coefficient (PCC) cutoff for each series-based network was about 0.73 (Figure S1). Since the number of samples for each coexpressed gene pair is at least 20, the correlation is statistically significant ($p\text{-value} < 0.001$ for $PCC = 0.73$ and sample size ≥ 20). About 60% of these series-based networks were inferred from leaf, seedling and root tissues (Fig. 2c). The other 40% included other tissues such as flowers, seeds and shoots (Fig. 2c). As can be seen in Fig. 2d, these series-based networks were inferred from samples in a variety of experimental contexts, such as the effect of gene mutations or abiotic stress. The breadth of data sources were included to better capture of coexpression networks in different conditions.

Coexpression networks in Arabidopsis are highly context-dependent. Similar to the idea of 'pan/core genome' in bacteria[35,36] and plants[37–39], we propose to use the term 'pan-network' for the union of all the 134 series-based networks, while 'core-network' to represent the intersection of a considerable fraction (10 or more out of 134) of these networks. Indeed, unlike a large number of core genes shared across the entire networks there were no edges present in all 134 of our networks. This is similar in spirit to a soft cutoff used by one of us[40] in detecting the core ("basic") genome of a bacterial species (*E. coli*).

The pan-network representing the union of all 134 of our individual networks contains 2,294,175 non-redundant edges and 18301 nodes/genes (Supplemental File 1). Every edge in the pan-network is characterized by its 'universality number', U, defined as the total number of our networks in which this edge was observed. More than 80% of the edges were observed in only one network (universality, $U = 1$), suggesting that gene networks in Arabidopsis operating under different conditions are drastically different from each other[25]. Compared to a more conventional approach in which the samples from all of 134 experiments are combined in one large table based on which a single coexpression network is calculated, our approach detects more edges (Figure S2a). In addition, our pan network has a better quality as it contains less false–positive interactions due to e.g. the batch effect. Indeed a higher fraction of edges in our pan-network than in the network based on the coexpression of all 134 samples are supported by experiments other than expression data (Figure S2b and c).

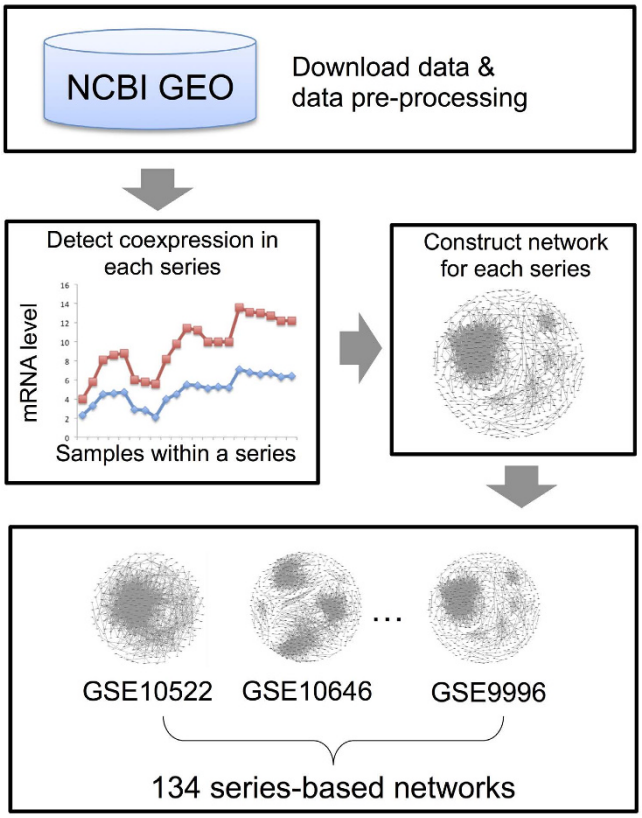


Figure 1. Our pipeline for construction of 134 GSE series-based coexpression networks in Arabidopsis.

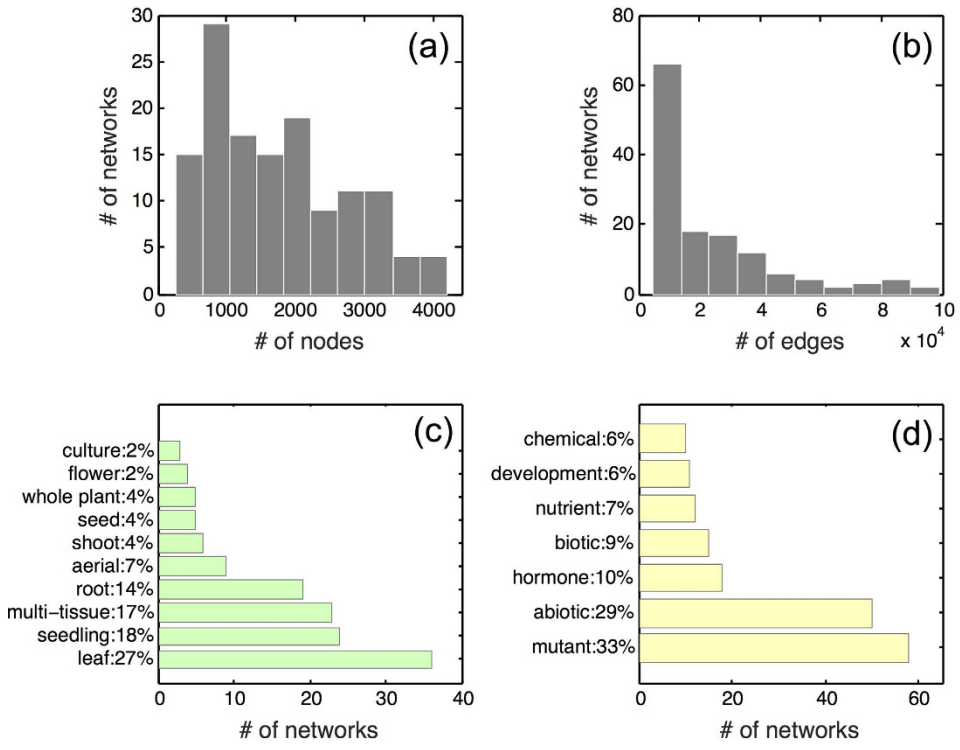


Figure 2. Basic statistics for 134 GSE series-based networks. (**a**) The histogram for the number of nodes in each series-based network; (**b**) The histogram for the number edges in each series-based network; (**c**) The bar chart of tissue types of the data source; (**d**) The bar chart of experimental conditions of the data source.

