Gene-Sharing Networks Reveal Organizing Principles of Transcriptomes in Arabidopsis and Other Multicellular Organisms

Song Li, Sona Pandey, Timothy E. Gookin, Zhixin Zhao, Liza Wilson, and Sarah M. Assmann

Biology Department, Pennsylvania State University, University Park, Pennsylvania 16802

Understanding tissue-related gene expression patterns can provide important insights into gene, tissue, and organ function. Transcriptome analyses often have focused on housekeeping or tissue-specific genes or on gene coexpression. However, by analyzing thousands of single-gene expression distributions in multiple tissues of Arabidopsis thaliana, rice (Oryza sativa), human (Homo sapiens), and mouse (Mus musculus), we found that these organisms primarily operate by gene sharing, a phenomenon where, in each organism, most genes exhibit a high expression level in a few key tissues. We designed an analytical pipeline to characterize this phenomenon and then derived Arabidopsis and human gene-sharing networks, in which tissues are connected solely based on the extent of shared preferentially expressed genes. The results show that tissues or cell types from the same organ system tend to group together to form network modules. Tissues that are in consecutive developmental stages or have common physiological functions are connected in these networks, revealing the importance of shared preferentially expressed genes in conferring specialized functions of each tissue type. The networks provide predictive power for each tissue type regarding gene functions of both known and heretofore unknown genes, as shown by the identification of four new genes with functions in guard cell and abscisic acid response. We provide a Web interface that enables, based on the extent of gene sharing, both prediction of tissue-related functions for any Arabidopsis gene of interest and predictions concerning the relatedness of tissues. Common gene-sharing patterns observed in the four model organisms suggest that gene sharing evolved as a fundamental organizing principle of gene expression in diverse multicellular eukaryotes.

INTRODUCTION

One intriguing feature of multicellular organisms is that multiple cell types harboring the same genomic constituents perform drastically different functions. A widely accepted model to explain the functional diversity of different tissues is that gene expression is selectively regulated in different tissues (Rokas, 2008). Two common ways of classifying genes are as tissue-specific genes or as housekeeping genes. A tissue-specific gene expresses in one particular tissue or cell type, whereas a housekeeping gene expresses in all tissues. Tissue-specific genes are believed to contribute mainly to the structural and functional diversification of different tissue types; thus, the functions of tissue-specific genes are often experimentally tested in the tissue in which the gene is expressed (Endo et al., 2008; Deveshwar et al., 2011; Xie et al., 2011).

While the dichotomy of tissue-specific and housekeeping is useful, many genes are neither specifically expressed in one tissue nor ubiquitously expressed in all tissues. For example, in an expression map of 15 cell types in the Arabidopsis thaliana root, 51 dominant expression patterns were found, and many of these patterns consisted of genes that are expressed in more than one tissue (Brady et al., 2007). In mammalian systems, a recent study has revealed a few hundred genes that are repressed in one particular tissue but not in others (Thorrez et al., 2011). Oftentimes, knocking out a particular gene interrupts the physiological functions of a few tissues but not those of other tissues, suggesting that the target gene was selectively expressed in a few tissues where the gene’s function was required. We specifically define the phenomenon that one gene functions in two different tissues as a gene-sharing event between the two tissues. We further restrict our analysis of gene sharing to genes that function in two or more but not all tissues.

With the help of high-throughput gene expression analysis, gene expression atlases have become available for dozens of tissues and cell types in many multicellular organisms, including the plant species Arabidopsis (Birnbaum et al., 2003; Schmid et al., 2005; Brady et al., 2007) and rice (Oryza sativa), Jiao et al.,
2009), and mammalian species, such as mouse (Mus musculus) and human (Homo sapiens; Su et al., 2004). These multispecies tissue expression profiles provide us with an opportunity to characterize gene-sharing events in a systematic fashion in both closely related and evolutionarily distant organisms. Most of the published high-throughput data sets from multicellular organisms contain gene expression data from organs, tissues, and cell types. For simplicity, we will usually refer to these sample sources as tissues hereafter, following the convention of Schmid et al. (2005) and Su et al. (2004). In plants and other multicellular organisms, the most common approaches for characterizing gene expression patterns have been coexpression network analysis and gene expression clustering (Usadel et al., 2009), which can be seen as a reduced representation of a coexpression network. Many computational tools (Toufighi et al., 2005; Hruz et al., 2008; Mutwil et al., 2008; Srinivasasainagendra et al., 2008; Pop et al., 2010; Obayashi et al., 2011) have been developed for mining gene coexpression data from different tissues for several plant species. Valuable predictions of gene functions have been made by coexpression analyses based on the guilt-by-association principle (Lee et al., 2010; Mutwil et al., 2011), in which annotations from genes of known function are passed to genes of unknown function that share similar expression patterns. However, gene coexpression analysis, which relies on correlations in gene expression levels, cannot readily be adopted to analyze gene sharing. For example, for a pair of genes that are expressed and function together in just two tissues, the two genes may not be coexpressed across other tissues and thus are unlikely to be identified in coexpression analyses. Accordingly, new analytical approaches are needed to systematically characterize gene-sharing events.

To address the critical but yet poorly characterized phenomenon of gene sharing, we developed an analytical pipeline specifically designed to identify gene-sharing events and to analyze the distribution of these events across tissues. Genes that are of interest here are those shared by some but not all tissues. In our pipeline, housekeeping genes are filtered out by the distributional features of their expression across multiple tissues and cell types (Figures 1A and 1B). Each gene that has passed this first filter is then associated with the tissues in which

Figure 1. Pipeline for Constructing GSNs.

(A) For each gene, expression levels are normalized to the same scale. Gene 1 is highly expressed in only one tissue type. Gene 2 is highly expressed in two tissue types. Gene 3 shows no obvious tissue association.

(B) Genes are filtered based on the property of their expression distributions. The density of expression levels of gene 1 has one large peak and one small peak, while gene 2 has three peaks and gene 3 has only one major peak. We chose to filter genes by expression kurtosis, a metric that can distinguish genes 1 and 2 from gene 3.

(C) For each gene that passes the kurtosis filter, gene-tissue associations are summarized in a binary matrix, in which 1 indicates preferential expression in tissue T.

(D) In the GSN, each tissue is represented by a node. The size of each node is proportional to the number of genes associated with that node. Pairs of nodes in the networks are connected by an edge if they share a significant number of commonly associated genes. Edge thickness is proportional to the tissue similarity score. For example, gene 2 (G2) and gene 5 (G5) are shared between tissue 2 (T2) and tissue 4 (T4). Gene 5 (G5) is shared between tissue 1 (T1) and tissue 2 (T2). The edge weight (as indicated by the width of the edges) between T1 and T2 is smaller than the edge weight between T2 and T4 because more genes are shared between T2 and T4 than between T1 and T2.

(E) Each module in a GSN is computationally defined and consists of a group of nodes that share more connections within the group than outside the group. In this example, T1, T2, and T4 comprise a module.
the gene is highly expressed (Figure 1C). The distribution of gene-sharing events between tissues is represented by a gene-sharing network (GSN), in which nodes represent tissue samples, and the size of each node is proportional to the number of genes associated with that node (Figure 1D). A pair of tissues in the networks is connected by an edge if a significant number of genes is shared by the two tissues, with edge strength determined by statistical significance of the number of shared genes. We applied modular analysis of networks (Raghavan et al., 2007) to the GSNs; this algorithm groups tissues together based on how many genes are shared between tissues (Figure 1E). We found that knockout mutants of genes that are associated with guard cells by our method display altered guard cell phenotypes, supporting that our pipeline can precisely identify novel genes that function in a specific cell type. We also showed for two genes shared between guard cells and roots that these genes have functions in both tissues, validating the notion that gene sharing between tissues provides an important new avenue to predict gene functions. Overall, our pipeline has uncovered several intriguing organizing principles for gene sharing in multicellular organisms across species and kingdoms.

RESULTS

Data Resource

Our pipeline was tested on four multicellular organisms for which the most comprehensive tissue expression data are available: two plant species (Arabidopsis and rice) and two animal species (human and mouse). For the Arabidopsis data set, we compiled and appropriately normalized gene expression data from 80 samples from three data sets: an Arabidopsis tissue expression atlas (Schmid et al., 2005), a cell type–specific expression atlas of Arabidopsis roots (Birnbaum et al., 2003; Brady et al., 2007), and a cell type–specific expression data set of Arabidopsis guard cells (Pandey et al., 2010). All three data sets were generated by ATH1 microarray with Columbia (Col) wild-type Arabidopsis without designated environmental stimuli or hormonal perturbation (see Methods; see Supplemental Figure 1, Supplemental Methods 1, and Supplemental References 1 online). We also analyzed gene expression data obtained by laser capture microdissection from rice tissues and cell types under normal physiological conditions (Jiao et al., 2009). For human and mouse, we used cell type–specific and organ gene expression data from the Novartis Research Foundation Gene Expression Database (GNF Atlas) (Su et al., 2004). Each of the rice, human, and mouse data sets was generated by an individual group, such that no further normalization was required.

