Running title: Characterizing coexpression modules in tomato

Corresponding author: Atsushi Fukushima
RIKEN Plant Science Center, Yokohama, Kanagawa 230-0045, Japan
e-mail address: a-fukush@psc.riken.jp
PHONE: +81 45 503 9491  FAX: +81 45 503 9489

Journal research area: Genome Analysis; Systems Biology
Title: Exploring Tomato Gene Functions Based on Coexpression Modules Using Graph Clustering and Differential Coexpression Approaches

Authors:
Atsushi Fukushima¹,*
Tomoko Nishizawa¹,
Mariko Hayakumo²,
Shoko Hikosaka²,
Kazuki Saito¹,³,
Eiji Goto²,
Miyako Kusano¹,⁴

*Corresponding author

Institutes:
¹RIKEN Plant Science Center, Yokohama, Kanagawa 230-0045, Japan
²Graduate School of Horticulture, Chiba University, Matsudo, Chiba 271-8510, Japan
³Graduate School of Pharmaceutical Sciences, Chiba University, Chiba, Chiba 263-8522, Japan
⁴Kihara Institute for Biological Research, Yokohama City University, Yokohama, Kanagawa 244-0813, Japan

E-mail addresses of co-authors:
Atsushi Fukushima:  a-fukush@psc.riken.jp
Tomoko Nishizawa:  guko2430@psc.riken.jp
Mariko Hayakumo:  hayakumo_mariko@graduate.chiba-u.jp
Shoko Hikosaka:  s-hikosaka@faculty.chiba-u.jp
Kazuki Saito:  ksaito@psc.riken.jp
Eiji Goto:  goto@faculty.chiba-u.jp
Miyako Kusano:  mkusano005@psc.riken.jp
Footnotes:
1This work was cofinanced by Elucidation of biological mechanisms of photoresponse and development of advanced technologies utilizing light sponsored by Ministry of Agriculture, Forestry and Fisheries (MAFF) (five-year project) and a Grant-in-Aid for Young Scientists (B; grant no. 23700355 to A.F.) from the Ministry of Education, Culture, Sports, Science and Technology (MEXT).

* Corresponding author; Atsushi Fukushima; e-mail: a-fukush@psc.riken.jp
The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.planphysiol.org) is: Atsushi Fukushima (a-fukush@psc.riken.jp).
**Abstract**
Gene-to-gene coexpression analysis provides fundamental information and is a promising approach for predicting unknown gene functions in plants. We investigated various associations in the gene expression of *Solanum lycopersicum* (tomato) to predict unknown gene functions in an unbiased manner. We obtained more than 300 microarrays from publicly available databases and own hybridizations, and here, we present tomato coexpression networks and coexpression modules. The topological characteristics of the networks were highly heterogenous. We extracted 465 total coexpression modules from the dataset by graph clustering, which allows users to divide a graph effectively into a set of clusters. Of these, 88% were assigned systematically by gene ontology terms. Our approaches revealed functional modules in the tomato transcriptome data; the predominant functions of coexpression modules were biologically relevant. We also investigated differential coexpressions among datasets consisting of leaf-, fruit-, and root samples to gain further insights into the tomato transcriptome. We now demonstrate that (1) duplicated genes, as well as metabolic genes, exhibit a small but significant number of differential coexpressions, and (2) a reversal of gene coexpression occurred in 2 metabolic pathways involved in lycopene and flavonoid biosynthesis. Independent experimental verification of the findings for 6 selected genes was done using quantitative real-time polymerase chain reaction. Our findings suggest that differential coexpression may assist in the investigation of key regulatory steps in metabolic pathways. The approaches and results reported here will be useful to prioritize candidate genes for further functional genomics studies of tomato metabolism.

**Keywords:** *Solanum lycopersicum*, gene coexpression, graph clustering, network module, differential coexpression, microarrays
Introduction

One of the major challenges of plant-systems biology is in understanding genotype-phenotype associations. In that context, biological networks can increase our understanding of how biomolecules interact to function in plants (Fukushima et al., 2009; Stitt et al., 2010). Large-scale data from genome-wide gene expression profiling with DNA microarrays are publicly available for many species, including *Arabidopsis thaliana*, *Oryza sativa* (rice), poplar, and some crops (Ogata et al., 2010; Tohge and Fernie, 2010). These data make it possible to use gene coexpression analyses to predict unknown gene functions (Aoki et al., 2007; Usadel et al., 2009). Using pair-wise measures, e.g. Pearson’s correlation coefficient, it is possible to generate a coexpression network in which nodes represent genes and edges represent significant correlations between expression patterns. Network representation facilitates the prioritization of candidate genes for further functional genomics studies based on the so-called “guilt-by-association” principle (Saito et al., 2008).

Generally, graph clustering algorithms include the hierarchical clustering, the density-based and local searches, and other optimization-based clustering as summarized by Wang and co-workers (Wang et al., 2010). In Arabidopsis and rice microarray datasets, such algorithms, including Markov clustering (Van Dongen, 2000) and DPClus (Altaf-Ul-Amin et al., 2006), were applied to find coexpression modules, which are clusters consisting of densely-connected coexpressed genes (Ma et al., 2007; Mentzen and Wurtele, 2008; Fukushima et al., 2009; Mao et al., 2009). These types of network-module-based approaches are now widely used in attempts to predict new genes involved in biological processes (Saito et al., 2008; Usadel et al., 2009; Mutwil et al., 2011). Other network-based approaches have been applied to annotate unknown genes (Horan et al., 2008), to explore possible genes involved in carbon/nitrogen-responsive machineries (Gutierrez et al., 2007), and to prioritize candidate genes for a wide variety of traits (Lee et al., 2010).

Subsets of highly coexpressed genes in *Solanum lycopersicum* (tomato) have been studied. Miozzi and colleagues, who investigated conserved coexpressions in the Solanaceae family, including tomato, tobacco, and potato, used expressed sequence tags (ESTs) for tomato and compared transcriptional presence/absence patterns with those of other species to explore functional relationships between genes (Miozzi et al., 2010). Aoki and his colleagues demonstrated that coexpression modules extracted by network-based clustering
across various developmental stages and organs facilitated the functional analysis of genes encoding flavonoid biosynthesis in tomatoes (Ozaki et al., 2010).

Coexpression patterns also change based on different conditions such as the genotype and tissue type; this has been termed differential coexpression (Choi et al., 2005; Gillis and Pavlidis, 2009). Differential coexpression identifies the rewired edges of coexpression networks and may reflect changes in transcriptome organization (Watson, 2006; Chia and Karuturi, 2010). For example, this feature has been used to identify disease-specific networks (de la Fuente, 2010). Differential metabolomic correlations were utilized in some metabolomics studies (Weckwerth et al., 2004; Morgenthal et al., 2006; Fukushima et al., 2011; Kusano et al., 2011). While the differential expression of genes and gene coexpressions in plants have been addressed in earlier microarray studies, their differential coexpression has not been addressed.

Here, we present tomato coexpression modules extracted by graph clustering from over 300 microarrays obtained from publicly available databases and 20 of our own hybridizations in our efforts to unravel associations between gene expression patterns in the tomato. We assessed the clustered modules systematically by Gene Ontology (GO) enrichment analysis. We also investigated differential coexpressions among datasets of leaf-, fruit-, and root samples, based on Fisher’s Z-transformation, to gain insight into the tomato transcriptome. The aims of our study were to (1) characterize coexpression modules in tomato by using graph clustering and GO enrichment analysis, (2) identify significant differential coexpressions between organs, and (3) provide clues for elucidating key regulatory networks in a given metabolism. The expression levels of the obtained gene pairs were validated by quantitative real-time polymerase chain reaction (qRT-PCR).

Results

Construction of a Gene Coexpression Network in Tomato

We collected 307 tomato GeneChips from publicly available databases and 20 of our own microarray datasets. A statistical quality check of all microarrays (see Methods) detected 307 high-quality GeneChips; these were used for further analyses (Figure 1). After normalization we manually classified them according to meta-data (Supplemental Data 1); 307 arrays included 17 experiments, with
23.5% involving a stress condition, 23.5% a genotype, and 17.6% a pathogen (Figure 2A). They also corresponded to different organs, including fruits (44%), leaves (25.1%), and roots (11.7%) (Figure 2B). Based on the organ type in the metadata, we divided the original expression data matrix (307 arrays × 9,989 probe-sets, defined as “all data”) into 3 sub-matrices that we called “leaves”, “fruits”, and “roots”. We used these matrices, as well as “all data”, in further analyses (Figure 1). The distribution of the 3 expression matrices appeared normal (Supplemental Figure S1A). To construct the coexpression networks, we calculated correlation matrices using Pearson’s correlation coefficients. The correlation coefficients for each organ were also in normal distribution (Supplemental Figure S1B).