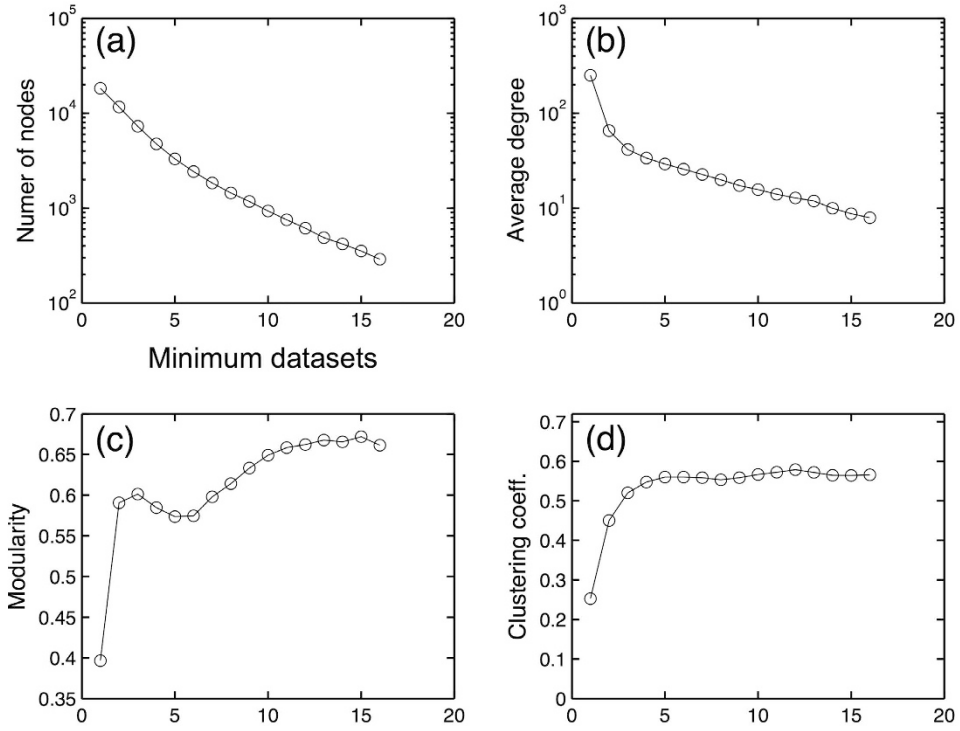


Figure 3. Cutoff selection for construction of the core-network. Edge universality, U_i, is given by the number of network datasets it was observed. The number of nodes (**a**), the average degree (**b**), the modularity (**c**) and the clustering coefficient (**d**) of the network constructed from edges with universalities U_i greater than or equal to the X-coordinate of the plot.

Co-expression edges between kinases and transcription factors (TF) are often of a particular biological interest as they may indicate gene regulation triggered by signaling pathways. To illustrate this below we discuss several examples of edges with experimental evidence of physical interactions[41]. Firstly, we observed that a guanylate kinase (AT3G57550) is coexpressed with a MYB TF (AT1G18570) in only one out of the 134 experiments (GSE40354, Supplemental File 1). The experimental context of the coexpression between those two genes involved treatment with bacterial elicitor[42]. Another edge deserving further investigation connects an F-box protein (AT1G47340) and SKP-1 (AT5G42190) (Supplemental File 1). F-box proteins are well known to interact with SKP-1 to degrade unwanted proteins[43], however, hundreds of F-box genes are encoded in the Arabidopsis genome[44]. It is critical to determine the specificity of the interactions between those F-box genes and their interacting partners. It is also important to understand under which environmental conditions the interaction happens[45]. Although previous experiments have already detected physical interactions between the edges discussed above[41], the results from our analysis suggest the design of further experiments to reveal the specificity of these interactions. In contrast to the edges that were observed in only one dataset (i.e. $U = 1$), the edges with larger values of U (Universality) were mostly formed between members of large multi-protein complexes (Supplemental File 1).

The core-network connects components of large molecular machines. A family of core networks of progressively increasing universality can be extracted from the pan-network by applying a strict cutoff on the universality of edges (e.g. a core-network formed by all edges existing in at least 5 datasets). As the cutoff value increases, the resulting network becomes smaller but more modular (Fig. 3). Modularity measures how well are these network modules separated from each other, while the clustering coefficient measures how tightly the neighbors of a node within a module are connected with each other. Both parameters are frequently reported for all types of biological networks, including protein-protein networks, metabolic networks and transcriptional networks[46] but (to the best of our knowledge) ours is the first study of their systematic dependence on edge universality.