Details of the Pipeline

The pipeline consists of four steps: (1) filtering out genes that are expressed across all tissues (housekeeping genes); (2) identifying gene-tissue associations by analyzing the distributional features of gene expression across multiple tissues and cell types; (3) identifying gene-sharing events and constructing GSNs, which represent how genes are shared between tissues; and (4) identifying modules in the GSNs.

In step 1, we applied a kurtosis filter to remove genes that are expressed across all tissues. Although many existing analytical approaches can identify gene activity patterns, detecting gene-sharing events faces unique challenges. Approaches that only detect single tissue-specific genes (Greller and Tobin, 1999) are not appropriate for detecting gene sharing, since single tissue-specific genes are not shared by any other tissue. An alternative approach (Liang et al., 2006), which restricts the maximum number of tissues with which a gene can be associated, is not appropriate since the restricted approach may underestimate the number of tissues with which a gene is associated. We chose an approach that measures the shape of a gene’s expression distribution, such that the number of tissues that a gene is associated with is not restricted. In a nutshell, if a gene is preferentially expressed (or repressed) in a small number of tissues, then the expression distribution of that gene will have a sharp peak with heavy tails on one or both sides of the peak. Such a distribution is called a high kurtosis or leptokurtic distribution and can be captured by kurtosis analysis (see Methods; see Supplemental Figure 2 online). The choice of kurtosis is also motivated by recent advances where kurtosis is found to be a superior metric over other conventional methods in detecting disease marker genes (Teschendorff et al., 2006; Hellwig et al., 2010).

In step 2, we identified gene-tissue associations. For all genes passing the kurtosis filter, we next asked in which tissue is each gene preferentially expressed. For each gene, tissue preferential expression is determined by a relative expression threshold Z such that the gene is associated with a tissue if the gene is expressed in that tissue at a level higher than Z. We determined the appropriate Z threshold based on an approach that measures the reduction of kurtosis for different Z threshold (see Methods; see Supplemental Figures 3, 4, and 5 online). At the end of this step, a gene is associated with a tissue if the gene’s expression is leptokurtic (Q < 0.005) and the relative expression of that gene is larger than three (Z ≥ 3) in that tissue type.

In step 3, gene-tissue associations identified by the previous step were used to define gene-sharing events and to construct GSNs (see Methods; Figures 1D and 1E). In the GSNs, each node represents a different tissue, and the size of each node is proportional to the number of leptokurtic genes associated with that node. Pairs of nodes in the networks are connected by an edge if they share a significant number of commonly associated genes, with significance defined by tissue (node) similarity scores (S scores) (see Methods; Figure 1D). Although a single gene-sharing event is computationally independent of any other gene-sharing event between the same two tissues, the converse is not true. If two tissues share significantly more genes than what would be expected by random, this would imply that the two tissues tend to have overlapping functions. To characterize the distribution of gene-sharing events across tissues, a weighted label propagating algorithm is used to identify modules in the GSNs as the final step of the pipeline (see Methods; Figure 1E).

Results from the Pipeline

In the first step of our pipeline, we examined for each species the distribution (histogram) of each single gene’s expression
have high classical coexpression (PCC) but are not coassociated by our method. Dashed lines in both plots indicate a relative expression threshold, as used in our analysis. Gene A is grouped with gene B but not with gene C by our approach, whereas gene A and gene C, but not genes A and B, are grouped together by classical coexpression analysis.

pattern across all samples, and we calculated expression kurtosis (see Methods) for every single gene across all samples, ranked genes accordingly (Figure 2) and applied the Anscombe test (Anscombe and Glynn, 1983) to identify genes with high kurtosis (we used a stringent threshold: Q < 0.005, where Q represents the false discovery rate [FDR]-adjusted P value; see Methods). We found that 62.1% of Arabidopsis genes and 39.6% of rice genes have leptokurtic distributions, and for mammalian systems, 78.8% of human genes and 79.9% of mouse genes have leptokurtic distributions. These results were confirmed by simulated background distribution (Figure 2A, dashed curves; see Methods). The smaller percentage of rice genes identified as leptokurtic compared with the other three species is likely due the fact that the rice data set contains fewer tissues than the other species.

In addition, out of the three theoretically possible leptokurtic patterns (heavy tail on the right, on the left, or on both sides of the sharp peak), we found that, in Arabidopsis, 94.6% of leptokurtic genes have a specific pattern consisting of a sharp peak with a rightward heavy tail (see Figure 1B, genes 1 and 2). By contrast, only 5.2% of leptokurtic genes have a heavy tail on left. We found a similar leptokurtic pattern of gene expression in all other organisms analyzed (Figure 2A, rice, human, and mouse): 99.4% of rice leptokurtic genes, 95.3% of human genes, and 97.3% of mouse genes have rightward heavy tails. These conserved high kurtosis and right-skewed distributions in all four organisms suggest a conserved active-when-necessary mechanism in the regulation of gene expression (see Discussion).

We also tested the commonly used analysis of variance approach (Brady et al., 2007; Krouk et al., 2009) for identifying gene-tissue associations. We found that analysis of variance analysis generates less specific results than the kurtosis based approach in identifying genes that are only preferentially expressed in a small number of tissues. As discussed in the Supplemental Methods 1 online (Section 2), rank products, another common method of combining gene expression data is not applicable in this case.

**Shared Genes Are More Common than Single Tissue-Specific Genes**

Since genes that passed the kurtosis filter include both single tissue-specific genes and genes that are preferentially expressed in a few tissues, one could hypothesize that most genes are single tissue-specific genes. We asked whether genes that passed our kurtosis filter tend to be single tissue-specific genes and found that this was not the case: In all four organisms, single tissue-specific genes are in the great minority among genes passing the kurtosis filter in each organism (Figure 2B; see Supplemental Figure 6 online), suggesting that shared genes are much more common than single tissue-specific genes.

---

**Figure 2.** Properties of Gene Expression Kurtosis in Four Organisms.

(A) Cumulative distribution of gene expression kurtosis for all genes in two plant species (Arabidopsis and rice) and two animal species (human and mouse). Solid curve: Cumulative distribution of expression kurtosis. The y axis shows the percentage of genes with kurtosis above the kurtosis values indicated on the x axis. Dashed curve: Cumulative distribution of expression kurtosis simulated from random backgrounds, which were generated from normal distributions using parameters (sample number, median, and MAD) obtained from the original gene expression data for all genes (see Methods). In each panel, the vertical line indicates the kurtosis threshold corresponding to an FDR threshold (Q) of 0.005.

(B) In all four organisms, the majority of genes with Q < 0.005 show expression in more than one tissue. Black regions, genes associated with only one tissue type; open regions, genes associated with more than one tissue type. The percentage of genes associated with more than one tissue type is given in each pie chart.

(C) Box plot of median coexpression levels of leptokurtic genes compared with control shows that high kurtosis genes are not highly coexpressed. HC, high coexpressed genes; HK, high kurtosis genes (Q < 1e-9). See Supplemental Methods 1 online for more information.

(D) Illustration showing why tissue coassociations are not captured by classical coexpression analysis. (a) Expression levels of gene A (open circles) and gene B (closed circles) are anticorrelated with each other in all tissues except in tissue eight. Genes A and B are coassociated with tissue eight by our method but have low classical coexpression (Pearson correlation coefficient [PCC]). (b) Gene A (open circles) and gene C (open squares), which are coexpressed in all tissues except in tissue eight, have high classical coexpression (PCC) but are not coassociated by our method. Dashed lines in both plots indicate a relative expression threshold, as used in our analysis. Gene A is grouped with gene B but not with gene C by our approach, whereas gene A and gene C, but not genes A and B, are grouped together by classical coexpression analysis.
We asked whether the shared genes identified by our method also tend to have high coexpression levels, and we found that genes grouped together by our pipeline tend to have lower coexpression levels than genes grouped by coexpression methods (Wilcoxon test, P value < 2e-16; Figure 2C; see Methods). Indeed, two genes that are coassociated by our method by virtue of their high expression only in one or a few tissues can have small coexpression levels (Pearson correlation coefficients) (Figure 2D), suggesting that gene-sharing events that are identified by our approach lead to a different way of generating gene groups compared with gene coexpression analysis (see Discussion).

**Arabidopsis GSN**

We analyzed the *Arabidopsis* GSN in detail and found that the tissue-sharing events captured by our pipeline provide insights into gene functions. The *Arabidopsis* GSN derived by our pipeline clearly captures changes of activities of genes through development, as represented by the distributions of gene-sharing events between adjacent and nonadjacent developmental stages of tissues and organs (Figures 3 and 4; see Methods; see Supplemental Figure 7 and Supplemental Data Set 1 online). For example, in the subnetwork containing seeds and siliques, or in the subnetwork containing shoot apical meristems and early flowers, the S scores are higher between tissues that are in consecutive developmental stages than between those that are not (Figures 3A and 3B). Keeping in mind that our networks are based solely on shared genes that passed the kurtosis threshold, the congruency between subnetwork structures and developmental hierarchies suggests that these tissue-preferentially shared genes play important functional roles in the nodes (tissues) with which they are associated. We tested this hypothesis further both in silico and experimentally.