The Topological Characteristics of Tomato Coexpression Networks Are Highly Heterogenous

We next constructed tomato coexpression networks. To evaluate whether they manifested the common properties of a complex network, such as power-law degree distribution, we investigated the topological characteristics of the network (e.g., degree distribution and average path length) (Table I). We found that the degree distribution of the coexpression network followed a power law (Figure 3A). In the case of the coexpression network with $r \geq 0.6$ ($p < 2.1 \text{e-31}$, Pearson’s correlation statistical test), the degree distribution of “all data” network followed a power-law with a degree exponent of $\gamma = 1.67$. Here, $P(k) \sim k^{-\gamma}$, where $\gamma$ represents the degree exponent. The average path length and the average clustering coefficient of this network were 2.65 and 0.45, respectively, implying small-world properties and high modularity in the network (see review by (Barabasi and Oltvai, 2004)). Figure 3B shows a partial coexpression network generated using “all data”; it was based on the list of first-order genes that neighbored and were coexpressed with the LeMADS-Rin gene, which encodes a MADS box transcriptional factor regulating fruit ripening-related genes (Giovannoni et al., 1995; Vrebalov et al., 2002). The network shows how the neighboring genes of LeMADS-Rin correlate with each other.

Graph Clustering and Gene Ontology Enrichment Analyses Reveal Functional Modules in the Tomato Transcriptome Dataset
To efficiently identify densely connected nodes in the coexpression network, i.e., a coexpression module, we used the graph clustering algorithm IPCA (Li et al., 2008). We detected 465 modules with at least 5 gene members in the tomato coexpression network constructed from “all data”, consisting of 9,797 nodes and 1,754,361 edges; they ranged in size from 5 to 68 genes (Table II and Supplemental Table S1). We counted the number of unknown genes per module (Figure 4 and Supplemental Table S2). The distribution of the percentage of unknown genes within a module showed a bimodal curve at 50% and 70%.

We then subjected the modules to GO enrichment analysis; for simplicity, we selected a GO term with the best \( p \)-value within the Biological Process (BP), Cellular Component (CC), and Molecular Function (MF) domains. For over-represented GO terms, Table III lists the top 10 modules repeatedly assigned by the same functional term (see also Supplemental Tables S1 and S3). Of the modules extracted, 88% had significantly enriched GO terms in BP, CC, and MF. The distribution of GO terms for each module in tomato is shown in Supplemental Table S3. There were 150, 110, and 51 modules in the specific functional categories of “negative regulation of ligase activity”, “DNA endoreduplication”, and “photosynthesis”, respectively.

**Predominant Functions of Coexpression Modules Show Biological Relevance**

Below, we focus on the details of the extracted modules in Table II. The modules selected for our analysis follow (see also Supplemental Table S1).

**DNA endoreduplication**

Module 1, consisting of 63 genes, was involved in “DNA endoreduplication” (false discovery rate (FDR) = 3.1e-2). It contained genes encoding cyclin A1 and A2, histone H4, B1-type cyclin-dependent kinase, ripening-regulated protein (Differential Display Tomato Fruit Ripening (DDTFR)-18), copper transport protein 1 (CTR1)-like protein kinase 2 (TCTR2), jasmonic acid 1 and 2 (LEJA1, LEJA2), cytosolic glutamine synthetase 1, and 12-oxophytodienoate reductase 3.

**Photosynthesis**

Module 4 (54 genes) was associated with “photosynthesis” (FDR =
It included genes encoding the photosystem I subunit II protein precursor, photosystem II 23 kDa protein, chlorophyll a/b-binding protein precursor, ribulose-1,5-bisphosphate carboxylase, ADP-glucose pyrophosphorylase large subunit, photosystem 1 reaction center protein subunit 2 (PSAD), and zeaxanthin epoxidase (ZEP), an enzyme important in the xanthophyll cycle and in abscisic acid (ABA) biosynthesis.

Response to cold
Module 435, consisting of 33 genes, was related to “response to cold” (FDR = 4.3e-2). This module included genes encoding cytosolic ascorbate peroxidase 2 and dehydroascorbate reductase 2. These genes are involved in the ascorbate-glutathione cycle that scavenges reactive oxygen species, particularly H$_2$O$_2$.

Jasmonic acid metabolic process
Module 457 (8 genes) was involved in the “jasmonic acid metabolic process” (FDR = 8.7e-3). It contained the LeMTS1 gene encoding tomato monoterpene synthase. A homeobox knotted-1-like gene, LeT6 (Chen et al., 1997), was also included in this module. The LeT6 gene is essential for meristem maintenance and the process of leaf initiation (see Discussion).

Comparative Analyses of Topological Characteristics in Tomato Transcriptome Coexpression in Different Organ Datasets
To obtain further insight into the tomato transcriptome, we employed a comparative approach to organ-specific coexpression. Using 3 sub-datasets, “leaves”, “fruits”, and “roots”, we first assessed the topological characteristics of each coexpression network. As shown in Figure 5A, the degree distribution of each network ($r \geq 0.6$) was indicative of a typical power-law distribution in a wide range of values for the degree $k$. We also calculated several graph-theoretic statistics including the number of edges, average path length, and clustering coefficient (for example, see (Barabasi and Oltvai, 2004)) (Table I). In the “roots” dataset, we observed that the highest number of edges and the highest clustering coefficient were 5,997,644 and 0.70, respectively. The average path length in 3 organs was smaller than 3, indicating that these networks are characterized by small-network property. To assess the degree of node overlap
among the 3 datasets, we investigated similarity scores for nodes based on their connection partners using the Jaccard coefficient, which assesses the interaction between two graphs by assessing the tendency that links simultaneously are present in both graphs. Figure 5B shows the relationship between the average of the Jaccard coefficient and the correlation coefficient, and it indicates that the larger the cut-off of the correlation coefficient, the smaller the average similarity. In this graph, we observed that the roots dataset had the highest Jaccard coefficient in all ranges.

**Direct Measures of Differential Coexpressions in the Tomato Transcriptome**

We next focused on changes in coexpression patterns between gene expression levels. We posited that a significant coexpression between given genes may be found under one condition but not another and that the changes elicited under one condition may be reversed under the other condition. An example of changed coexpressions is shown in Figure 6. The correlation coefficient between Les.107.1.S1_at (SGN-U565450, cyclin A2) and Les.1081.1.S1_at (SGN-U577270, putative ribosomal protein) was negative ($r = -0.66$) in leaves but, positive ($r = 0.71$) in fruits. To compare coexpression patterns among organs, we first visualized the 3 correlation matrices using pseudo-heatmaps (Supplemental Figure S2). While genes involved in photosynthesis were highly coexpressed in leaves and fruits, in roots they showed no remarkable coexpression within the pathway. For other pathways, root data displayed overall positive coexpressions.

To obtain the difference between coexpressions in the 3 organs, we used Fisher’s Z-transformation. This approach can test directly for differential coexpression by testing, for example, the null hypothesis $H_0: r^L = r^F$. Here, $r^L$ and $r^F$ indicate that coexpression is calculated over the leaves and fruits samples, respectively (see Methods). In 3 comparisons, “fruits” vs. “roots”, “leaves” vs. “fruits”, and “leaves” vs. “roots”, the numbers of significantly different correlation pairs were 753,133, 826,190, and 395,814, respectively (FDR < 1e-10) (Figure 7). Full lists of the differential coexpressions are shown in Supplemental Data 2. Using GO terms, we characterized the gene pairs with significant differential coexpressions (Table IV and Supplemental Table S4). For example, in the transition from $r > 0.7$ in leaves to $r < -0.7$ in fruits, the differential coexpression
included genes associated with cell wall modifications (FDR = 7.37e-03), while in the opposite transition ($r < -0.7$ in leaves to $r > 0.7$ in fruits), it included genes involved in flower development (FDR = 5.32e-04). In the case of the GO term “flower development”, we observed 17 annotated probes including $LeMADS-Rin$ and other genes encoding the MADS-box protein (Supplemental Table S4).