We used $U = 10$ as the cutoff to determine the core-network used in the rest of our study (marked red in Fig. 4), which contains 7326 non-redundant edges among 935 genes. The exact value of the cutoff defining the core-network is always somewhat arbitrary. Our choice was inspired by the network analysis shown in Fig. 3. Both clustering coefficient and especially modularity of the core network appear to saturate above the $U = 10$ threshold. Interacting genes in thus defined core network are likely to be statistically significantly correlated even in datasets where they didn't make the PCC cutoff. Indeed, the average PCC of core-network edges in datasets where they did not pass the PCC cutoff is still significantly higher than the average PCC of edges within the pan-network (student's t-test, *p-value* = 0, see Figure S3). We refer to the set of edges present in the pan-network but not universal enough to be included in the core network as "condition-specific" (marked green in Fig. 4). The degree distribution of the core-network approximately follows a power-law (scale-free) pattern with the exponent

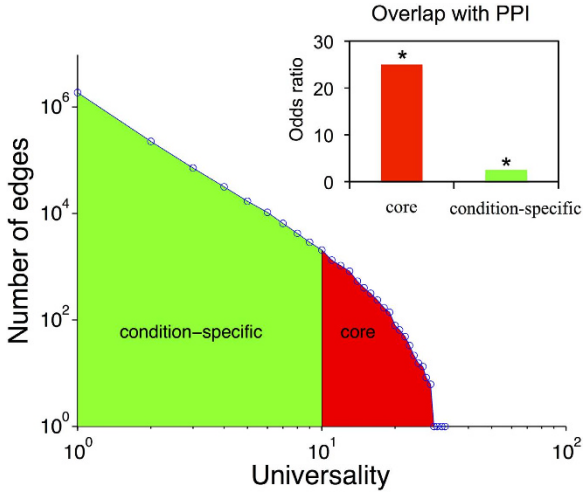


Figure 4. The pan-network contains both core and condition-specific edges. The overlap with the protein-protein interaction (PPI) network for both types of edges are statistically significant ($p\text{-}value < 0.001$). The PPI data was downloaded from BioGRID version 3.4.132 (http://thebiogrid.org/).

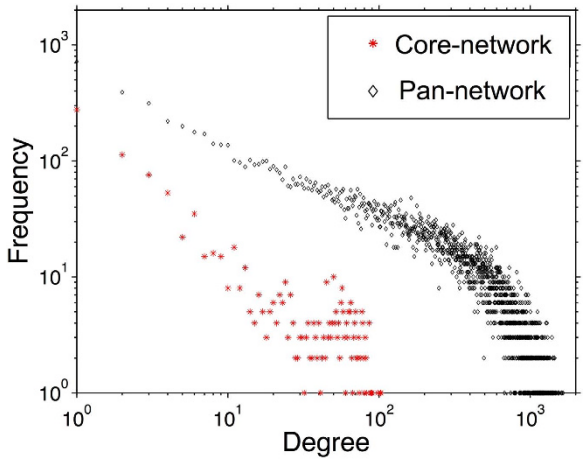


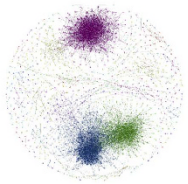
Figure 5. Degree distribution for pan- and core-networks. The y-axis is the number of nodes with a particular degree.

-1.5 transitioning into exponential cutoff above 50 (Fig. 5, Supplemental File 2)[47,48]. Many of its edges connect parts of large multi-protein complexes such as the ribosome and photosystems (Supplemental File 3). In fact, the network is enriched for physically interacting (binding) proteins[49] ($p\text{-}value < 0.001$, Fig. 4). The modules of this network were also highly enriched in genes characterized by functional categories 'translation' or 'photosynthesis' ($p\text{-}value < 10^{-10}$, Supplemental File 4). These facts are consistent with earlier observation that the genes involved in large molecular machines tend be coexpressed under many conditions[13].

The pan-network is enriched in condition-specific biological processes. Since each expression dataset usually has its own experimental design addressing a specific biological question[50], an edge detected in one dataset may not be detected in another[21]. As more and more evidence supports condition-specific networks in animals[28–30], we hypothesized that biological processes that are active in a small set of specific contexts would form the bulk of our pan-network. First of all, although the edges with smaller values of U contained fewer direct physical Protein-Protein Interactions (PPIs) compared with the core-network, PPIs are still overrepresented among pan-network edges ($p\text{-}value < 0.001$, Fig. 4). This suggests that biologically meaningful connections exist among co-expressed genes which are less likely to be detected by the traditional methods[16]. The degree distribution for the pan-network approximately followed a power-law pattern (exponent $= -0.5$ transitioning to the exponential cutoff above 500) (Fig. 5, Supplemental File 2). Besides the support from physical interactions, we wondered if more evidence could be found to reveal the biological significance of the pan-network.

With an average degree more than 200 (Supplemental File 2), an overview of the pan-network showed a densely connected large central component. However, we were able to detect network communities (i.e. modules)

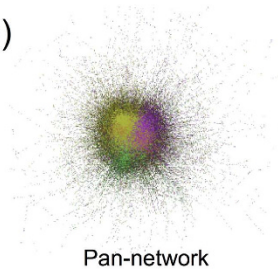
(a)



Core-network

Module ID	Represented enrichment	p-value
C4	translation	1.03E-202
C6	photosynthesis	1.74E-49
C7	translation	5.72E-28
C13	photosynthesis	7.77E-23
C22	response to chitin	3.99E-18
C35	response to radiation	8.43E-14
C58	cellular polysaccharide catabolic process	5.04E-08
C42	ribonucleoprotein complex biogenesis	7.14E-07

(b)



Pan-network

P4	regulation of cell cycle	5.39E-15
P5	phenylpropanoid metabolic process	1.12E-15
P2	plastid membrane organization	2.18E-15
P3	regulation of cell communication	2.93E-06
P11	programmed cell death	8.30E-06
P15	pollen wall assembly	4.73E-05

Figure 6. The functional enrichment for network modules in the core-network (**a**), and in pan-network (**b**). The most over-represented biological processes were shown for each module in the core-network while the biological processes that only enriched in pan-network modules were listed. See Supplemental File 4 for details.

using a scalable algorithm[51] implemented in Gephi 0.9 graph visualization software package[52]. In fact, the modules from the core-network and pan-network were identified using the same method to allow for an apples-to-apples comparison. We found two (out of six) modules in the pan-network enriched in 'regulation of cell cycle' ($p\text{-}value = 5.39 \times 10^{-15}$) and 'regulation of cell communication' (2.93×10^{-6}), respectively (Fig. 6). Interestingly, those biological processes were not detected among core-network modules ($p\text{-}value > 0.01$, Supplemental File 4).