To test this hypothesis in silico, we further studied gene function using Gene Ontology (GO) analysis for genes that appear in the nodes of the above-mentioned two subnetworks (see Methods) and found that these sets of leptokurtic genes are enriched in many GO annotations that agree well with known functions of the corresponding tissues (Figures 3C and 3D). For example, “pectinesterase” (GO:0030599) and “inhibitor of pectinesterase” activity (GO:0046910) are enriched in early silique development, reflecting cell wall restructuring during early fruit development, while “seed development” (GO:0048316), “negative regulation of seed germination” (GO:0010187), and “embryonic development ending in seed dormancy” (GO:0009793) become enriched in the later stages of seed development (Figure 3C). A second example is found in the shoot apex subnetwork (Figure 3D). Genes annotated as “response to jasmonic acid” and “salicylic acid” (GO:0009753 and GO:0009751) are enriched in the earlier stages of the shoot apex and gradually disappear in early flowers and late stage flowers. “Positive regulators of flower development” (GO:0009911) become enriched in early flower samples and are reduced in late-stage flower samples. In late-stage flower samples, “sexual reproduction function” (GO:0019953) becomes enriched (Figure 3D), consistent with the fact that reproduction occurs in the mature flower. Given that GO functions of leptokurtic genes suffice to reflect known biological functions of these developmental stages, other enriched GO groups (for example, “β-galactosidase complex” [GO:0009341] is enriched in early stage of seed development; “lactose catabolic process” [GO:0019513] is enriched in later stages of flower development) represent new discoveries of genes and gene

![Figure 3. Gene-Sharing Subnetworks Recapitulate Development.](image)
functions that likely also participate in these developmental stages. For more examples, see Supplemental Figures 8 and 9 online.

We constructed a complete GSN for all the Arabidopsis tissues (Figure 4). A higher Q (Q < 1e-09) was used to make the visualization less crowded, and the corresponding Z thresholds were determined anew using the reduction of kurtosis method (see Methods and Supplemental Figure 4 online). In the complete Arabidopsis GSN, we observed clear modularity (i.e., more connections between nodes within a group than connections between nodes in different groups). Eight modules were computationally identified in the Arabidopsis GSN by applying a weighted label propagating algorithm (see Methods). Remarkably, each mathematically derived module is seen to correspond to a group of nodes with similar organ origination. For example, in the Arabidopsis network, the two largest modules are root system and floral organs and siliques, which contain 24 nodes and 16 nodes, respectively, most of which are from organ systems contained within these structures. The two smallest modules are stem module and shoot apex module; each contains two samples from stems or shoot apical meristems, respectively. Intermodule connections are also biologically meaningful. For example, the node of senescing leaves in the rosette leaves module (module 6) is connected with the node of sepals at stage 15, which is in the module of floral organs (module 8), indicating a significant number of gene-sharing events between senescing leaves and sepals.

Sometimes a whole organ and a single cell type within that organ both appear in the network. In other cases, the same tissue from consecutive developmental stages is sampled and is included as two nodes in the network. This sample nesting and partial overlap could contribute to some connections in the GSN. Because our pipeline is designed to find genes shared between tissues, one would expect that the pipeline can identify connections between nested or overlapping tissues. We found that 27% of edges found by our pipeline are indeed between nested tissues or overlapping tissues, and the number of edges found by our pipeline is more than that found by randomly selecting edges between any pairs of tissues (Fisher exact test, P < 2 × 10^{-6}). In addition to the edges between nested and overlapping tissues, however, 73% of the connections in the Arabidopsis GSN (Figure 4; see Supplemental Data Set 1 online) are not due to tissue nesting or overlapping but nevertheless have similar or higher edge weights as the edges between nested tissues.

![Figure 4. Complete Arabidopsis GSN.](image-url)

Arabidopsis GSN, obtained using the parameters Q < 1e-09, Z = 10, and S > 20. Modules were defined computationally using a weighted label propagating algorithm. Modules: (1) roots, (2) seeds, (3) seedling green parts, (4) stems, (5) young leaves and vegetative shoot apex, (6) rosette leaves, senescing leaves, and cauline leaves, (7) shoot apex, (8) floral organs and siliques. Detailed node annotations, edge lists, and edge weights can be found in Supplemental Data Set 1 online.
tissues. We tested whether the commonly used coexpression method (Carter et al., 2004) generates a network similar to that of Figure 4 (see Supplemental Methods 1 online). We found that 75% of Arabidopsis GSN edges are different from those found in the corresponding coexpression network, supporting our conclusion (Figure 2) that GSN and coexpression networks capture different properties of the expression data (see Discussion).

**Experimental Analysis of Leptokurtic Genes Associated with Guard Cells**

To experimentally evaluate our pipeline vis-à-vis gene function identification, we focused on examining genes that are associated with plant guard cells. Guard cells are specialized cells that play pivotal roles in plant drought response by regulating the apertures of stomatal pores in response to water limitation, as signaled by the stress hormone abscisic acid (ABA).

First, we found that leptokurtic genes associated with guard cells are enriched in genes with recognized guard cell functions (Table 1): 28 out of 93 genes with known functions in guard cells (Zhao et al., 2008) are present in the list of 2345 guard cell leptokurtic genes (Fisher test, P < 6.3e-8). Our analysis, based on the standardized transcriptome data set of normal Arabidopsis tissues, reveals that most of these genes are shared between guard cells and other tissues: Only three genes are guard cell specific (i.e., expressed above a basal level solely in guard cell samples), whereas the remaining 25 genes are shared between guard cells and other tissues. This result suggests that our pipeline can preferentially identify gene-sharing events. It also suggests that methods that only identify single tissue-specific genes (Greller and Tobin, 1999) will miss significant numbers of genes with important functional roles in the specific biological system of interest.

Second, the functional significance of guard cells’ leptokurtic genes was evaluated experimentally. Quantitative RT-PCR (qRT-PCR) was performed on seven different Arabidopsis tissues for 14 leptokurtic genes associated with guard cells by our in silico analysis of kurtosis. qRT-PCR showed that all 14 genes have strong preferential expression in guard cells (Figure 5), supporting the conclusion that our pipeline can correctly identify tissue preferential expression. The fact that several of the genes with relatively lower kurtosis values also have nonnegligible expression levels in other tissues in qRT-PCR analysis (Figure 5) experimentally confirms that gene-sharing events are identified by our pipeline.

Third, we pursued functional analysis for 10 of these genes (see Methods). These genes were selected solely on the basis of their leptokurtic association with guard cells, coupled with listed availability of two T-DNA insertional mutants. For six of these genes (At1g20880, At1g31335, At1g65020, At3g51760, At4g01880, and At5g60410), we could not verify two independent mutant lines lacking full-length transcripts, so we focused our efforts on the remaining four genes (Figure 6): At1g50400, a putative mitochondrial import receptor subunit TOM40 homolog 2; At2g21080, annotated as a plant-specific unknown protein; At1g11100, an SNF2 domain-containing and helical domain containing protein; and At5g60410 (SUMO E3 ligase [SIZ1]) (Mura et al., 2009). None of these four genes had been previously reported to function in guard cells.

T-DNA knockouts of all four genes resulted in altered guard cell physiology compared with the Col wild type (Figure 6). Two mutants (At2g21080 and At1g11100) showed enhanced light stimulation of stomatal opening, while one mutant (At5g60410 [siz1]) showed reduced opening compared with the wild type. One mutant showed hypersensitivity to ABA inhibition of stomatal opening (At5g60410 [siz1]), one mutant showed hypersensitivity in ABA promotion of stomatal closure (At1g50400), and one mutant was hypersensitive in both ABA responses (At1g11100). Morphologically, guard cells of mutants At2g21080 and At1g11100 were significantly longer than the wild type, while siz1 had shorter guard cells than the wild type (Figure 6).

**In Silico and Experimental Analysis of Genes Shared between Guard Cells and Other Tissues**

Because most genes identified by our pipeline are shared between tissues, we asked whether gene-sharing events between other tissues and guard cells identified by our in silico approach are biologically meaningful. We identified tissues that share a significant number of common genes with guard cells by their S scores (see Methods) and found 14 such tissues (Figure 7). While many genes that passed our pipeline are associated with...
Figure 6. Experimental Validation of Predictions from the Arabidopsis GSN.

(A) Schematic of the local genomic structure and T-DNA insertion sites of mutant Arabidopsis lines used for phenotypic analyses. The genomic structure of each gene is drawn to scale. The lengths of all gene models are normalized. Open bars indicate 5′ and 3′ untranslated regions, solid black bars indicate exons, and black lines indicate introns. T-DNA insertions are not drawn to scale.