**Duplicated Genes Show a Small But Significant Number of Differential Coexpressions**

We further investigated whether differential coexpressions exist between duplicated genes. We first classified the all probe-set “target” sequences into similarity clusters (referred to as gene families; see Methods). Consequently, 1677 obtained clusters were regarded as gene families in this analysis. For 1677 gene families including duplicated genes we calculated the number of significantly different coexpressions (FDR < 1e-10). The numbers of duplicated genes that were differentially coexpressed among organs is listed in Table V and Supplemental Table S5. Overall, we were able to observe a small but significant number of differential coexpression relationships (Table V). For example, a gene family, Markov Cluster (MC)100, was characterized by the GO terms “photosynthesis, light harvesting (FRD = 6.1e-5)” and “protein-chromophore linkage (FDR = 6.1e-5)”. In MC100, there were 19 differentially coexpressed pairs between leaves and roots, while there were 41 pairs between fruits and roots.

To gain further insights into genes with significant differential coexpression, we investigated the number of significant differential coexpressions (FDR < 1e-10) for each of the 256 metabolic pathways from the LycoCyc (Bombarely et al., 2011) (Table VI and Supplemental Table S6). For example, there were 15 differential coexpressions in “glycolysis IV (plant cytosol)” between fruits and roots but only 3 differential coexpressions between leaves and roots. The coexpression patterns in biosynthetic pathways associated with the Calvin cycle, sugar degradation, photorespiration, and the tricarboxylic acid (TCA) cycle, as well as glycolysis, were usually different among organs. In comparisons between leaves and fruits, but not in other comparisons, there were 3 differentially coexpressed pairs in anthocyanine biosynthesis. On the other hand, gene pairs in flavonoid biosynthesis were different when compared to roots. We also studied duplicated genes encoding isozymes that
were primarily derived from LycoCyc. Most of these genes were inferred from computational analysis without any human curation. As shown in Supplemental Table S7, we observed quite a few pairs with significantly different coexpression (FDR < 1e-10) between duplicated genes (isoforms). For example, the number of significantly different coexpression genes encoding beta-fructofuranosidase was 4 of 528 possible gene pairs.

**Differential Coexpressions Provide Clues of Key Regulatory Steps in Metabolic Pathways**

To interpret differential coexpression in metabolic pathways, we investigated, in detail, 2 biosynthetic pathways, lycopene and flavonoid, in a proof-of-concept study.

**Lycopene biosynthesis**

The changes in coexpressions between organs involved in carotenoid biosynthesis are of particular interest because this pathway contributes to the production of lycopene and β-carotene, which can be utilized as indicators of tomato quality. The mapping of differential coexpressions between leaves and fruits onto a lycopene biosynthesis pathway (Figure 8) revealed significantly different coexpressions in this pathway. The coexpression of genes encoding 2 isoforms of phytoene synthase (PSY) [Les.3171.4.A1_at (SGN-U580527)] and abscisic acid stress ripening 1 (Asr1) [Les.4930.1.A1_at (SGN-U581076)] was significantly different between leaves and fruits [(I) in Figure 8]. In leaves, their expressions were positively correlated ($r = 0.75$); there was no correlation in fruits ($r = -0.09$). The expression of the *psy1* gene [Les.3171.4.A1_at (SGN-U580527)] and that of the gene encoding ζ-carotene desaturase (zs) [Les.20.1.S1_at (SGN-U568537)] exhibited a highly positive correlation ($r = 0.82$) in fruits and a weak negative correlation in leaves ($r = -0.22$) [(II) in Figure 8]. The zs gene [Les.20.1.S1_at (SGN-U568537)] and the gene encoding LELYOCYC lycopene epsilon-cyclase [Les.3771.1.S1_at (SGN-U567885)] were mildly coexpressed in leaves ($r = 0.58$); in fruits, the coexpression was mildly negative ($r = -0.49$) [(III) in Figure 8]. We also observed differential coexpressions in genes associated with lutein biosynthesis and abscisic acid biosynthesis pathways [(IV) in Figure 8]. The coexpression between the gene encoding ε-ring hydroxylase [Les.4915.1.S1_at (SGN-U562951)] and the
short-chain dehydrogenase/reductase (SDR) homolog [LesAffx.68802.1.S1_at (SGN-U580225)] was highly positive ($r = 0.72$) in fruits and, mildly negative ($r = -0.31$) in leaves.

**Flavonoid biosynthesis**

There were 3 differential coexpression patterns in flavonoid biosynthesis (Figure 9). The gene encoding flavanone 3-hydroxylase (F3H) [Les.2278.1.S1_at (SGN-U563669)] and the gene encoding 4-coumarate-CoA ligase [Les.5848.1.A1_at (SGN-U579683)] exhibited a highly positive correlation ($r = 0.89$) in fruits and a weak negative correlation in roots ($r = -0.23$) [(I) in Figure 9]. Similarly, the gene encoding flavanone 3-hydroxylase (F3H) [Les.2278.1.S1_at (SGN-U563669)] and the gene for naringenin-chalcone synthase (Chs2) [Les.3649.1.S1_at (SGN-U580856)] showed a highly positive correlation ($r = 0.89$) in fruits and a weak negative correlation in roots ($r = -0.16$) [(II) in Figure 9]. The last pair manifested pattern that was in direct contrast to the one observed between the gene encoding chalcone isomerase [Les.3218.1.S1_at (SGN-U579009)] and putative naringenin-chalcone synthase [Les.4911.1.S1_at (SGN-U581366)], which showed a positive correlation ($r = 0.50$) in fruits and a highly negative correlation ($r = -0.72$) in roots [(III) in Figure 9].

**Independent Experimental Verification of The Microarray Datasets Using qRT-PCR**

To independently confirm the expression profiling data, 6 expressed genes identified in our microarray results were assayed with qRT-PCR at 7 experimental conditions using the same RNA sample that was used for our original experiment consisting of 20 microarrays (see Methods). We chose 6 genes that were involved in photosynthesis, flavonoid biosynthesis, and carotenoid biosynthesis, which corresponded with 3 types of expression pattern: (1) leaf-specific (2) fruit-specific and (3) moderate (Figure 10A). We verified that the expression patterns were consistent between microarray and qRT-PCR analyses and showed very good reproducibility (Supplemental Figure S3). Thus, we can state the followings: (i) the significant coexpression in Module 4, which includes the genes zep and psad, was reproducible ($r = 0.84$, $p = 3.63e-6$ in Figure 10) and (ii) regarding flavonoid genes, the calculation based on fruit samples resulted in significant coexpression ($r = 0.84$, $p = 0.0343$). This result
supports, at least in part, our approach (see also Figure 9). These verification results suggest that the resulting coexpressions derived from 307 collected datasets, as well as each sub-dataset (“leaves”, “fruits”, and “roots”), are relevant, although we must note coexpression between genes with low-level expression.

Discussion
We presented a comprehensive coexpression network analysis based on over 300 tomato microarrays and investigated the topological properties of the network. The degree distribution of the coexpression network followed the heterogeneous power law. The sub-datasets from 3 organs, leaf-, fruit-, and root samples, also exhibited high heterogeneity and modularity (e.g., power-law degree distributions and small-world properties). This result indicates that the properties of the network are consistent with earlier reports on typical coexpression networks (Stuart et al., 2003). We constructed genome-wide tomato coexpression networks with a correlation cut-off value of $r \geq 0.6$ ($p < 2.1 \times 10^{-31}$, Pearson’s correlation statistical test) using “all data”. Our threshold selection was supported, at least partially, by earlier studies. Aoki and colleagues showed that the Arabidopsis coexpression network had a minimal network density at a cut-off $r$ value ranging from 0.55 to 0.60 (Aoki et al., 2007). Using simulated datasets, Elo and colleagues demonstrated that a cut-off value of $r = 0.6$ yielded both low error and high reproducibility (Elo et al., 2007).