The most connected nodes, i.e. hubs, are generally the focus of the network analysis. For instance, hubs in protein-protein interaction networks are more likely to be essential genes[53,54]. Hubs in coexpression networks are often considered to be the most informative genes[13,55,56]. Approximate power-law degree distribution observed in our analysis confirms the existence of hubs in both pan-network and core-network (Fig. 5). Most of the hubs in core-network are ribosomal genes, while the hubs in the pan-network represent a broad spectrum of functional categories, such as aminotransferase (AT3G49680), or ferredoxin (AT1G10960) (Table 1, Supplemental File 2). Genes involved in 'response to abiotic stimulus' were enriched among the top 200 most connected genes in the pan-network ($p\text{-}value = 1.6 \times 10^{-10}$). In addition, chaperonin genes were also enriched ($p\text{-}value < 0.01$). Chaperonins play a critical role in helping plants fight against environmental stresses by reestablishing the normal conformation of proteins. This may explain their potential ability to interact with many different genes under different conditions[57]. For instance, AT1G55490, encoding a subunit of chloroplasts chaperonins, was coexpressed with 1605 genes in the pan-network. These instances of coexpression were from 49 different experiments in total (Supplemental file 5). In conclusion, we demonstrated that a broad spectrum of condition-specific biological processes can be revealed by the pan-network analysis.

Discussion

Networks of interactions between different genes are key to our understanding of cellular mechanisms. Large-scale network data for plants are still very limited and are expensive to generate experimentally. Coexpression networks inferred from gene expression profiling data allows one to study interactions between genes (or proteins they encode) albeit indirectly. Based on 'guilt-by-association', coexpression network analysis provides great power in providing clues for gene functions in plants[58]. It also suggests candidates for the design of both high-throughput[59,60] as well as more focused low-throughput experiments[3,4,7,8]. Thousands of expression profiling experiments are available for Arabidopsis in the GEO[50]. In our previous study we reported a principle component analysis of ~7000 expression samples from more than 300 publications in Arabidopsis[61]. The developmental stage, growth conditions and the tissues of the mRNA samples used in each study are highly variable, since each of these experiments was designed to answer a specific biological question[50]. In this study, we assume the coexpression network inferred from a given experiment represents a part of the overall gene regulatory network perturbed by these specific changes in environmental or intrinsic conditions. We found relatively small number of edges (core network) that can be repeatedly detected in multiple conditions. Differences between functional enrichment within modules in pan- and core-networks emphasize the importance of biological context in coexpression analysis.

Recent studies have shown that coexpression networks in animals are highly condition-specific[28–30]. For example, by comparing mice of different ages, Southworth *et al.* found many coexpression gene modules cannot be detected in older mice[28]. Network rewiring was first revealed through comparisons between different cellular states, such as healthy and cancerous tissues[21]. Our study showed plant coexpression networks are also under dramatic changes under different conditions. To complement our pan/core-network analysis focused

	Degree	Gene description
Top 10 hubs in pan-network		
AT1G33040	1524	nascent polypeptide-associated complex subunit alpha-like
AT2G37660	1527	NAD(P)-binding Rossmann-fold superfamily protein
AT1G15820	1529	light harvesting complex photosystem II subunit 6
AT3G49680	1549	branched-chain aminotransferase 3
AT5G46110	1551	Glucose-6-phosphate/phosphate translocator-related
AT1G55490	1605	chaperonin 60 beta
AT1G74470	1607	Pyridine nucleotide-disulphide oxidoreductase family protein
AT3G56940	1615	dicarboxylate diiron protein, putative (Crd1)
AT4G01150	1642	located in thylakoid
AT1G10960	1648	ferredoxin 1
Top 10 hubs in core-network		
AT2G46820	86	photosystem I P subunit
AT3G60770	86	Ribosomal protein S13/S15
AT4G03280	86	photosynthetic electron transfer C
AT4G21280	86	photosystem II subunit QA
AT2G36170	88	Ubiquitin supergroup; Ribosomal protein L40e
AT1G42970	89	glyceraldehyde-3-phosphate dehydrogenase B subunit
AT5G16130	90	Ribosomal protein S7e family protein
AT1G26880	98	Ribosomal protein L34e superfamily protein
AT1G54780	102	thylakoid lumen 18.3 kDa protein
AT3G23390	103	Zinc-binding ribosomal protein family protein

Table 1. The hubs in pan-network and core-network[a]. [a]See Supplemental File 2 for a full list of nodes.

on individual edges s, we further calculated the preservation of co-expressed gene modules in each dataset[62]. Consistent with the results shown in the above, the most preserved modules are enriched in 'photosynthesis' ($p\text{-}value < 10^{-10}$, Supplemental File 6). The modules which can only be detected in one dataset are enriched in more specific biological processes, such as 'pollen exine formation' ($p\text{-}value < 10^{-10}$) and 'nucleosome organization' ($p\text{-}value = 3.9{*}10^{-7}$) (Supplemental File 6). In the plant community, the context-specificity of gene coexpression has been usually ignored[5,8,13,63,64]. Our search of existing literature revealed that most previous meta-analysis studies of plant expression data (except for a few notable examples discussed below) combined datasets from different labs to detect pairs of universally co-expressed genes (Supplemental File 7). Using 15 rice gene expression datasets, Childs *et al.* compared coexpression networks generated from the combined expression data against those from individual datasets. They found networks from individual datasets to contain specific but potentially informative gene modules[65]. Lee *et al.* detected coexpression relationships based on individual datasets instead of one combined dataset for Arabidopsis as well as for rice[9,66]. Although Lee *et al.* successfully predicted gene functions based on the individual networks they constructed, the differences between networks from different labs were not systemically evaluated in their study.