(B) RT-PCR showing absence of full-length transcripts from the knockout mutant lines. Insertion sites and absence of full-length transcripts for these genes were confirmed by genomic PCR and RT-PCR, respectively. Gene-specific primers flanking the insertion sites were used in RT-PCR to confirm absence of full-length transcripts in the T-DNA insertional lines. Wild-type Col-0 was used as a positive control for each primer pair.

(C) and (D) Stomatal phenotypes of mutants of four guard cell leptokurtic genes. ABA inhibition of light-stimulated stomatal opening (C) and ABA induction of stomatal closure (D) (three replicates, 300 stomata, error bars represent ± s.e.). Labels above the white bars, * or ** represent significant difference from wild-type (WT) Col in light-stimulated stomatal opening. Labels above the black bars, * or ** represent significant difference from wild-type Col in response to ABA (50 μM). *P value < 0.05; **P value < 0.01.
guard cells and guard cells’ 14 first neighbor tissues (Figure 7; also see http://genesharingnetworks.org), only genes that are associated with guard cells, and not genes associated with any of the other 14 tissues, are enriched in the GO term “stomatal movement” (GO:0010118). This result indicates the specificity of our analysis.

We found that significant numbers of genes are shared between guard cells and three samples corresponding to seeds at different developmental stages (Figure 7). Both guard cells and seeds are major targets of ABA action, which regulates both stomatal movement and seed dormancy/germination. GO analysis again confirmed that shared genes represent shared functions (e.g., the term “response to ABA stimulus” [GO:0009737] is enriched in guard cells and in one of the seed samples).

We observed significant numbers of genes shared between guard cells and root epidermal cells, including root hair cells, nonhair epidermal cells, and lateral root caps (Figure 7). The fact that both guard cells and other epidermal cells play major roles in plant defense against pathogens and in regulation of water balance suggests that genes associated with both guard cells and root epidermal cells may contribute to the generic functions of epidermal tissues. We tested this hypothesis by GO analysis (see Methods) and found that genes shared between guard cells and root epidermal cells are enriched in GO category “response to chitin” (GO:0010200), a well-known elicitor for plant defense response (15), and in “hydrogen peroxide function” (GO:0042542), which is well documented as a secondary messenger in stress responses of both guard cells and roots (Kwak et al., 2003).

Genes shared between guard cells and two out of the three root epidermal cell types are also enriched in GO category “response to water deprivation” (GO:0009414). These results are consistent with the hypothesis that connections between guard cells and root epidermal cells are partly due to genes with epidermal tissue functions. The gene sharing between guard cells and non-epidermal root cell types, such as pericycle and endodermal cells, and between guard cells and senescing leaves or flower sepals (Figure 7) is unexpected based on current knowledge of the physiologies of these cell or tissue types, suggesting that our method may also yield clues regarding previously unidentified functional similarities between cells or tissues.

Because the GSN approach successfully identified genes involved in GO physiological processes known to be shared between nodes, we hypothesized that genes with functions in several nodes can be predicted from the GSN. After identifying the four novel genes with guard cell functions, we found that one of them, At5g60410 (SIZ1), is also associated with several root cell types by our method. This result is consistent with the observation that siz1 mutants are hypersensitive to ABA inhibition of root elongation (Miura et al., 2009). We also found that the kurtosis of SIZ1 is lower than that of other genes in our qRT-PCR analyses (Figure 5), consistent with known pleiotropic effects of siz1 (Miura et al., 2009). Among the mutants we characterized experimentally, At1g11100 is also associated with both guard cells and root cells in our guard cell subnetwork (Figure 7), leading us to hypothesize that At1g11100 would also have a root-related function. Indeed, we found that At1g11100 mutants are hypersensitive to ABA inhibition of root elongation (Figure 6F). Thus, both published results from siz1 mutants and our new results indicate that our method accurately captures tissue-related functions inferred for these genes from our GSN.

Comparison of Arabidopsis and Human GSNs

To evaluate whether our GSN representation is broadly applicable in capturing biologically meaningful information, we constructed and analyzed the human GSN. In the human network (see Supplemental Figure 10 online), we found further support for our conclusion that shared genes confer organ identity and relatedness; for example, connections between fetal liver and liver, and between fetal lung and lung, are stronger than connections with other tissues in the network. Using the weighted label propagating algorithm, we identified seven modules in the human GSN, and, as for Arabidopsis, each module corresponded to a group of tissue and cell types with related biological functions (see Supplemental Figure 10, Supplemental Methods 1, and Supplemental Data Set 2 online). We also found biologically intriguing connections between connected nodes in different modules (see Supplemental Methods 1 and Supplemental Figures 11 and 12 online). Finally, we found that 82% of connections in human GSN are not due to tissue nesting (see Supplemental Data Set 2 online), and 86% of human GSN connections are not found by coexpression analysis, further confirming that our pipeline provides a unique approach to identify biologically meaningful gene-sharing events.

Investigating Functions of Leptokurtic Genes in Plants and Animals

Because computational analyses of both Arabidopsis and human networks strongly support the importance of gene-sharing events in determining tissue functions in multicellular organisms,
DISCUSSION

Leptokurtic Shared Genes Confer Tissue Identity and Relatedness

Traditional analyses of gene expression patterns in multicellular organisms have focused primarily on the identification of coexpressed genes (Usadel et al., 2009) or tissue-specific genes (Greller and Tobin, 1999). Here, we show that leptokurtic genes, which are preferentially expressed in just a few tissues, are ~10 times more common than single tissue-specific genes in all four multicellular organisms analyzed: Arabidopsis, rice, human, and mouse. The number of true single tissue-specific genes (Figure 2B) may actually be even smaller, since not all cell types have been sampled by expression profiling in any of the four organisms. The observed paucity of single tissue-specific genes leads us to propose that, although some tissue-specific functions can be ascribed to single tissue-specific genes, most tissue-specific functions are actually conferred through combinations of the functions of shared tissue-preferential genes.

Leptokurtic genes exhibit an organ-based modularity in expression pattern (Figure 4), unexpectedly suggesting that organs, not individual cell types, are the major determinants of gene expression patterns. A similar idea was noted in a principal component analysis of organ expression patterns (Schmid et al., 2005). An example consistent with this conclusion is the fact that while the connections between guard cells (a type of epidermal cell) and root epidermal cells have an average S score of 75.3, which indicates nonnegligible similarities between these epidermal tissues, the connections between different epidermal cell types of the root have a much larger average S score of 330. Together, these results imply that epidermal tissues in different organs share expression of fewer genes than epidermal tissues in the same organ. Because the available microarray (ATH1) data sets for Arabidopsis include relatively few samples from single cell or tissue types of aboveground organs, additional data from such samples will be required to further address this hypothesis.

We found that edges between nesting and overlapping tissues are enriched in our GSN, supporting that our pipeline is able to identify biologically meaningful connections between tissues. Among the 73% of connections in Arabidopsis GSN that are not between two nested or between tissues in consecutive developmental stages, we found many cases where a single cell type is connected with a complex tissue or an organ that comprise mixtures of cell types. For example, we found that guard cells and seeds are connected (Figure 7). One may hypothesize that such connections occur because some genes are highly expressed in guard cells and one particular cell type in the seeds. One can also hypothesize that many genes that are highly expressed in guard cells are highly expressed in all cell types in seeds. The connections between simple cell types and complex tissues can guide further experimental validation of the shared genes (e.g., by high-resolution experimental procedures, such as in situ hybridization).

GSNs Provide Different Information Than Coexpression Analysis

We found that pairs of leptokurtic genes that are coassociated with exactly the same tissues by our analysis are not highly correlated in their expression levels when assessed by standard methods of coexpression analysis (Figures 2C and 2D). This result suggests that our approach can identify gene expression patterns that are not captured by coexpression analysis. For genes that only function in a subset of tissues, our approach,
which only considers expression levels in that subset of tissues, may provide more functionally relevant predictions for a given gene than coexpression analysis, which is based on gene expression levels across all tissues (Figure 2D).

Coexpression analysis can also be used to create tissue networks; in that case, a pair of tissues is connected if the Pearson correlation coefficient of the expression levels of all genes in the pair of tissues exceeds a predetermined threshold (Ueda et al., 2004). As a comparison to our GSN, we analyzed genes in the pair of tissues that exceed a predetermined threshold, sharing events that is used to derive the tissue similarity is also a sharp peak with a rightward heavy tail. We also provide in silico and experimental evidence that leptokurtic genes associated with a tissue are likely to have physiological function in that tissue. Together, these results suggest that most genes are only activated when the function of the gene is required by the gene-tissue. Together, these results suggest that most genes are only activated when the function of the gene is required by the gene-associated tissue (i.e., active when necessary). It is interesting that the alternative scenario, where a gene is only repressed if the gene’s absence is biologically necessary (i.e., a leftward heavy-tailed pattern), is not widely observed in any of the four organisms, suggesting that the rightward heavy-tailed pattern may represent a general phenomenon for multicellular organisms. This distribution is in some ways analogous to a just-in-time transcription program in single-celled organisms, such as Escherichia coli (Zaslaver et al., 2004), in which enzymes in a metabolic pathway are transcribed sequentially as their substrates become available. Both just-in-time transcription and the active-when-necessary principle can be regarded as mechanisms that maximize the efficiency of the transcriptional program because cellular resources are not allocated to transcribe a gene when the gene product is not necessary.