We clustered densely coexpressed genes by graph clustering, a non-targeted approach that divides the tomato coexpression network into gene modules. We detected 465 coexpression modules and assessed them by GO term enrichment analysis. A coexpression module-based approach applied to 67 microarrays from 24 different tissues in tomato provided strong predictions for gene functions, such as the pathways involved in flavonoid biosynthesis (Ozaki et al., 2010). Based on the evolutionary conservation of coexpressions in the Solanaceae family, Miozzi and colleagues developed the data-mining tool, ORTom, which predicts functional relationships among tomato genes (Miozzi et al., 2010). The Tomato Functional Genomics Database (TFGD) offers a web-based retrieval tool for coexpressed genes (Fei et al., 2011). There are several differences between our approach and these previous studies: First, our study involved a much larger number of microarrays (>300) and covered wider experimental conditions, including various stress experiments. Second, we
introduced differential coexpression into our plant transcriptome study. In addition, the extracted modules carried many unknown genes (Figure 4). Finally, the expression profiles of 6 selected genes were independently verified by qRT-PCR (Supplemental Figure S3 and Figure 10).

The coexpression-module-based approach by graph clustering is an important approach to facilitate predictions (Mentzen and Wurtele, 2008; Atias et al., 2009; Fukushima et al., 2009; Mao et al., 2009). In our previous works (Fukushima et al., 2009; Fukushima et al., 2011), we used DPclus-based graph clustering to characterize gene coexpression networks and metabolomic correlation networks. Although we believe that DPclus is also useful for gene coexpression networks, it has a limitation on the number of nodes (limitation: \( n < 5,000 \)). To this end, we chose IPCA (Li et al., 2008) that enables us to perform the graph clustering for larger-scale networks in this study. We demonstrated that 88% of all coexpression modules in the tomato were assigned by GO terms, although the modules to which we assigned GO terms overlapped markedly. The tomato genome sequencing project and enhanced genomic annotations in tomato (Barone et al., 2008; Bombarely et al., 2011) will improve such characterizations of coexpression modules. We highlighted 4 modules, Module 1, Module 4, Module 435, and Module 457 (Table II). The coexpression between genes encoding cyclin B1 and histone H4 in Module 1 (GO term: DNA endoreduplication) has been observed in Arabidopsis [See ATTED-II database; (Obayashi et al., 2011)]. Module 4 (GO term “photosynthesis”) is consistent with previous reports in Arabidopsis and rice (Mentzen and Wurtele, 2008; Fukushima et al., 2009). Genes in Module 435 (GO term “response to cold”) are reasonable in the sense that abiotic stresses, including cold stress, are related to oxidative stress that perturbs almost all functions in a plant cell. Module 457 (GO term “Jasmonic acid metabolic process”) was also relevant because it included the \( LeMTS1 \) gene. According to van Schie and colleagues, the expression of the gene in tomato leaves was induced by jasmonic acid treatment, and this gene encoded a linalool synthase (van Schie et al., 2007). We also observed the \( LeT6 \) gene (Chen et al., 1997) in this module. Although both genes were highly coexpressed, their relationship is currently unclear. Using the \( arf6arf8 \) double mutant, Tabata and his colleagues showed that Arabidopsis auxin response factors 6 and 8 regulate jasmonic acid biosynthesis and floral organ development via class 1-KNOX genes (Tabata et al., 2009). This mutant is similar to that found in plants harboring a mutation associated with the
biosynthetic pathway of jasmonic acid, which causes morphologic abnormality in floral organ development. Our clustering results may suggest a yet unknown transcriptional coordination between jasmonic acid biosynthesis and KNOX genes in tomato. The predominant functions of the coexpression modules detected here show biological relevance. Our top-down approach revealed good candidates for functional modules in the tomato transcriptome. That is, when using the “guilt-by-association” principle, such a coexpression-module-based approach can help to predict the function of unknown genes within a module, although coexpressed genes are not necessarily involved in the same biological process.

We studied differential coexpressions among tomato leaves, fruits, and roots using Fisher’s Z-transformation. Direct measurements showed that duplicated genes as well as metabolic genes exhibit a small but significant number of differential coexpressions. Gene clusters assigned the GO term “flower development” as an example of switching coexpression, included LeMADS-Rin (Table IV and Supplemental Table S4). Regarding the link between the identification of differential coexpression and the identification of coexpressed gene modules, we describe that these are complementary relationship. The reason of this statement is 2-fold: Because a gene coexpression can be interpreted as a ‘fingerprint’ of the underlying transcriptional network, it can be used to compare 2 physiological states of cellular systems. That is, differential coexpression approach leads to a systematic difference in transcriptomic levels among such organs. Secondly, the comparative coexpression study here provides a way to use the observed coexpressions to find additional information about the plant transcriptome. It is obvious that identifying coexpressed gene modules is not sufficient, because differentially coexpressed gene groups also reflect the changes in underlying transcriptional regulation.

We also demonstrated the existence of several pairs within LycoCyc pathway (enzymatic isoforms) that had significantly different coexpression between duplicated genes. In our proof-of-concept study, we investigated various patterns of differential coexpressions in 2 biosynthetic pathways associated with lycopene and flavonoid. In the lycopene biosynthesis pathway, we identified 4 significant differences in coexpression between leaves and fruits (Figure 8). Of the genes involved, psy encodes a key regulatory enzyme (Cazzonelli and Pogson, 2010). Positive coexpression of the psy and zds genes
in tomato fruits is highly likely from a proof-of-concept point of view regarding differential coexpression approaches. The differential coexpression between the 2 isoforms of the \textit{psy} gene may reflect organ-specific regulation on the transcription level. We found that 2 genes encoding the \( \varepsilon \)-ring hydroxylase and the SDR homolog, exhibited markedly different coexpression patterns in leaves and fruits. In fruits, the high coexpression (\( r = 0.73 \)) between the genes encoding \( \varepsilon \)-ring hydroxylase and SDR was partly consistent with the observation that 2 abscisic acid-deficient mutants, \textit{sitiens} and \textit{flacca}, exhibit a decrease in lutein levels (Galpaz et al., 2008). Our approach suggests that the correlation between ABA and lutein levels in tomato plants is attributable to their transcriptional coordination.

Based on transcriptome coexpression analysis, several flavonoid biosynthetic genes in Arabidopsis had characterized (Luo et al., 2007; Yonekura-Sakakibara et al., 2007; Yonekura-Sakakibara et al., 2008; Yonekura-Sakakibara et al., 2012). Flavonoids in plants play significant roles in many biological processes such as protecting plants against ultraviolet light, providing pigmentation in fruits and flowers, and in protecting against diseases and pests in higher plants (Buer et al., 2010). In the tomato, the total flavonoid concentration in roots is lower than that in leaves and fruits (Zornoza and Esteban, 1984); among the 3 organs, tomato fruits manifest the highest concentration of total flavonoids. Earlier studies had detected up to 70 flavonoids including naringenin chalcone, kaempferol, and quercetin in tomato fruits (cv. Micro-Tom) (Moco et al., 2006; Iijima et al., 2008). Our results showed strong coexpression patterns among genes encoding chalcone synthase, chalcone isomerase, and flavanone 3-hydroxylase in fruits (Figure 9). However, there was no coexpression between these genes in roots. In addition, we found a reversal of coexpression between genes encoding chalcone synthase and chalcone isomerase. This suggests a pronounced change in the underlying transcriptional regulation of the flavonoid pathway. Taken together, these observations suggest that at least in lycopene and flavonoid pathways, these differential coexpressions reflect a key regulatory mechanism among organs.

Because a marked change in coexpression patterns among organs may reflect underlying transcriptional changes directly or indirectly, the genes related to these differential coexpressions may act distinctly in those organs. Differential coexpression analysis represents a good data-mining approach to prioritizing candidate genes in further functional genomics studies on computationally
assigned genes encoding an enzyme, for example. Additional work is needed to evaluate the conservation of differential coexpression patterns among species, and more efficient tools for their assessment in plants would be helpful. For example, like Kappa-view 4 (Sakurai et al., 2011), the mapping of differential coexpressions onto metabolic pathways is highly useful in the field of plant research. We offer our extensive analysis of the coexpression networks to aid researchers in the selection and prioritization of candidate genes in studies on tomato functional genomics.