Our collection of 134 co-expression networks in Arabidopsis based on individual experimental series in GEO database can be used to answer the question of whether one should combine multiple expression datasets before constructing the co-expression network and if yes, which datasets can be best grouped together. In principle, by combining together similar series one can get the best of both worlds: increased statistical power to detect significantly correlated genes can be gained without losing condition-specific edges. To shed light on this problem we constructed a 134×134 matrix of similarities between our set of networks. The similarity was estimated using two different measures. The first similarity matrix shown in Fig. 7a and made available for download as Supplemental File 8 was constructed in the spirit of pan- and core-network analysis. It quantifies the fraction of edges shared between a pair of networks (See Methods). While clusters of similar networks, corresponding mostly to identical tissue types (empirical $p\text{-}value < 10^{-5}$ based on permutation tests), are visible already in this measure, the contrast between similar and different networks can be made even sharper (Fig. 7) if one uses an alternative similarity measure based on shared modules of co-expressed gene across a pair of networks detected by the WGCNA algorithm's[62] method 'modulePreservation' (Supplemental File 9). We used this similarity measure to construct the 'network of networks' connecting pairs of networks with average module similarity score above 10 (Fig. 7). Densely interconnected modules in this 'network of networks' represent good candidates for series that can be integrated without significant loss of condition-specific edges. We plan to investigate pros and cons of this approach in a follow up study.

Our analysis demonstrated that coexpression networks inferred from different microarray datasets share relatively small number of common edges, while at the same time maintaining a large number of condition-specific edges. We constructed a pan-network to represent the union of all detected co-expressed edges among 134 datasets. We also proposed the concept of core-network representing edges detected in multiple datasets. Compared to the pan-network, the core-network is more modular and enriched in genes from large multi-protein complexes. The hubs of the pan-network include genes that play a role in response to a variety of environmental

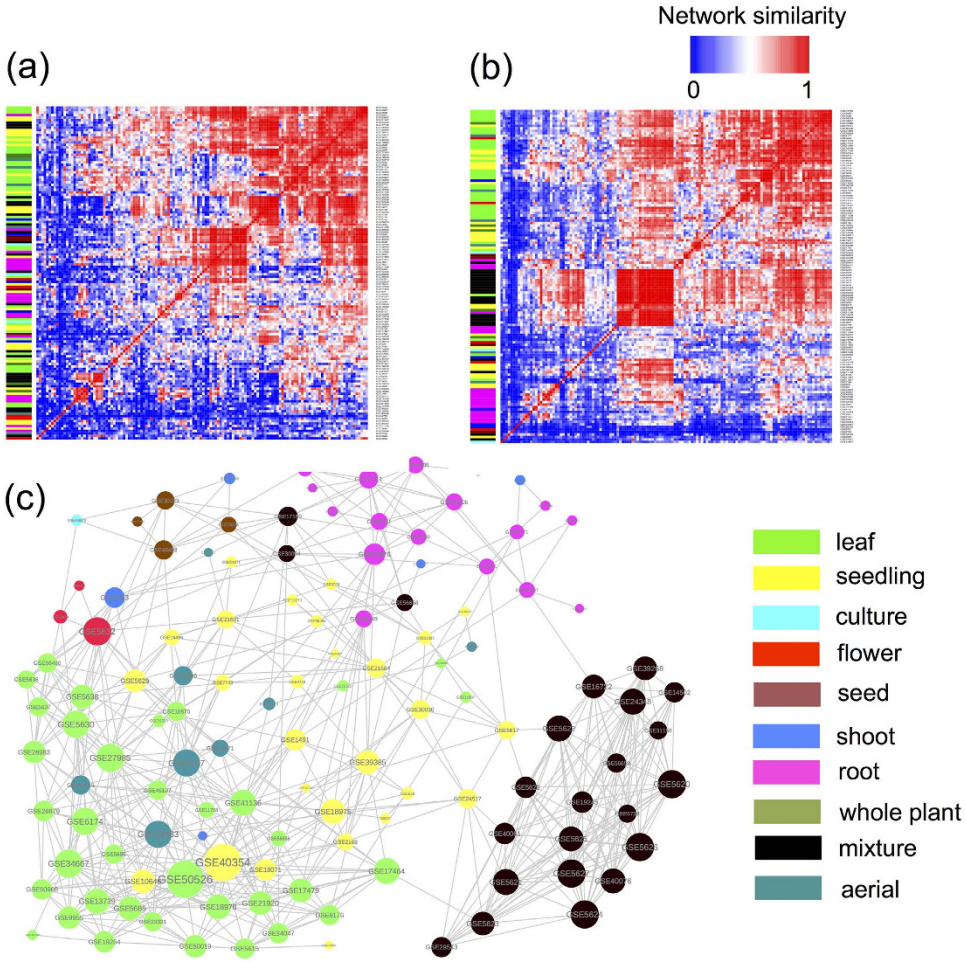


Figure 7. Microarray datasets that are generated from the same tissue type tend to form similar coexpression networks. The clustering of 134 series-based networks based on the similarity between their edges (**a**) and overlap between modules (**b**). The 'network of networks' (**c**) connecting pairs of networks with module similarity scores above 10. Network nodes (**c**) and matrix rows (**a,b**) are colored according to the tissue type with the color guide shown in the figure.

stimuli. In comparison, the modules within the pan-network are enriched in signaling and regulatory functions. We also considered several measures of similarities between individual coexpression networks and constructed 'network of networks' connecting similar networks to each other. We anticipate concepts of pan- and core- coexpression networks to provide a useful description of gene regulation architecture in a variety of species and we are currently working on extending our analysis to model organisms other than Arabidopsis.

Methods

Microarray data. All the expression datasets in this study were based on the Affymetrix platform GPL198. Only datasets containing at least 20 samples were used[67]. The CEL files of 134 expression datasets were downloaded from GEO (see Supplemental File 10) and normalized by MAS5.0[68]. The probesets were converted into TAIR gene locus ID based on the annotation file for GPL198. We only used 21678 probesets each of which has a unique mapping on a single Arabidopsis gene.