The observed active-when-necessary principle does not prescribe the underlying transcriptional or posttranscriptional mechanisms that generate the expression pattern. For example, we tested whether known microRNA targets (Backman et al., 2008) are enriched in the high kurtosis genes with heavy right tail or left tail and found no evidence of enrichment in either case (P > 0.4 for both cases; see Supplemental Methods 1 online).

Table 1. Leptokurtic Genes with Known Functional Roles in Guard Cells

<table>
<thead>
<tr>
<th>TAIR IDa</th>
<th>Gene Nameb</th>
<th>Tissuec</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT1G66340</td>
<td>ETR1</td>
<td>1</td>
</tr>
<tr>
<td>AT4G00400</td>
<td>GPA78</td>
<td>1</td>
</tr>
<tr>
<td>AT1G12860</td>
<td>SCREAM2</td>
<td>1</td>
</tr>
<tr>
<td>AT3G24140</td>
<td>FMA</td>
<td>2</td>
</tr>
<tr>
<td>AT1G62400</td>
<td>HT1</td>
<td>3</td>
</tr>
<tr>
<td>AT2G18960</td>
<td>OST2</td>
<td>3</td>
</tr>
<tr>
<td>AT1G12480</td>
<td>SLAC1</td>
<td>4</td>
</tr>
<tr>
<td>AT5G46240</td>
<td>KAT1</td>
<td>5</td>
</tr>
<tr>
<td>AT3G53720</td>
<td>ATCHX20</td>
<td>7</td>
</tr>
<tr>
<td>AT2G20875</td>
<td>EFP1</td>
<td>7</td>
</tr>
<tr>
<td>AT3G26744</td>
<td>ICE1</td>
<td>7</td>
</tr>
<tr>
<td>AT4G26080</td>
<td>ABI1</td>
<td>8</td>
</tr>
<tr>
<td>AT1G08810</td>
<td>MYB60</td>
<td>8</td>
</tr>
<tr>
<td>AT4G33950</td>
<td>OST1</td>
<td>8</td>
</tr>
<tr>
<td>AT3G11820</td>
<td>SYD121</td>
<td>8</td>
</tr>
<tr>
<td>AT4G16110</td>
<td>ARR2</td>
<td>9</td>
</tr>
<tr>
<td>AT4G17615</td>
<td>CBL1</td>
<td>10</td>
</tr>
<tr>
<td>AT4G18290</td>
<td>KAT2</td>
<td>11</td>
</tr>
<tr>
<td>AT1G78390</td>
<td>NCED9</td>
<td>12</td>
</tr>
<tr>
<td>AT5G57050</td>
<td>ABI2</td>
<td>13</td>
</tr>
<tr>
<td>AT5G37500</td>
<td>GORK</td>
<td>13</td>
</tr>
<tr>
<td>AT2G29940</td>
<td>PDR3</td>
<td>14</td>
</tr>
<tr>
<td>AT4G34000</td>
<td>ABF3</td>
<td>15</td>
</tr>
<tr>
<td>AT1G04110</td>
<td>SDO1</td>
<td>15</td>
</tr>
<tr>
<td>AT5G53210</td>
<td>SPCH</td>
<td>15</td>
</tr>
<tr>
<td>AT5G07180</td>
<td>ERL2</td>
<td>18</td>
</tr>
<tr>
<td>AT1G80080</td>
<td>TMM</td>
<td>21</td>
</tr>
<tr>
<td>AT1G22690</td>
<td>pGC1</td>
<td>34</td>
</tr>
</tbody>
</table>

aThe unique identifier for each Arabidopsis gene and is obtained from the TAIR website (www.Arabidopsis.org).
bThe short name assigned to the gene.
cThe number of tissues the gene is preferentially associated with based on our approach.

as to which genes are likely to function in the given tissues of interest.

Online Web Resource

Many genes identified by our methods currently are annotated as "unknown" or have no annotations related to the functions of the associated tissue. Our results suggest that these genes have a high likelihood to play important roles in the tissues with which they are associated and indicate to the experimentalist exactly which tissues should be scrutinized for functional analysis of such genes. To provide a facile way for experimental biologists to explore the computational results from our work, we provide an online browser at http://genesharingnetworks.org for the Arabidopsis GSN. This website enables users to click on any tissue of interest and find the list of genes that are preferentially expressed in that tissue. The user can also retrieve the subnetwork that contains all tissues that share a significant number of genes with a selected tissue of interest. Lists of genes that are shared between pairs of tissues can be retrieved from the networks by clicking on the edges. Finally, for each gene in the gene list, one can find how many tissues share this gene. These different query schemes can help experimentalists to identify shared genes of interest in a fast and intuitive fashion. More instructions are available at the website.

Leptokurtic Gene Expression in Both Plants and Animals

One of the most intriguing discoveries from our study is that in all four organisms, the observed leptokurtic gene expression overwhelmingly has a specific conserved pattern consisting of a sharp peak with a rightward heavy tail. We also provide in silico and experimental evidence that leptokurtic genes associated with a tissue are likely to have physiological function in that tissue. Together, these results suggest that most genes are only activated when the function of the gene is required by the gene-associated tissue. It is interesting that the alternative scenario, where a gene is only repressed if the gene’s absence is biologically necessary (i.e., a leftward heavy-tailed pattern), is not widely observed in any of the four organisms, suggesting that the rightward heavy-tailed pattern may represent a general phenomenon for multicellular organisms. This distribution is in some ways analogous to the just-in-time transcription program in single-celled organisms, such as Escherichia coli (Zaslaver et al., 2004), in which enzymes in a metabolic pathway are transcribed sequentially as their substrates become available. Both just-in-time transcription and the active-when-necessary principle can be regarded as mechanisms that maximize the efficiency of the transcriptional program because cellular resources are not allocated to transcribe a gene when the gene product is not necessary.

The observed active-when-necessary principle does not prescribe the underlying transcriptional or posttranscriptional mechanisms that generate the expression pattern. For example, we tested whether known microRNA targets (Backman et al., 2008) are enriched in the high kurtosis genes with heavy right tail or left tail and found no evidence of enrichment in either case (P > 0.4 for both cases; see Supplemental Methods 1 online).
GO Analysis for Human and Arabidopsis Leptokurtic Genes

<table>
<thead>
<tr>
<th>GO Annotation</th>
<th>Annotated</th>
<th>Selected</th>
<th>Expected</th>
<th>P Value</th>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>To 10 Human</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0016021</td>
<td>5072</td>
<td>1150</td>
<td>1025.51</td>
<td>6.58E-03</td>
<td>Integral to membrane</td>
</tr>
<tr>
<td>GO:0005886</td>
<td>4201</td>
<td>1029</td>
<td>849.4</td>
<td>1.29E-07</td>
<td>Plasma membrane</td>
</tr>
<tr>
<td>GO:0005576</td>
<td>2088</td>
<td>769</td>
<td>422.17</td>
<td>7.18E-41</td>
<td>Extracellular region</td>
</tr>
<tr>
<td>GO:0048513</td>
<td>1973</td>
<td>577</td>
<td>401.46</td>
<td>8.16E-03</td>
<td>Organ development</td>
</tr>
<tr>
<td>GO:0005887</td>
<td>1785</td>
<td>453</td>
<td>360.91</td>
<td>5.25E-08</td>
<td>Integral to plasma membrane</td>
</tr>
<tr>
<td>GO:0008283</td>
<td>1414</td>
<td>359</td>
<td>287.71</td>
<td>2.62E-03</td>
<td>Cell proliferation</td>
</tr>
<tr>
<td>GO:0005509</td>
<td>1123</td>
<td>302</td>
<td>225.89</td>
<td>9.09E-09</td>
<td>Calcium ion binding</td>
</tr>
<tr>
<td>GO:0006629</td>
<td>1000</td>
<td>298</td>
<td>203.48</td>
<td>8.10E-04</td>
<td>Lipid metabolic process</td>
</tr>
<tr>
<td>GO:0004888</td>
<td>1162</td>
<td>297</td>
<td>233.73</td>
<td>1.02E-03</td>
<td>Transmembrane receptor activity</td>
</tr>
<tr>
<td>GO:0051239</td>
<td>941</td>
<td>289</td>
<td>191.47</td>
<td>7.33E-03</td>
<td>Regulation of multicellular organismal process</td>
</tr>
<tr>
<td>Top 10 Arabidopsis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0012505</td>
<td>2617</td>
<td>735</td>
<td>399.03</td>
<td>8.56E-76</td>
<td>Endomembrane system</td>
</tr>
<tr>
<td>GO:0003700</td>
<td>1428</td>
<td>257</td>
<td>218.75</td>
<td>2.31E-03</td>
<td>Transcription factor activity</td>
</tr>
<tr>
<td>GO:0016301</td>
<td>1236</td>
<td>230</td>
<td>189.34</td>
<td>3.13E-03</td>
<td>Kinase activity</td>
</tr>
<tr>
<td>GO:0006468</td>
<td>859</td>
<td>180</td>
<td>129.45</td>
<td>1.41E-06</td>
<td>Amino acid phosphorylation</td>
</tr>
<tr>
<td>GO:0005975</td>
<td>739</td>
<td>150</td>
<td>111.37</td>
<td>2.00E-03</td>
<td>Carbohydrate metabolic process</td>
</tr>
<tr>
<td>GO:0005576</td>
<td>364</td>
<td>111</td>
<td>55.5</td>
<td>3.70E-10</td>
<td>Extracellular region</td>
</tr>
<tr>
<td>GO:0004553</td>
<td>327</td>
<td>98</td>
<td>50.09</td>
<td>1.08E-06</td>
<td>Hydrolase activity</td>
</tr>
<tr>
<td>GO:0009535</td>
<td>294</td>
<td>74</td>
<td>44.83</td>
<td>2.61E-04</td>
<td>Chloroplast thylakoid membrane</td>
</tr>
<tr>
<td>GO:0048046</td>
<td>271</td>
<td>72</td>
<td>41.32</td>
<td>9.18E-07</td>
<td>Apoplast</td>
</tr>
<tr>
<td>GO:0004091</td>
<td>229</td>
<td>70</td>
<td>35.08</td>
<td>3.64E-04</td>
<td>Carboxylesterase activity</td>
</tr>
</tbody>
</table>