Materials and Methods

Plant Material and Growth Conditions

We developed a plant irradiation system based on applying light-emitting diodes (LEDs) to a leaf. We used this system to investigate changes in transcript profiles of tomato leaves and fruits grown under different light conditions. Seeds from tomatoes (*Solanum lycopersicum*, cv. Reiyo) were sown in 72-cell plant trays (Takii Seed, Japan) and grown in a commercial soil mix (Napura Soil Mixes, Yanmar, Japan) for 2 weeks, in a growth chamber (MKV DREAM, Japan) at 25°C/20°C (light/dark) and 900 ppm CO₂ concentration with a light/dark cycle of 16 h/8 h for 2 weeks at Chiba University, Matsudo, Japan. Then, seedlings were transferred to pots one by one (final size: 2.4 L) and grown in a growth chamber (Asahi Kogyosha, Japan). Photosynthetic photon flux (PPF) level in the growth chamber was adjusted to 450-500 μmol m⁻² s⁻¹ when we measured at the meristem of each tomato plant (light source: Ceramic metal halide lamps). After plant flowering in Summer 2010, we removed leaves and trusses from each plant, with the exception of the second truss, a leaf below the second truss and the meristem. LED irradiation at 0, 200, and 1000 μmol m⁻² s⁻¹ was directly applied to the leaf using a plant irradiation system (humidity, 70%; CO₂ concentration, 900 ppm). Plant material was harvested after 0, 1, and 2 weeks. Twenty samples were analyzed (14 leaf samples and 6 pericarp samples), and 2-3 biological replicates were used.

GeneChip Microarray Analysis

Analysis was performed using the Affymetrix GeneChip® Tomato Genome Array (Santa Clara, CA, USA) according to the manufacturer’s instructions. This microarray includes more than 9,000 transcripts and does not cover all of the
approximately 35,000 genes encoded in the tomato genome, which are located largely in euchromatic regions (Van der Hoeven et al., 2002). RNA was extracted using the standard procedure of Affymetrix. Our own data have been deposited in the Gene Expression Omnibus (GEO) (Barrett et al., 2011) and are accessible through the GEO series accession number GSE35020.

**Quantitative Real-Time PCR Analysis**
Total RNA was extracted using the RNeasy Plant Mini Kit (Qiagen, Valencia, CA, USA). Reverse transcription of each total DNase-treated (Qiagen) RNA sample was performed using the SuperScript III First-Strand Synthesis System for RT-PCR (Invitrogen, Carlsbad, CA, USA). qRT-PCR with the first-strand cDNA using the Fast SYBR Green Master Mix (Applied Biosystems, Foster City, CA, USA) was performed on the ABI StepOnePlus Real Time PCR system (Applied Biosystems). qRT-PCR primers used in the present study are listed in Supplementary Table S8.

**Data Collection and Preprocessing**
We collected 307 GeneChips from publicly available databases including the GEO, ArrayExpress (Parkinson et al., 2011), and Tomato Functional Genomics Database (TFGD) (Fei et al., 2011), and included 20 of our own hybridizations (GEO accession, GSE35020), resulting in a total of 327 arrays. The collected dataset contains 17 experiments (see also Figure 2 and Supplemental Data 1). The raw CEL data were normalized by the robust multichip average (RMA) (Irizarry et al., 2003) with Bioconductor (Gentleman et al., 2004). The resulting values were normalized to the same range by means-based scale normalization. Probe-sets with the prefix “RPTR” or “AFFX” were removed. Using the detection call (present/absent) from the MAS5 algorithm (http://www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf) with default settings, we excluded all probe-sets with absent calls in all samples. For the visual inspection of the quality of microarrays based on chip pseudo-images of residual weights and signed residuals, we used the R packages affy (Gautier et al., 2004) and affyPLM (Gentleman et al., 2005). We next performed a quality check of the microarrays based on the Kolmogorov-Smirnov goodness-of-fit statistic $D$ (Persson et al., 2005) and discarded all GeneChips with $D \geq 0.15$, resulting in 307 high-quality datasets for
Constructing Coexpression Networks and Topological Analyses

Using Pearson’s correlation coefficient (\( r \)), we calculated correlation matrices of sub-datasets classified by organ names (leaves, fruits, and roots) as well as “all data” (Figure 1). The number of arrays corresponding to the leaves, fruits, and roots were 77, 135, and 36, respectively. A correlation matrix was calculated for the remaining 9,989 probe-sets. All topological analyses, such as calculation of the Jaccard coefficient, were performed in R with the “igraph” package (Csardi and Nepusz, 2006). Coexpression networks were visualized using the Cytoscape software program (Shannon et al., 2003).

Graph Clustering and Functional Enrichment Analysis

To extract coexpression modules, we utilized the graph clustering algorithm IPCA (Li et al., 2008), an extension of the original DPclus algorithm (Altaf-Ul-Amin et al., 2006). DPclus is based on the combined periphery and density of graphs to extract dense subgraphs. The parameters we used were: modules with sizes smaller than 5 (\( S = 5 \)), a shortest path length of 2 (\( P = 2 \)), and a parameter “\( T_{in} \)” of 0.6 (\( T = 0.6 \)).

We used BiNGO (Maere et al., 2005), a tool to assess over- or under-representation in a set of genes, for the analysis of significantly over-represented GO categories among coexpression modules detected by graph clustering. We selected a GO term with the best \( p \)-value within the BP, CC, and MF domains. The Benjamini and Hochberg correction (Benjamini and Hochberg, 1995) for false discovery rate (FDR) was applied in functional enrichment analysis.

Calculating Differential Coexpressions

We used the term “differential coexpression” to describe significant correlations between given genes found under one-, but not another, condition. This definition also included instances in which the correlations were changed to the opposite direction under 2 conditions (for an example, see Figure 6). It provided a means to identify associations that were dramatically changed by mutations, organs, or treatments. For example, the correlation between gene expressions was calculated over the leaves sample, \( r_L \), and over the fruits sample, \( r_F \). In this case, the differential coexpression could be elucidated by testing the
null-hypothesis $H_0$: $r^L = r^F$. To test whether 2 coexpressions were significantly different from one another, the correlations were transformed by Fisher’s $Z$-transformation. If there were 2 correlations with sample sizes $n_1$ and $n_2$, respectively, they were each transformed into Fisher’s $z$ values, $z = 1/2[\ln(1+r)/(1-r)]$. Under the null hypothesis that the population correlations are equal, the $Z$ value, $Z = |z_1 - z_2|/\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}$, has approximately a normal distribution. We also calculated the local false-discovery rate for all correlation tests with the ‘fdrtool’ package (Strimmer, 2008).

**Definition of duplicated genes**
The tomato target sequences from the Affymetrix website (http://www.affymetrix.com/Auth/analysis/downloads/data/Tomato.target.zip) were utilized in an all-against-all BLASTX search (cutoff threshold, E-value < 1e-5). We used the Markov chain clustering (MCL) algorithm (http://micans.org/mcl/; Van Dongen, 2000) to assign the target sequences to clusters. The 1677 obtained clusters were regarded as gene families in this analysis.

**Gene Annotation**
Annotation information (Release 31) for tomato was downloaded from the Affymetrix website (http://www.affymetrix.com/Auth/analysis/downloads/na31/ivt/Tomato.na31.anno t.csv.zip). Gene ontology of tomato genes was based on information from TFGD (Fei et al., 2011). To assign the Sol Genomics Network (SGN) unigene (Mueller et al., 2005; Bombarely et al., 2011), similarity searches between Affymetrix probe-set ID and SGN unigene were performed using BLASTN (Altschul et al., 1990) with a threshold E-value of 1e-10. For functional categories of central metabolism, we used the MapMan flat file (Thimm et al., 2004) from the website (http://mapman.gabipd.org/web/guest/mapmanstore). We also used the SolCyc (LycoCyc) database (Bombarely et al., 2011) for the classification of metabolic genes.

**Acknowledgment**
We gratefully acknowledge the tomato research community for providing their
microarray data through NCBI GEO, ArrayExpress, and TFGD databases. We also thank Tetsuya Sakurai and Kenji Akiyama for computational assistance, Takashi Osanai for technical support for qRT-PCR instrument and Ursula Petralia for editing our manuscript.

Literature Cited


Galpaz N, Wang Q, Menda N, Zamir D, Hirschberg J (2008) Abscisic acid deficiency in the tomato mutant high-pigment 3 leading to increased plastid number and higher fruit lycopene content. Plant J 53: 717-730


synthetase 1:1 in coordinating metabolic balance in rice. Plant J


Figure Legends

Figure 1: Work-flow for extracting coexpression modules and for differential coexpression analyses among organs. Coexpression modules for each gene were generated by graph clustering without regard to functional properties. (1) Quality checks of microarrays were performed with robust regression
techniques and the Kolmogorov-Smirnov goodness-of-fit statistic $D$ (See Methods). We discarded 20 arrays with low-quality scores ($D \geq 0.15$). Probe sets with the prefix “AFFX” and “RPTR” were excluded. (2) We rejected 220 probes with the detection call “absent” across all samples.