Filtering biologically relevant genes. We manually grouped samples within each dataset into different replicate groups based on the metadata provided in the dataset. Then we applied ANOVA (Analysis of variance) to identify genes that are differentially expressed between replicate groups ($p\text{-}value < 0.01$). If there were fewer than 3000 differentially expressed genes, the top 3000 genes ranked by *p-value* were used. We then applied the following standards to exclude genes with low expression. 1) For a gene to be included, at least one of its expression abundance values in a dataset was identified as expressed (i.e. present) by MAS5.0; and 2) For a gene to be included, at least one of its expression abundance values in a dataset was higher than the 90% percentile of the abundance of transposable element on the same array[69]. Those biologically relevant genes were then used to build series-based coexpression networks.

Building the series-based network for each GEO dataset. For each GEO dataset, we calculated the Pearson Correlation Coefficient (PCC) between the expression profiles of two genes. All possible pairs were

calculated. Then, the top 0.1% pairs with the highest correlations were used to build the network for each dataset (i.e. $p\text{-}value < 0.001$)[34]. A recent study showed that genes within the same metabolic pathway tend to have positive correlations[70]. In addition, taking into account negatively correlated gene pairs complicates the interpretation of our results. Therefore, we focused our analysis only on positively correlated gene pairs.

Enrichment test. The GO annotation data were downloaded from GeneOntology web site (http://geneontology.org/gene-associations/gene_association.tair.gz, July 18, 2014). The annotations inferred from the expression profile (i.e. IEP) were removed to avoid the possibility of 'self-fulfilling prophecy'. All the daughter nodes were recursively mapped to the mother node based on 'is_a' relationship. Only the GO terms that are not broadly associated with too many genes were used according to the Bonferroni correction:

$$\left(\frac{\#\text{ of genes associated with the GO term}}{\#\text{ of genes in the genome}}\right)^2 < \frac{0.05}{\#\text{ of GO terms in the genome}} \quad (1)$$

The Fisher's exact test was utilized to calculate the significance of the enrichment for each GO term, followed by Benjamini–Hochberg correction.

Calculation of network similarity between different datasets. We first used the fraction of overlapped edges between two networks to measure their similarity (Supplemental File 8) which was based on Jaccard index,

$$\frac{\text{Ei} \cap \text{Ej}}{\text{Ei} \cup \text{Ej}} \quad (2)$$

where *Ei* and *Ej* are the edges in network *i* and *j*, respectively. Another measure of network similarity was calculated as follows (Supplemental File 9). First, densely interconnected network modules were detected by Weighted Gene Co-Expression Network Analysis (WGCNA) software for each one of our 134 networks[71]. Second, the method, 'modulePreservation' within WGCNA was utilized to calculate the preservation of each module in another dataset[62]. Then the average Z-summary score of all the modules shared between a pair of networks was used to represent their similarity. This score was normalized between 0 and 1 for visualization purpose by,

$$1 - \frac{max - S_{i,j}}{max - min} \quad (3)$$

where $S_{i,j}$ is the similarity (i.e average Z-summary) between a pair of datasets. *max* and *min* represent the largest and smallest similarity, respectively. A cutoff of $S_{i,j} > 10$ was applied in order to keep the network pairs with the strongest similarity when be visualized[62]. If a network has more than 10 neighbors, only the first 10 neighbors were shown. If a network has no neighbor, it was not shown. For more information on the calculation of network similarity using WGCNA, see Supplemental File 11.

References

1. Kim, S. K. *et al.* A gene expression map for Caenorhabditis elegans. *Science* **293,** 2087–2092 (2001).
2. Stuart, J. M., Segal, E., Koller, D. & Kim, S. K. A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302,** 249–255 (2003).
3. Yonekura-Sakakibara, K., Tohge, T., Niida, R. & Saito, K. Identification of a flavonol 7-O-rhamnosyltransferase gene determining flavonoid pattern in Arabidopsis by transcriptome coexpression analysis and reverse genetics. *J. Biol. Chem.* **282,** 14932–14941 (2007).
4. Mentzen, W. I., Peng, J., Ransom, N., Nikolau, B. J. & Wurtele, E. S. Articulation of three core metabolic processes in Arabidopsis: fatty acid biosynthesis, leucine catabolism and starch metabolism. *BMC Plant Biol.* **8,** 76; doi: 10.1186/1471-2229-8-76 (2008).
5. Gigolashvili, T. *et al.* The plastidic bile acid transporter 5 is required for the biosynthesis of methionine-derived glucosinolates in Arabidopsis thaliana. *Plant Cell* **21,** 1813–1829 (2009).
6. Alejandro, S. *et al.* AtABCG29 is a monolignol transporter involved in lignin biosynthesis. *Curr. Biol.* **22,** 1207–1212 (2012).
7. Vanholme, R. *et al.* Caffeoyl shikimate esterase (CSE) is an enzyme in the lignin biosynthetic pathway in Arabidopsis. *Science* **341,** 1103–6 (2013).
8. Pick, T. R. *et al.* PLGG1, a plastidic glycolate glycerate transporter, is required for photorespiration and defines a unique class of metabolite transporters. *Proc. Natl. Acad. Sci. USA* **110,** 3185–90 (2013).
9. Lee, I., Ambaru, B., Thakkar, P., Marcotte, E. M. & Rhee, S. Y. Rational association of genes with traits using a genome-scale gene network for Arabidopsis thaliana. *Nat. Biotechnol.* **28,** 149–156 (2010).
10. Hwang, S., Rhee, S. Y., Marcotte, E. M. & Lee, I. Systematic prediction of gene function in Arabidopsis thaliana using a probabilistic functional gene network. *Nat. Protoc.* **6,** 1429–1442 (2011).
11. Berardini, T. Z. *et al.* Functional Annotation of the Arabidopsis Genome Using Controlled Vocabularies1. *Plant Physiol.* **135,** 745–755 (2004).
12. Heyndrickx, K. S. & Vandepoele, K. Systematic Identification of Functional Plant Modules through the Integration of Complementary Data Sources. *Plant Physiology* **159,** 884–901 (2012).
13. Mao, L., Van Hemert, J. L., Dash, S. & Dickerson, J. A. Arabidopsis gene co-expression network and its functional modules. *BMC Bioinformatics* **10,** 346; doi: 10.1186/1471-2105-10-346 (2009).
14. Atias, O., Chor, B. & Chamovitz, D. A. Large-scale analysis of Arabidopsis transcription reveals a basal co-regulation network. *BMC Syst. Biol.* **3,** 86; doi: 10.1186/1752-0509-3-86 (2009).
15. Wang, S. *et al.* Genome-scale identification of cell-wall related genes in Arabidopsis based on co-expression network analysis. *BMC Plant Biology* **12,** 138; doi: 10.1186/1471-2229-12-138 (2012).
16. Lee, H. K., Hsu, A. K., Sajdak, J., Qin, J. & Pavlidis, P. Coexpression analysis of human genes across many microarray data sets. *Genome Res.* **14,** 1085–1094 (2004).
17. Weirauch, M. T. In *Applied Statistics for Network Biology* 215–250 (Wiley-VCH Verlag GmbH & Co. KGaA, 2011), doi: 10.1002/9783527638079.ch11