GO enrichment analysis was carried out using the eliminate count method in topGO (Alexa et al., 2006). All enriched GO terms with P value < 0.01 were obtained, and these terms were then ranked from greatest to least based on the “Selected” column. The top 10 GO annotations with the maximum number of “Selected” high kurtosis genes among all other significant GO annotations are shown here. This way of ranking GO categories guarantees that generic terms (annotated to many genes) are included such that our interpretation of GO enrichment represents generic biological functions. Only one GO category (extracellular region GO:0005576) is enriched in both human and Arabidopsis high kurtosis genes.

- Number of genes annotated by this GO term in the relevant genome.
- Number of high kurtosis genes (Q < 1e-9) annotated by each GO term. This column is called “Significant” by the original topGO software but is changed here to “Selected” for clarity.
- Expected number of genes annotated by this GO term.
- P value is calculated using the eliminate count method with P value of 0.01 as an eliminating parameter.

A short description of each GO term.

Therefore, we conclude that the high kurtosis distribution is not mainly caused by microRNA-mediated gene repression. How the mechanisms of transcript production and degradation are harnessed to create the heavy right-tailed distributions discovered here and disfavor heavy left-tailed distributions is an intriguing question for future research.

Functions of Leptokurtic Shared Genes

Since gene sharing is, by definition, only possible in multicellular organisms, we were interested in generic functions of leptokurtic shared genes. According to the genome expansion theory of multicellularity (Kaiser, 2001; Rokas, 2008), genes with functions of transcriptional regulation, cell–cell communication, and cell adhesion were particularly crucial to the evolution of multicellularity, and gene families conferring these functions proliferated in multicellular organisms via gene duplication. While some gene families apparently proliferated in parallel with the evolution of multicellularity in animals, there is evidence that genes involved in development of the extracellular matrix and cell–cell adhesion of multicellular animals were already present in preanimal genomes of unicellular organisms (Rokas, 2008). Therefore, it is intriguing that GO annotations pertaining to the extracellular region (GO:0005576) are enriched in both human and Arabidopsis shared genes (Table 2), suggesting that these shared genes may have been important for the evolution of multicellularity in both kingdoms. Similarly, several enriched GO annotations in both organisms are related to cell–cell communication and adhesion, albeit in an organism-dependent manner. For instance, the enriched GO annotation “apoplast” (GO:0048046) corresponds to a plant-specific extracellular matrix and diffusion area outside the plasma membrane that is central to plant-specific cell–cell adhesion and communication. Likewise, in human, we find that cell–cell communication-related functions, such as “transmembrane receptor activity” (GO:0004888), are enriched in shared genes. Our enriched GO annotations further suggest a unique importance of the endomembrane system in plant multicellularity, which may be rooted in the fact that intercellular communication of plants is partly mediated by an endomembrane system that remains connected in mature cells through plasmodesmata.

In conclusion, our analytical pipeline revealed new organizing principles of gene expression in multicellular organisms. Our pipeline and associated website provide intuitive ways to explore...
gene expression patterns in multicellular organisms and formulate hypotheses about gene functions within and across tissues. In this first implementation of our transcriptomic approaches, we deliberately focused on transcriptomes of wild-type tissues without environmental perturbation. In the future, it also will be of interest to assess the systems properties we have uncovered in the context of physiological and genetic regulatory events. In addition, the tissue network method is readily applicable to analysis and visualization of high-dimension data sources from other -omic resources, ranging from quantitative proteomic to metabolomic data sets.

**METHODS**

**Microarray Data Sets**

For Arabidopsis thaliana gene expression data, raw CEL files from AtGenExpress (ATGE data set) (Schmid et al., 2005) were obtained from the ME00319 data set at The Arabidopsis Information Resource (TAIR) ftp site: ftp://ftp.Arabidopsis.org/home/tair/Microarrays/Datasets/ExpressionSet_ME00319/. The root data set (ROOT data set) was obtained from GSE5749 and GSE8934 data sets in the Gene Expression Omnibus database (Birnbaum et al., 2003; Brady et al., 2007). Guard cell and whole leaf microarray data sets were generated in our laboratory (GC data set) (Pandey et al., 2010). A total of 237 raw CEL files from 80 tissues were analyzed for Arabidopsis. For rice (Oryza sativa), gene expression data for 42 tissue types (Jiao et al., 2009) were obtained from the rice atlas project from http://bioinformatics.med.yale.edu/riceatlas/overview.jspx. Gene expression data from 73 human (Homo sapiens) tissues and 61 mouse (Mus musculus) tissues with gcma-condensed data sets (Su et al., 2004) were downloaded from http://wombat.gnf.org/index.html.

To compare microarray results, we applied a linear model-based strategy to remove any systematic differences between studies. We showed that commonly used normalization methods are insufficient to remove study-specific effects, such that additional adjustments are required (see Supplemental Methods 1 online). We validated microarray data by qRT-PCR and found high consistency between our renormalized data by qRT-PCR and found high consistency between our renormalized data (see Supplemental Methods 1 online). We validated microarray data by qRT-PCR and found high consistency between our renormalized data set (Pandey et al., 2010). A total of 237 raw CEL files from 80 tissues were analyzed for Arabidopsis. For rice (Oryza sativa), gene expression data for 42 tissue types (Jiao et al., 2009) were obtained from the rice atlas project from http://bioinformatics.med.yale.edu/riceatlas/overview.jspx. Gene expression data from 73 human (Homo sapiens) tissues and 61 mouse (Mus musculus) tissues with gcma-condensed data sets (Su et al., 2004) were downloaded from http://wombat.gnf.org/index.html.

To compare microarray results, we applied a linear model-based strategy to remove any systematic differences between studies. We showed that commonly used normalization methods are insufficient to remove study-specific effects, such that additional adjustments are required (see Supplemental Methods 1 online). We validated microarray data by qRT-PCR and found high consistency between our renormalized results and our qRT-PCR data (Figure 5).

**Statistical Tests of Kurtosis**

We centered gene expression by median and median-absolute-deviation (MAD) to obtain a relative expression level Z score, as calculated by the following formula:

\[
Z_g = \frac{(E_g - \text{MAD}(E_g))}{\text{MAD}(E_g)}
\]

\(E_g\) is the expression level of gene \(g\) in tissue \(i\), \(E_g\) is a vector of all gene expression levels in all tissue samples. Kurtosis is calculated for each gene using the “moments” package (http://rss.acs.unt.edu/Rdoc/library/moments/html/00Index.html) with the following formula:

\[
K = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^4
\]

\(x_i\) is the gene expression in the \(i\)th tissue type, and \(\bar{x}\) is the mean expression of the same gene. The Anscombe test for kurtosis was performed using the Anscombe.test function in the R moments package. We define FDR < 0.5% as a threshold to select genes with leptokurtic distributions. This threshold is more stringent than most published microarray analyses. The stringent threshold is necessary because a large number of tissue conditions are included in our analysis, and higher FDR may increase the number of false positives.