**Figure 2:** Pie charts with a classification of the experiments and organs collected in this study. GeneChips (n = 327) were from publicly available databases including NCBI GEO, ArrayExpress, TFGD, and our own data (see Methods). The 17 experiments contained in the dataset were classified into 9 experimental (A) and 10 organ (B) categories.

**Figure 3:** Topological overview of the tomato coexpression network. A, The degree distribution of the network $P(k)$ at various correlation thresholds (Pearson’s correlation coefficient, $r$, ranging from 0.5 to 0.95) is shown; $k$ indicates connectivity and $P(k)$ indicates the connectivity distribution. B, A partial coexpression network ($r \geq 0.6$, $p < 2.1e-31$) in “all data”. The network shows how genes neighboring *LeMADS-Rin* (orange circle) correlate with each other. This undirected graph consists of nodes (red circles) and links (purple edges), indicating genes and positive correlations between genes.

**Figure 4:** Distribution of the number of unknown genes within a module.

**Figure 5:** Topological overview of organ-specific coexpression networks. A, The degree distribution of the coexpression network ($r \geq 0.6$) for 3 organs: leaves, fruits, and roots. B, Relationship between the average of the Jaccard coefficient (JC) and the correlation coefficient. The former can measure the degree overlap between 2 networks as the ratio of the intersection to the union of the networks.

**Figure 6:** A typical example of differentially coexpressed genes in different datasets. The correlation between the expression levels of 2 genes was significantly different in “leaves” and “fruits” (FDR = 9.27e-14). The axes represent relative gene expressions. Note that ‘differential coexpression’ was completely different from differential expression, and the mean level of given genes was significantly different between 2 organs (see Methods).

**Figure 7:** Distribution of significantly different coexpressed genes between
leaves and fruits (A), leaves and roots (B), and fruits and roots (C). This calculation was based on Fisher’s Z-transformation (see Methods). Thresholds correspond to Pearson’s correlation coefficient (PCC). Black and gray bars show the number of transitions from positive to negative correlations and from negative to positive correlations, respectively.

**Figure 8:** Differential coexpressions mapped onto the lycopene biosynthesis pathway. Solid arrows show the reaction steps. Orange dashed and green broken lines represent intensified correlations in fruits and leaves, respectively. L, leaves and F, fruits.

**Figure 9:** Differential coexpressions mapping to the flavonoid biosynthesis pathway. Solid arrows represent reaction steps. Orange dashed and green broken lines indicate intensified correlations in fruits and roots, respectively. CHS, chalcone synthase; CHI, chalcone isomerase; F3H, flavanone-3-hydroxylase; F, fruits; R, roots.

**Figure 10:** Quantitative real-time PCR (qRT-PCR) assessment of microarray results. A, Boxplots of expression values measured by qRT-PCR across 7 experimental conditions, including leaf and fruit samples. Six genes associated with biosynthetic pathways were involved in photosynthesis, *photosystem 1 reaction center protein subunit 2* (*psad*) [SGN-U580167] and *zeaxanthin epoxidase* (*zep*) [SGN-U569421]; flavonoids, *flavanone 3-hydroxylase* (*F3H*) [SGN-U563669] and *chalcone synthase 2* (*chs2*) [SGN-U580856]; and carotenoids, *phytoene synthase 1* (*psy1*) [SGN-U590947] and *zeta-carotene desaturase* (*zds*) [SGN-U568537]. B, Coexpression patterns between the genes, as determined by qRT-PCR (blue circles). Correlation between qRT-PCR and microarray analyses was also verified as shown in Supplemental Figure 3. w, week; L, leaves; F, fruits.

### Tables

**Table I:** Topological statistics of tomato coexpression networks ($r \geq 0.6$).

<table>
<thead>
<tr>
<th>Datasets</th>
<th># of nodes</th>
<th># of edges</th>
<th>Average path length</th>
<th>Clustering coefficient</th>
<th>Average degree</th>
<th>Degree exponent</th>
</tr>
</thead>
</table>

---

Downloaded from on January 6, 2018 - Published by www.plantphysiol.org
Copyright © 2012 American Society of Plant Biologists. All rights reserved.
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th>2.81</th>
<th>0.59</th>
<th>531.5</th>
<th>2.37</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leaves</td>
<td>9,952</td>
<td>2,644,732</td>
<td>2.81</td>
<td>0.59</td>
<td>531.5</td>
<td>2.37</td>
</tr>
<tr>
<td>Fruits</td>
<td>9,867</td>
<td>1,782,674</td>
<td>2.92</td>
<td>0.50</td>
<td>361.3</td>
<td>2.07</td>
</tr>
<tr>
<td>Roots</td>
<td>9,924</td>
<td>5,997,644</td>
<td>2.62</td>
<td>0.70</td>
<td>1208.7</td>
<td>2.29</td>
</tr>
<tr>
<td>All data</td>
<td>9,797</td>
<td>1,754,361</td>
<td>2.65</td>
<td>0.45</td>
<td>358.1</td>
<td>1.67</td>
</tr>
</tbody>
</table>

**Table II:** Postulated physiological function of coexpression modules. We show 15 selected modules with 5 or more genes. Annotations are based on GO functional categories (FDR < 0.05). See also Supplemental Table S1.

<table>
<thead>
<tr>
<th>Module #</th>
<th># of genes</th>
<th>Predominant function in biological process (best p-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>63</td>
<td>DNA endoreduplication (2.9e-2)</td>
</tr>
<tr>
<td>4</td>
<td>54</td>
<td>Photosynthesis (1.6e-3)</td>
</tr>
<tr>
<td>52</td>
<td>65</td>
<td>Response to organic substance (2.8e-2)</td>
</tr>
<tr>
<td>90</td>
<td>54</td>
<td>Protein metabolic process (2.2e-2)</td>
</tr>
<tr>
<td>107</td>
<td>57</td>
<td>Response to gibberellin stimulus (1.3e-2)</td>
</tr>
<tr>
<td>126</td>
<td>36</td>
<td>Cell cycle process (9.2e-3)</td>
</tr>
<tr>
<td>174</td>
<td>64</td>
<td>Pathogenesis (4.7e-2)</td>
</tr>
<tr>
<td>219</td>
<td>47</td>
<td>Protein farnesylation (4.6e-2)</td>
</tr>
<tr>
<td>291</td>
<td>49</td>
<td>Response to organic substance (3.6e-2)</td>
</tr>
<tr>
<td>309</td>
<td>19</td>
<td>Apoptosis (3.0e-2)</td>
</tr>
<tr>
<td>340</td>
<td>7</td>
<td>Jasmonic acid metabolic process (4.1e-3)</td>
</tr>
<tr>
<td>354</td>
<td>37</td>
<td>Response to ethylene stimulus (1.1e-2)</td>
</tr>
<tr>
<td>435</td>
<td>33</td>
<td>Response to cold (4.3e-2)</td>
</tr>
<tr>
<td>441</td>
<td>5</td>
<td>Primary root development (1.2e-4)</td>
</tr>
<tr>
<td>457</td>
<td>8</td>
<td>Jasmonic acid metabolic process (8.7e-3)</td>
</tr>
</tbody>
</table>

**Table III:** Distribution of GO terms for each module in the tomato dataset. Significance levels are set at FDR < 0.05. This list shows only the top 10 occurrences in this study. “freq” represents occurrences of the modules with specific GO terms. BP, biological process; CC, cellular component; MF, molecular function. Detailed information is shown in Supplemental Table S3.

<table>
<thead>
<tr>
<th>BP terms</th>
<th>freq</th>
<th>CC terms</th>
<th>freq</th>
<th>MF terms</th>
<th>freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>negative regulation of</td>
<td>150</td>
<td>chloroplast thylakoid</td>
<td>59</td>
<td>water channel activity</td>
<td>33</td>
</tr>
</tbody>
</table>
Table IV: Significantly over-represented GO terms associated with differential coexpressions among organs (FDR < 0.05) when the threshold is $r = 0.7$. Number in parenthesis indicates FDR. Detailed information is shown in Supplemental Table S4. BP, biological process; CC, cellular component; MF, molecular function; n.s., not significant.
### Table V: The number of coexpressions of duplicated genes with significant difference among organs. n.s., not significant. See also Supplemental Table S5.