18. Ramasamy, A., Mondry, A., Holmes, C. C. & Altman, D. G. Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS Medicine* **5,** 1320–1332 (2008).
19. Fare, T. L. *et al.* Effects of atmospheric ozone on microarray data quality. *Anal. Chem.* **75,** 4672–4675 (2003).
20. Chen, C. *et al.* Removing batch effects in analysis of expression microarray data: An evaluation of six batch adjustment methods. *PLoS One* **6,** e17238 (2011).
21. De la Fuente, A. From 'differential expression' to 'differential networking' - identification of dysfunctional regulatory networks in diseases. *Trends Genet.* **26,** 326–33 (2010).
22. Roguev, A. *et al.* Conservation and rewiring of functional modules revealed by an epistasis map in fission yeast. *Science* **322,** 405–410 (2008).
23. Bandyopadhyay, S. *et al.* Rewiring of genetic networks in response to DNA damage. *Science* **330,** 1385–9 (2010).
24. Choi, J. K., Yu, U., Yoo, O. J. & Kim, S. Differential coexpression analysis using microarray data and its application to human cancer. *Bioinformatics* **21,** 4348–55 (2005).
25. Ideker, T. & Krogan, N. J. Differential network biology. *Mol. Syst. Biol.* **8,** 565; doi: 10.1038/msb.2011.99 (2012).
26. Amar, D., Safer, H. & Shamir, R. Dissection of regulatory networks that are altered in disease via differential co-expression. *PLoS Comput. Biol.* **9,** e1002955 (2013).
27. Guénolé, A. *et al.* Dissection of DNA Damage Responses Using Multiconditional Genetic Interaction Maps. *Mol. Cell* **49,** 346–358 (2013).
28. Southworth, L. K., Owen, A. B. & Kim, S. K. Aging mice show a decreasing correlation of gene expression within genetic modules. *PLoS Genet.* **5,** e1000776 (2009).
29. Hudson, N. J., Reverter, A. & Dalrymple, B. P. A differential wiring analysis of expression data correctly identifies the gene containing the causal mutation. *PLoS Comput. Biol.* **5,** e1000382 (2009).
30. Anglani, R. *et al.* Loss of connectivity in cancer co-expression networks. *PLoS One* **9,** e87075 (2014).
31. Charng, Y.-Y. *et al.* A heat-inducible transcription factor, HsfA2, is required for extension of acquired thermotolerance in Arabidopsis. *Plant Physiol.* **143,** 251–262 (2007).
32. Okushima, Y., Mitina, I., Quach, H. L. & Theologis, A. AUXIN RESPONSE FACTOR 2 (ARF2): A pleiotropic developmental regulator. *Plant J.* **43,** 29–46 (2005).
33. Usadel, B. *et al.* Co-expression tools for plant biology: Opportunities for hypothesis generation and caveats. *Plant, Cell and Environment* **32,** 1633–1651 (2009).
34. Bergmann, S., Ihmels, J. & Barkai, N. Similarities and differences in genome-wide expression data of six organisms. *PLoS Biol.* **2,** E9 (2004).
35. Lapierre, P. & Gogarten, J. P. Estimating the size of the bacterial pan-genome. *Trends in Genetics* **25,** 107–110 (2009).
36. Medini, D., Donati, C., Tettelin, H., Masignani, V. & Rappuoli, R. The microbial pan-genome. *Current Opinion in Genetics and Development* **15,** 589–594 (2005).
37. Hansey, C. N. *et al.* Maize (zea mays L.) genome diversity as revealed by rna-sequencing. *PLoS One* **7,** e33071 (2012).
38. Hirsch, C. N. *et al.* Insights into the maize pan-genome and pan-transcriptome. *Plant Cell* **26,** 121–35 (2014).
39. Golicz, A. A., Batley, J. & Edwards, D. Towards plant pangenomics. *Plant Biotechnology Journal*. **14,** 1099–105; doi: 10.1111/pbi.12499 (2015).
40. Dixit, P. D., Pang, T. Y., Studier, F. W. & Maslov, S. Recombinant transfer in the basic genome of *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* **112,** 9070–9075 (2015).
41. Arabidopsis Interactome Mapping Consortium. Evidence for network evolution in an Arabidopsis interactome map. *Science* **333,** 601–7 (2011).
42. Tintor, N. *et al.* Layered pattern receptor signaling via ethylene and endogenous elicitor peptides during Arabidopsis immunity to bacterial infection. *Proc. Natl. Acad. Sci. USA* **110,** 6211–6 (2013).
43. Schulman, B. A. *et al.* Insights into SCF ubiquitin ligases from the structure of the Skp1-Skp2 complex. *Nature* **408,** 381–386 (2000).
44. Kuroda, H. *et al.* Classification and expression analysis of Arabidopsis F-box-containing protein genes. *Plant Cell Physiol.* **43,** 1073–1085 (2002).
45. Skaar, J. R., Pagan, J. K. & Pagano, M. Mechanisms and function of substrate recruitment by F-box proteins. *Nat. Rev. Mol. Cell Biol.* **14,** 369–81 (2013).
46. Albert, R. Scale-free networks in cell biology. *J. Cell Sci.* **118,** 4947–4957 (2005).
47. Barabási, A.-L. & Albert, R. Emergence of scaling in random networks. *Science* (*80-.*). **286,** 11 (1999).
48. Barabási, A.-L. & Bonabeau, E. Scale-free networks. *Sci. Am.* **288,** 60–69 (2003).
49. Chatr-Aryamontri, A. *et al.* The BioGRID interaction database: 2015 update. *Nucleic Acids Res*. **43,** D470–D478 (2015).
50. Barrett, T. *et al.* NCBI GEO: Archive for functional genomics data sets - Update. *Nucleic Acids Res*. **41,** D991–5, doi: 10.1093/nar/gks1193 (2013).
51. Blondel, V. D., Guillaume, J., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech.* P10008 (2008).
52. Bastian, M., Heymann, S. & Jacomy, M. Gephi: An Open Source Software for Exploring and Manipulating Networks. *Third Int. AAAI Conf. Weblogs Soc. Media* 361–362, doi: 10.1136/qshc.2004.010033 (2009).
53. Jeong, H., Mason, S. P., Barabási, a. L. & Oltvai, Z. N. Lethality and centrality in protein networks. *Nature* **411,** 41–42 (2001).
54. Zotenko, E., Mestre, J., O'Leary, D. P. & Przytycka, T. M. Why do hubs in the yeast protein interaction network tend to be essential: Reexamining the connection between the network topology and essentiality. *PLoS Comput. Biol.* **4,** e1000140 (2008).
55. Horvath, S. & Dong, J. Geometric interpretation of gene coexpression network analysis. *PLoS Comput. Biol.* **4,** e1000117 (2008).
56. Azuaje, F. J. Selecting biologically informative genes in co-expression networks with a centrality score. *Biol. Direct* **9,** 12; doi: 10.1186/1745-6150-9-12 (2014).
57. Wang, W., Vinocur, B., Shoseyov, O. & Altman, A. Role of plant heat-shock proteins and molecular chaperones in the abiotic stress response. *Trends in Plant Science* **9,** 244–252 (2004).
58. He, F., Karve, A. A., Maslov, S. & Babst, B. A. Large-Scale Public Transcriptomic Data Mining Reveals a Tight Connection between the Transport of Nitrogen and Other Transport Processes in Arabidopsis. *Front. Plant Sci*. **7,** 1207, doi: 10.3389/fpls.2016.01207 (2016).
59. Chae, L., Lee, I., Shin, J. & Rhee, S. Y. Towards understanding how molecular networks evolve in plants. *Curr. Opin. Plant Biol*. **15,** 177–184 (2012).
60. Jiménez-Gómez, J. M. Network types and their application in natural variation studies in plants. *Curr. Opin. Plant Biol*. **18,** 80–86 (2014).
61. He, F. *et al.* Large-scale atlas of microarray data reveals the distinct expression landscape of different tissues in Arabidopsis. *Plant J.* **86,** 472–80 (2016).
62. Langfelder, P., Luo, R., Oldham, M. C. & Horvath, S. Is my network module preserved and reproducible? *PLoS Comput. Biol.* **7,** e1001057 (2011).
63. Mutwil, M. *et al.* PlaNet: combined sequence and expression comparisons across plant networks derived from seven species. *Plant Cell* **23,** 895–910 (2011).
64. Gu, Y. *et al.* Identification of a cellulose synthase-associated protein required for cellulose biosynthesis. *Proc. Natl. Acad. Sci. USA* **107,** 12866–71 (2010).