In Figure 2A, for each of the four organisms, the cumulative distribution of kurtosis of all genes is plotted as a solid curve. To compare with the observed kurtosis distribution, we simulate a random background distribution for each gene in all four organisms. In the simulation for each gene, \(N\) random numbers from a normal distribution were generated as simulated expression in \(N\) samples in each of the four microarray data sets. For each gene, the median and MAD of microarray data were used as parameters to generate normally distributed random numbers, and kurtosis of the simulated expressions was calculated. Median and MAD are used here because they are robust estimators of the location and spread of a distribution. Cumulative distribution of each simulated kurtosis was plotted as a dashed line in Figure 2A. On average, < 0.18% of genes from random backgrounds are above our kurtosis threshold. The simulated percentage agrees qualitatively with the FDR threshold and confirms that the Anscombe test does not overestimate the number of leptokurtic genes.

We ranked genes based on their expression kurtosis from the greatest to the least and applied the Anscombe test (Anscombe and Glynn, 1983) to identify genes that have excessive Kurtosis (as measured by FDR-adjusted P value \(Q < 0.005\)). In Arabidopsis (Figure 2A), we find 14,161 probe sets out of 22,810 probe sets (62.1%), corresponding to 14,528 genes out of a total of 23,200 (62.6%) as annotated by TAIR downloaded in April, 2009 (Bioconductor ath1121501.db version 2.2.13), exceed this Q threshold. In rice (Figure 2A), we find 2503 probe sets out of 6320 probe sets (39.6%), corresponding to 2498 genes out of a total of 6275 (39.8%) as annotated by the original authors (Jiao et al., 2009), exceed this Q threshold. In human, 15,657 probe sets out of 18,854 probe sets (78.8%), corresponding to 10,665 genes out of a total of 12,837 (83.1%) genes as annotated by Ensembl in March, 2009 (Bioconductor hg133a.db version 2.2.12), exceed this Q threshold. In mouse, we find 28,900 probe sets out of 36,182 probe sets (79.9%), corresponding to 13,239 genes out of a total of 16,018 (82.6%) as annotated by the original authors using Ensembl identifiers (Su et al., 2004), exceed this Q threshold.

We chose not to use other measurements of the shapes of distributions, such as mean and variances or skewness, because they are less informative than kurtosis for selecting tissue preferential expression patterns (see Supplemental Figure 2 online for an example). Genes with high average expression levels or high variation in expression levels do not necessarily have tissue preferential expression. We also noticed that most genes have positive skewness in expression (data not shown). Positive skewness is a weaker indicator than kurtosis for tissue preferential expression because a distribution with positive skewness only means that the distribution has a central peak leaning toward lower expression levels.

**Gene Selection Based on Reduction of Kurtosis**

We define gene-tissue association using a threshold on the Z score. If a gene has a leptokurtic expression pattern, and high expression in some tissues is removed from the calculation of kurtosis for that gene, then the expression distribution will have a reduced kurtosis (see Supplemental Figure 3 online). Therefore, we associate genes with tissues based on whether a given gene’s expression is higher than Z in a particular tissue, then remove that gene expression in that tissue and calculate kurtosis again. If we use an arbitrarily large Z, moderately expressed genes will not be associated with tissues, even though their expression distribution can be still leptokurtic. On the other hand, using a small Z can significantly decrease the kurtosis calculated from expression levels below Z, but gene-tissue associations identified by such small Z thresholds will be less specific. To obtain a Z threshold that balances the above two issues, we plotted the cumulative reduction of kurtosis curve for any given kurtosis threshold using several different Z threshold values. We then chose a threshold for Z such that for 99% of genes, after removing the expression values above the threshold, expression kurtosis was reduced. In
this way, we obtain a threshold that is dependent on all data available. We define FDR-adjusted P value as Q, and using this method, we found that for Q < 0.005, the corresponding Z threshold is 3, and for Q < 1e-09, the corresponding Z threshold is 10 (see Supplemental Figures 4A and 4B online).

The choice of Z score is also biologically meaningful. We first selected genes that are associated with guard cells with different integer Z scores from 1 to 5. We then tested whether these gene sets are enriched in "stomatal movement function" (GO:0010118). The enrichment score is plotted against different Z thresholds in Supplemental Figure 5 online. We found that guard cell-associated genes are enriched in "stomatal movement function" when the Z threshold is above 3. This result suggests a Z threshold smaller than 3 is not sufficiently specific. A higher Z threshold is also undesirable as fewer genes can be identified for experimental validation.

Comparing Pearson Correlation Coefficients for Shared Genes and Coexpressed Genes

We first identified 158 groups of leptokurtic genes (Q < 0.005) with genes in each group associated with exactly the same subset of tissues (see Supplemental Methods 1 online for details of methods and analysis). We also generated the same number of groups of genes based on coexpression method. We compared the median coexpression values for the genes within each group identified by our pipeline to the groups of genes identified by the coexpression method, and we found that genes grouped together by our pipeline tend to have lower coexpression levels than genes grouped by coexpression methods (Wilcoxon test, P value < 2e-16; Figure 2C).

Tissue (Node) Similarity Score, Tissue Network, and GO Analysis

The tissue (node) similarity (S) score is defined as the negative log10 of the Q value. For any pair of tissues, a P value of Fisher's exact test was calculated based on the number of leptokurtic genes that are highly expressed in each tissue, number of leptokurtic genes that are highly expressed in both tissues, and the sum of common and unique leptokurtic genes between the two tissues. For any pair of tissues, a small P value indicates that the observed number of genes that are associated with both tissues is larger than one would expect by randomly selecting genes from both tissues. FDR correction was used on all calculated p values to obtain a Q value.

In GSN, nodes represent sample types, with size of the node proportional to the number of leptokurtic genes associated with that sample. Two nodes are connected with an edge in Figures 3 and 4 if S is larger than a specified cutoff. We plotted the number of connected components and the number of edges of the networks as a function of threshold S and chose a threshold S such that both networks still have one large connected component (see Supplemental Figure 7A online), and the numbers of edges in the networks start to slowly decrease (see Supplemental Figure 7B online). The choice of S ensures that the network shown in Figure 4 (main text) represents the characteristic structure of the tissue networks.

GO enrichment analysis was performed using the eliminate count method in topGO software (Alexa et al., 2006) and annotation packages hgu133a.db_2.2.12 and ath1121501.db_2.2.13 in Biocoductor version 2.4 (http://bioconductor.org/packages/2.4/data/annotation/). The eliminate count method relies on the graph structure of GO annotations for the calculation of multicomparison adjusted P values for GO enrichment analysis. When the hierarchical structure of GO terms is not accounted for, sometimes, both a child term and a parent term are called significantly enriched by classical Fisher's exact test. In such cases, however, the enrichment of the parent term is simply the result of the number of genes found in the child term. The main reason for choosing the eliminate-count method is to avoid such bias. All enriched GO terms with P value < 0.01 were obtained. In the GO analyses of Figure 3 and Supplemental Figures 8, 9, and 12 online, the resulting eliminate-count P values were converted using negative log10 transformation such that a smaller P value becomes a larger number. The transformed scores were used in hierarchical clustering of GO terms, and results are plotted as heat maps.

Module-Finding Algorithms

We tested three different community-finding algorithms: a fast greedy algorithm (Clauset et al., 2004), a random walk-based algorithm (http://arxiv.org/abs/physics/0512106), and a weighted label propagating algorithm (Raghavan et al., 2007) as implemented in the igraph package (http://igraph.sourceforge.net/index.html). These algorithms were selected because they can detect community in weighted graphs. The first two algorithms terminate based on the criterion of maximum modularity measurement, while the last algorithm is based on numerical simulation. We found that the label propagating algorithm is more appropriate for our data set for two reasons. First, this algorithm assigns each node to modules based on module identity of its first neighbors. In our tissue network, nodes share common genes with their first neighbors but not necessarily with their second neighbors. Second, because this algorithm is not based on modularity of the whole network, small modules that have strong local connections (i.e., for the Arabidopsis network, module 4 with stem first internode and stem second internode) can still be found by the algorithm, while such modules will be undetectable by other algorithms. The ability to find small modules is important for our data sets because some specific tissue types are only sampled in a small number of experiments (e.g., two tissues for Arabidopsis module 4: stems).

Validation of Guard Cell Leptokurtic Genes Using Real-Time qRT-PCR

Total RNA was isolated from tissue samples using the TriZol reagent (Invitrogen), treated with RNase-free DNase I and purified using the RNAeasy kit (Qiagen). For quality control, each RNA sample was analyzed by a bioanalyzer (Agilent Technologies). Purified high-quality RNA samples were used for cDNA synthesis using the Superscript III first-strand synthesis kit (Invitrogen). cDNA was diluted at a concentration of 1:100, aliquoted, and kept at 4°C throughout each experiment to avoid discrepancy in the data because of freeze-thaw cycles. qRT-PCR was performed using premix containing SYBR-Green intercalating dye (Bio-Rad). Actin was used as an internal control (Charrier et al., 2002). The position of the oligonucleotides used for real-time PCR was chosen so that the size of all PCR products was between 100 and 150 bp. The suitability of the oligonucleotide sequences in terms of efficiency of annealing was evaluated in advance using the Primer 3 program. qRT-PCR experiments were repeated thrice independently, and the data were averaged. The data obtained were analyzed with IQ5 software (Bio-Rad). Primers for each gene are given in Supplemental Table 2 online.