<table>
<thead>
<tr>
<th># of significantly different pairs</th>
<th># of genes within duplicated gene cluster (Markov Cluster: MC)</th>
<th>GO terms “Biological Process” (p-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leaves vs. Fruits</td>
<td>9</td>
<td>12</td>
</tr>
<tr>
<td>Fruits</td>
<td></td>
<td>MC665</td>
</tr>
<tr>
<td></td>
<td></td>
<td>n.s.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Leaves vs. fruits</th>
<th>Leaves vs. roots</th>
<th>Fruits vs. roots</th>
</tr>
</thead>
<tbody>
<tr>
<td>positive</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BP</td>
<td>cell wall modification</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td></td>
<td>(7.37e-3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td>MF</td>
<td>enzyme inhibitor activity</td>
<td>Calcium ion binding</td>
<td>n.s.</td>
</tr>
<tr>
<td></td>
<td>(4.03e-7)</td>
<td>(3.15e-2)</td>
<td></td>
</tr>
<tr>
<td>negative</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BP</td>
<td>flower development</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td></td>
<td>(5.32e-4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC</td>
<td>n.s.</td>
<td>thylakoid</td>
<td>n.s.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(6.28e-3)</td>
<td></td>
</tr>
<tr>
<td>MF</td>
<td>enzyme inhibitor activity</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td></td>
<td>(4.21e-04)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leaf</td>
<td>Root</td>
<td>Gene</td>
<td>Function</td>
</tr>
<tr>
<td>------</td>
<td>------</td>
<td>------</td>
<td>----------</td>
</tr>
<tr>
<td>5</td>
<td>21</td>
<td>MC7</td>
<td>Cellular amino acid derivative biosynthetic process</td>
</tr>
<tr>
<td>4</td>
<td>12</td>
<td>MC893</td>
<td>Regulation of transporter activity</td>
</tr>
<tr>
<td>4</td>
<td>9</td>
<td>MC141</td>
<td>Brassinosteroid metabolic process</td>
</tr>
<tr>
<td>4</td>
<td>16</td>
<td>MC584</td>
<td>Cellular response to auxin stimulus</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>MC533</td>
<td>Cellular cell wall organization</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>MC1117</td>
<td>Oxidation reduction</td>
</tr>
<tr>
<td>3</td>
<td>13</td>
<td>MC813</td>
<td>Ciliary or flagellar motility</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>MC310</td>
<td>n.s.</td>
</tr>
</tbody>
</table>

Leaves vs. Roots

<table>
<thead>
<tr>
<th>Leaf</th>
<th>Root</th>
<th>Gene</th>
<th>Function</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>16</td>
<td>MC100</td>
<td>Photosynthesis, light harvesting</td>
<td>6.1e-5</td>
</tr>
<tr>
<td>6</td>
<td>21</td>
<td>MC7</td>
<td>Cellular amino acid derivative biosynthetic process</td>
<td>1.7e-12</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>MC10</td>
<td>Cellular response to auxin stimulus</td>
<td>2.4e-9</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>MC596</td>
<td>Response to cold</td>
<td>2.0e-3</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>MC748</td>
<td>Regulation of epidermal cell division</td>
<td>1.8e-2</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>MC1622</td>
<td>Ethylene metabolic process</td>
<td>4.5e-13</td>
</tr>
<tr>
<td>3</td>
<td>12</td>
<td>MC786</td>
<td>Phenylpropanoid biosynthetic process</td>
<td>1.3e-4</td>
</tr>
<tr>
<td>3</td>
<td>15</td>
<td>MC493</td>
<td>Regulation of protein metabolic process</td>
<td>1.6e-5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>----------------------</td>
<td>-----------------------------------------------------------------</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>MC19</td>
<td>cellular response to auxin stimulus (1.2e-6); auxin mediated signaling pathway (1.2e-6)</td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>16</td>
<td>MC100</td>
<td>photosynthesis, light harvesting (6.1e-5); protein-chromophore linkage (6.1e-5)</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>21</td>
<td>MC7</td>
<td>cellular amino acid derivative biosynthetic process (1.7e-12)</td>
<td>regulation of protein metabolic process (1.6e-5)</td>
</tr>
<tr>
<td>7</td>
<td>15</td>
<td>MC493</td>
<td>ubiquitin-dependent protein catabolic process (5.0e-3)</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>14</td>
<td>MC296</td>
<td>response to chitin (3.9e-15)</td>
<td>cellular response to hydrogen peroxide (4.4e-21); hydrogen peroxide catabolic process (4.4e-21)</td>
</tr>
<tr>
<td>4</td>
<td>12</td>
<td>MC181</td>
<td>brassinosteroid metabolic process (2.0e-6)</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>9</td>
<td>MC141</td>
<td>gluconeogenesis (4.2e-3)</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>MC19</td>
<td>cellular response to auxin stimulus (1.2e-6); auxin mediated signaling pathway (1.2e-6)</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>MC1622</td>
<td>ethylene metabolic process (4.5e-13)</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>MC641</td>
<td>gluconeogenesis (1.9e-3)</td>
<td></td>
</tr>
</tbody>
</table>

**Table VI:** The number of coexpressed pairs with significant differences between leaves and fruits in metabolic pathway-related genes. FDR < 1e-10. We used the LycoCyc database for classification of genes. See also Supplemental Table S6.
<table>
<thead>
<tr>
<th>Gene Number</th>
<th>ID</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>134</td>
<td>sucrose degradation to ethanol and lactate anaerobic</td>
</tr>
<tr>
<td>15</td>
<td>100</td>
<td>glycolysis IV (plant cytosol)</td>
</tr>
<tr>
<td>14</td>
<td>105</td>
<td>fructose degradation to pyruvate and lactate (anaerobic)</td>
</tr>
<tr>
<td>13</td>
<td>103</td>
<td>Calvin cycle</td>
</tr>
<tr>
<td>12</td>
<td>93</td>
<td>glycolysis I</td>
</tr>
<tr>
<td>12</td>
<td>90</td>
<td>glycolysis V</td>
</tr>
<tr>
<td>8</td>
<td>71</td>
<td>gluconeogenesis</td>
</tr>
<tr>
<td>8</td>
<td>78</td>
<td>glucose heterofermentation to lactate I</td>
</tr>
<tr>
<td>7</td>
<td>78</td>
<td>glycolysis II</td>
</tr>
<tr>
<td>6</td>
<td>71</td>
<td>glucose fermentation to lactate II</td>
</tr>
<tr>
<td>6</td>
<td>62</td>
<td>photorespiration</td>
</tr>
<tr>
<td>5</td>
<td>35</td>
<td>Entner-Doudoroff pathway II non-phosphorylative</td>
</tr>
<tr>
<td>5</td>
<td>35</td>
<td>Entner-Doudoroff pathway III semi-phosphorylative</td>
</tr>
<tr>
<td>5</td>
<td>38</td>
<td>UDP-glucose conversion</td>
</tr>
<tr>
<td>4</td>
<td>15</td>
<td>folate transformations</td>
</tr>
<tr>
<td>4</td>
<td>26</td>
<td>formylTHF biosynthesis I</td>
</tr>
<tr>
<td>4</td>
<td>38</td>
<td>formylTHF biosynthesis II</td>
</tr>
<tr>
<td>4</td>
<td>55</td>
<td>salvage pathways of purine and pyrimidine nucleotides</td>
</tr>
<tr>
<td>4</td>
<td>28</td>
<td>TCA cycle</td>
</tr>
<tr>
<td>4</td>
<td>25</td>
<td>TCA cycle variation VIII</td>
</tr>
<tr>
<td>3</td>
<td>20</td>
<td>asparagine degradation I</td>
</tr>
<tr>
<td>3</td>
<td>46</td>
<td>flavonoid biosynthesis</td>
</tr>
<tr>
<td>3</td>
<td>69</td>
<td>tRNA charging pathway</td>
</tr>
<tr>
<td>2</td>
<td>37</td>
<td>chlorophyllide a biosynthesis</td>
</tr>
<tr>
<td>2</td>
<td>11</td>
<td>glycine degradation I</td>
</tr>
<tr>
<td>2</td>
<td>41</td>
<td>jasmonic acid biosynthesis</td>
</tr>
<tr>
<td>2</td>
<td>18</td>
<td>phenylalanine degradation I</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
<td>reductive tricarboxylic acid cycle</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
<td>ribose degradation</td>
</tr>
<tr>
<td>2</td>
<td>38</td>
<td>triacylglycerol degradation</td>
</tr>
<tr>
<td>2</td>
<td>14</td>
<td>tyrosine degradation</td>
</tr>
</tbody>
</table>
1  4  &gamma;-glutamyl cycle
1  24  arginine biosynthesis II (acetyl cycle)
1  27  carotenoid biosynthesis
1  16  cysteine biosynthesis I
1  19  dTDP-L-rhamnose biosynthesis I
1  18  folate polyglutamylation I
1  17  formaldehyde assimilation I (serine pathway)
1  4  glutathione biosynthesis
1  10  glycine betaine degradation
1  11  glycine biosynthesis I
1  38  lipoxygenase pathway
1  9  ornithine biosynthesis
1  15  phenylpropanoid biosynthesis
1  72  purine nucleotides de novo biosynthesis I
1  39  respiration (anaerobic)
1  12  salvage pathways of purine nucleosides
1  16  salvage pathways of purine nucleosides II plant
1  39  suberin biosynthesis
1  39  sucrose biosynthesis
1  50  sucrose degradation I
1  33  sulfate assimilation III
1  13  superpathway of serine and glycine biosynthesis II