65. Childs, K. L., Davidson, R. M. & Buell, C. R. Gene coexpression network analysis as a source of functional annotation for rice genes. *PLoS One* **6,** e22196 (2011).
66. Lee, I. *et al.* Genetic dissection of the biotic stress response using a genome-scale gene network for rice. *Proc. Natl. Acad. Sci. USA* **108,** 18548–53 (2011).
67. Li, W. *et al.* Integrative analysis of many weighted Co-Expression networks using tensor computation. *PLoS Comput. Biol.* **7,** e1001106 (2011).
68. Bolstad, B. M., Irizarry, R. A., Åstrand, M. & Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19,** 185–193 (2003).
69. Rodgers-Melnick, E. *et al.* Contrasting patterns of evolution following whole genome versus tandem duplication events in Populus. *Genome Res.* **22,** 95–105 (2012).
70. Chae, L., Kim, T., Nilo-Poyanco, R. & Rhee, S. Y. Genomic signatures of specialized metabolism in plants. *Science* **344,** 510–3 (2014).
71. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9,** 559, doi: 10.1186/1471-2105-9-559 (2008).

Acknowledgements

We thank Shinjae Yoo, Daifeng Wang, Mark Gerstein, Sunita Kumari and Doreen Ware for helpful discussions. We also appreciate editing provided by Claudia Lutz from the University of Illinois at Urbana-Champaign. This work was supported by grants PM-031 from the Office of Biological Research of the U.S. Department of Energy.

Author Contributions

S.M. and F.H. conceived the study. F.H. performed analysis. S.M. and F.H. wrote the paper.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: He, F. and Maslov, S. Pan- and core- network analysis of co-expression genes in a model plant. *Sci. Rep.* **6**, 38956; doi: 10.1038/srep38956 (2016).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.