Verification of T-DNA Insertional Mutants

The T-DNA Express database (http://signal.salk.edu/cgi-bin/tdnaexpress; Alonso et al., 2003) was searched for the availability of T-DNA insertion lines in Arabidopsis genes showing leptokurtotic expression in guard cells. Four genes, At1g50400, At5g60410 (previously characterized and named szf1; Miura et al., 2005), At2g21080, and At1g11100 were selected for further characterization as they had at least two independent and homoygous insertion lines that we could confirm at the time of the analysis (Figure 6). T-DNA left border LBB1 and gene-specific primers were used to confirm the reported insertion sites. Gene-specific primers flanking each of the insertion sites were used to confirm the homozygosity of the insertions by genomic DNA PCR. DNA isolated from wild-type Col plants was used as controls in PCR reactions. Gene-specific primers flanking
the insertion sites were used in RT-PCR (Figure 6B) to confirm absence of full-length transcripts in the T-DNA insertional lines. The sequences of the primers used for PCRs are listed in Supplemental Table 2 online.

Functional Assays

For stomatal aperture measurements, plants were grown under identical conditions as for guard cell isolation for microarray analysis (Pandey et al., 2010). For each sample, two fully expanded leaves from 5-week-old plants were harvested just before the beginning of the light period. Ex- cised leaves were placed in a six-well Petri dish, abaxial side down in 5 mL of buffer solution. The composition of the solution for assays of ABA inhibition of stomatal opening was 10 mM KCl, 7.5 mM iminodiacetic acid, and 10 mM MES, pH 6.15, and the composition of the solution for assays of ABA promotion of stomatal closure was 20 mM KCl, 5 mM MES, and 1 mM CaCl₂, pH 6.15. Excised leaves from wild-type Col and the homozygous mutant plants were kept in darkness for 2 h and then transferred to light (450 μmol m⁻² s⁻¹) for 2.5 h in the presence of ABA (50 μM) to study inhibition of opening. For promotion of closure experiments, excised leaves were first kept in light for 2 h followed by addition of ABA (50 μM) and further incubation in light for 2.5 h. Cover slips (two per well) were added on top of the leaves to keep them evenly submerged in the buffer solutions. Ethanol was used as the solved control.

The epidermis of the leaves was peeled after the end of incubation periods and mounted on a glass slide. The peels were photographed using a digital camera mounted to a Nikon Diaphot 300 microscope. Ten images were recorded per leaf (20 per sample). Apertures were measured using Image J software (rsweb.nih.gov/j), with a micrometer image (photographed with each sample) used as a scale. At least 100 apertures were measured for each sample. Three double-blind biological replicates of the experiment were performed, and data were averaged.

At least 30 aperture lengths were measured from epidermal peels and data were averaged. For the measurement of stomatal lengths (Figure 6E), the measurements were performed along the length of the aperture between the two guard cells. For each of the two alleles of each mutant, the average stomatal length was compared with Col (the wild type) by a t test. All six alleles (two alleles for each mutant of At2g21080, At1g11100, and szt1) are statistically significantly different from Col (P < 0.01 for each of the six comparisons).

Root elongation assays were performed as described by Pandey et al. (2006) with slight modifications. Seeds were plated and stratified in the dark for 48 h at 4°C as described by Pandey et al. (2006). Plates were then transferred to long-day growth conditions of 16 h light/8 h darkness (0.120 μmol m⁻² s⁻¹) for 24 h. Seeds were then transferred to fresh control (ethanol) or ABA (1 μM) media plates. Growth was recorded 8 d after transfer to plates. Two replicates were performed, and data were averaged. Results marked by an asterisk are significantly different from Col at P ≤ 0.01 (Student’s t test).

Accession Numbers

Sequence data from this article can be found in the TAIR database (www.Arabidopsis.org) under the following accession numbers: At2g22320, At1g64010, At2g21080, At3g51760, At4g16820, At1g11100, At4g28460, At1g50400, At2g19180, At5g35320, At4g11330, At5g60410 (SIZ1), At1g20880, At2g27200, At5g36720 (PHYB), At3g18780 (ACTIN 2), At1g19240 (ACTIN8), At1g31335, At1g65020, and At4g01880.

Supplemental Data

The following materials are available in the online version of this article.

Supplemental Figure 1. Study-Specific Effects in RMA-Normalized Expression Data and in Data after Adjustment by a Nested Linear Model Method.

Supplemental Figure 2. Mean, Variance, Skewness, and Kurtosis for Two Typical Genes.

Supplemental Figure 3. Cumulative Distribution of Kurtosis before and after Z Threshold.

Supplemental Figure 4. Reduction of Kurtosis Plots for Arabidopsis and Human Data.

Supplemental Figure 5. Enrichment of Arabidopsis Guard Cell Functional Genes.

Supplemental Figure 6. Leptokurtic Genes Have Multitissue Expression Patterns.

Supplemental Figure 7. Graph Theoretical Properties of Tissue Networks with Q = 1e-09 and Z = 10 and Different S Values.

Supplemental Figure 8. Complete GO Annotation Analysis for Genes Appearing in the Seed Subnetwork.

Supplemental Figure 9. Complete GO Annotation Analysis for Genes Appearing in the Shoot Apex Subnetwork.

Supplemental Figure 10. Human Gene-Sharing Network.

Supplemental Figure 11. Human Liver and Fetal Liver Subnetwork.

Supplemental Figure 12. GO Analysis for Human Liver and Fetal Liver Subnetwork.

Supplemental Table 1. Coexpression Analysis: 14 Samples That Have the Highest Correlation Coefficient with Guard Cells.

Supplemental Table 2. Primers Used for Real-Time PCR Analysis of Guard Cell-Abundant Genes and for Genotyping of Knockout Lines.


Supplemental Data Set 1. Annotations for Arabidopsis GSN (Figures 3 and 4), Which Include Two Sheets in a Single Excel File.

Supplemental Data Set 2. Annotations for Human GSN (See Supplemental Figure 10 Online), Which Include Two Sheets in a Single Excel File.

Supplemental References 1. Supplemental References for the Supplemental Data.

ACKNOWLEDGMENTS

This research was supported by National Science Foundation Grants NSF-MCB-0209694 and NSF-MCB-0618402 to S.M.A. We thank Reka Albert for helpful discussions on network analysis.

AUTHOR CONTRIBUTIONS

S.P., T.E.G., Z.Z., and L.W. performed microarray experiments and identified T-DNA insertional mutants. S.P. and L.W. performed the qRT-PCR and mutant phenotype analysis. S.L. implemented the algorithms and analyzed the data. S.L., S.P., and S.M.A. wrote the article with input from all authors.

Received December 11, 2011; revised February 29, 2012; accepted March 16, 2012; published April 18, 2012.

REFERENCES


Gene-Sharing Networks Reveal Organizing Principles of Transcriptomes in *Arabidopsis* and Other Multicellular Organisms
Song Li, Sona Pandey, Timothy E. Gookin, Zhixin Zhao, Liza Wilson and Sarah M. Assmann
*Plant Cell* 2012;24;1362-1378; originally published online April 18, 2012;
DOI 10.1105/tpc.111.094748

This information is current as of June 29, 2017

<table>
<thead>
<tr>
<th>Supplemental Data</th>
<th>/content/suppl/2012/03/22/tpc.111.094748.DC1.html</th>
</tr>
</thead>
<tbody>
<tr>
<td>References</td>
<td>This article cites 42 articles, 25 of which can be accessed free at: /content/24/4/1362.full.html#ref-list-1</td>
</tr>
<tr>
<td>eTOCs</td>
<td>Sign up for eTOCs at: <a href="http://www.plantcell.org/cgi/alerts/ctmain">http://www.plantcell.org/cgi/alerts/ctmain</a></td>
</tr>
<tr>
<td>CiteTrack Alerts</td>
<td>Sign up for CiteTrack Alerts at: <a href="http://www.plantcell.org/cgi/alerts/ctmain">http://www.plantcell.org/cgi/alerts/ctmain</a></td>
</tr>
<tr>
<td>Subscription Information</td>
<td>Subscription Information for <em>The Plant Cell</em> and <em>Plant Physiology</em> is available at: <a href="http://www.aspb.org/publications/subscriptions.cfm">http://www.aspb.org/publications/subscriptions.cfm</a></td>
</tr>
</tbody>
</table>