Supplemental Materials

Supplemental Table S1: List of 465 coexpression modules in the tomato dataset (“all data”). The modules were extracted with the graph clustering algorithm IPCA (see Methods). We used a Pearson correlation value of $r \geq 0.6$ as a cut-off threshold. The header over each module (orange color) includes the cluster number, the number of members, and significant gene ontology terms (BP, CC, and MF). Each column (green color) represents probe annotations such as the SGN-ID and the UniGene-ID. n.s., not significant.

Supplemental Table S2. List of unknown genes for each of the 465 modules in the tomato dataset “all data”.

**Supplemental Table S3:** Distribution of GO terms for each module in the tomato dataset “all data”. The level of significance is set at FDR < 0.05.

**Supplemental Table S4:** Significantly over-represented GO terms associated with differential coexpressions among organs (FDR < 0.05). The values in the parenthesis represent best p-value from GO enrichment analysis. n.s., not significant.

**Supplemental Table S5:** The number of coexpression of duplicated genes with significant differences between leaves and fruits. (A) “leaves” vs. “fruits”, (B) “leaves” vs. “roots”, and (C) “fruits” vs. “roots”. FDR < 1e-10.

**Supplemental Table S6:** The number of coexpressed pairs with significant differences between organs (FDR < 1e-10). We used the LycoCyc database for classification of genes.

**Supplemental Table S7:** The number of coexpressed pairs with significantly different coexpressions between duplicated genes (isoforms) within the LycoCyc database ver1.0. FDR < 1e-10.

**Supplemental Table S8:** qRT-PCR primers used in this study.

**Supplemental Figure S1:** The distribution of the expression matrix (A) and correlation coefficients (B) for the 3 organs are based on 9k genes calculated using the RMA algorithm.

**Supplemental Figure S2:** Correlation heatmaps for 4 datasets, “all data”, “leaves”, “fruits”, and “roots”. The values represent Pearson’s correlation coefficients.

**Supplemental Figure S3:** Correlation of expression values measured by quantitative real-time PCR (qRT-PCR) and microarray. r, Pearson’s correlation coefficient.
**Supplemental Data S1:** Meta-data for the microarrays used in this study.

**Supplemental Data S2:** List of differential coexpressions between the datasets of “leaves” and “fruits”, “leaves” and “roots”, and “fruits” and “roots”. FDR < 1e-10. Comparisons are (A) leaves vs. fruits [positive correlations ($r \geq 0.6$) in leaves and negative correlations ($r \leq -0.6$) in fruits], (B) leaves vs. fruits [negative correlations ($r \leq -0.6$) in leaves and positive correlations ($r \geq 0.6$) in fruits], (C) leaves vs. roots [positive correlations ($r \geq 0.6$) in leaves and negative correlations ($r \leq -0.6$) in roots], (D) leaves vs. roots [negative correlations ($r \leq -0.6$) in leaves and positive correlations ($r \geq 0.6$) in roots], and (E) fruits vs. roots [positive correlations ($r \geq 0.6$) in fruits and negative correlations ($r \leq -0.6$) in roots].
Figure 1: Work-flow for extracting coexpression modules and for differential coexpression analyses among organs. Coexpression modules for each gene were generated by graph clustering without regard to functional properties. (1) Quality checks of microarrays were performed with robust regression techniques and the Kolmogorov-Smirnov goodness-of-fit statistic $D$ (See Methods). We discarded 20 arrays with low-quality scores ($D \geq 0.15$). Probe sets with the prefix “AFFX” and “RPTR” were excluded. (2) We rejected 220 probes with the detection call “absent” across all samples.

Fukushima et al. Figure 1
Figure 2: Pie charts with a classification of the experiments and organs collected in this study. GeneChips (n = 327) were from publicly available databases including NCBI GEO, ArrayExpress, TFGD, and our own data (see Methods). The 17 experiments contained in the dataset were classified into 9 experimental (A) and 10 organ (B) categories.
Figure 3: Topological overview of the tomato coexpression network. A, The degree distribution of the network $P(k)$ at various correlation thresholds (Pearson’s correlation coefficient, $r$, ranging from 0.5 to 0.95) is shown; $k$ indicates connectivity and $P(k)$ indicates the connectivity distribution. B, A partial coexpression network ($r \geq 0.6$, $p < 2.1\text{e}-31$) in “all data”. The network shows how genes neighboring $\text{LeMADS-Rin}$ (orange circle) correlate with each other. This undirected graph consists of nodes (red circles) and links (purple edges), indicating genes and positive correlations between genes.

Fukushima et al. Figure 3
Figure 4: Distribution of the number of unknown genes within a module.

Fukushima et al. Figure 4
Figure 5: Topological overview of organ-specific coexpression networks. A, The degree distribution of the coexpression network ($r \geq 0.6$) for 3 organs: leaves, fruits, and roots. B, Relationship between the average of the Jaccard coefficient (JC) and the correlation coefficient. The former can measure the degree overlap between 2 networks as the ratio of the intersection to the union of the networks.
Figure 6: A typical example of differentially coexpressed genes in different datasets. The correlation between the expression levels of 2 genes was significantly different in “leaves” and “fruits” (FDR = 9.27e-14). The axes represent relative gene expressions. Note that ‘differential coexpression’ was completely different from differential expression, and the mean level of given genes was significantly different between 2 organs (see Methods).
Figure 7: Distribution of significantly different coexpressed genes between leaves and fruits (A), leaves and roots (B), and fruits and roots (C). This calculation was based on Fisher's Z-transformation (see Methods). Thresholds correspond to Pearson's correlation coefficient (PCC). Black and gray bars show the number of transitions from positive to negative correlations and from negative to positive correlations, respectively.
**Figure 8:** Differential coexpressions mapped onto the lycopene biosynthesis pathway. Solid arrows show the reaction steps. Orange dashed and green broken lines represent intensified correlations in fruits and leaves, respectively. L, leaves and F, fruits.

Fukushima et al. Figure 8
Figure 9: Differential coexpressions mapping to the flavonoid biosynthesis pathway. Solid arrows represent reaction steps. Orange dashed and green broken lines indicate intensified correlations in fruits and roots, respectively. CHS, chalcone synthase; CHI, chalcone isomerase; F3H, flavanone-3-hydroxylase; F, fruits; R, roots.

Fukushima et al. Figure 9
Figure 10: Quantitative real-time PCR (qRT-PCR) assessment of microarray results. A, Boxplots of expression values measured by qRT-PCR across 7 experimental conditions, including leaf and fruit samples. Six genes associated with biosynthetic pathways were involved in photosynthesis, photosystem 1 reaction center protein subunit 2 (psad) [SGN-U580167] and zeaxanthin epoxidase (zep) [SGN-U569421]; flavonoids, flavone 3-hydroxylase (F3H) [SGN-U563669] and chalcone synthase 2 (chs2) [SGN-U580856]; and carotenoids, phytoene synthase 1 (psy1) [SGN-U590947] and zeta-carotene desaturase (zds) [SGN-U568537]. B, Coexpression patterns between the genes, as determined by qRT-PCR (blue circles). Correlation between qRT-PCR and microarray analyses was also verified as shown in Supplemental Figure 3. w, week; L, leaves; F, fruits.