Mol. Cells *34*, 295-304, September 30, 2012
DOI/10.1007/s10059-012-0117-z
pISSN: 1016-8478 eISSN: 0219-1032

Molecules and Cells
http://molcells.org

Metabolic Module Mining Based on Independent Component Analysis in *Arabidopsis thaliana*

Xiao Han[1,3], Cong Chen[2,3], Tae Kyung Hyun[1,3], Ritesh Kumar[1], and Jae-Yean Kim[1,*]

Independent Component Analysis (ICA) has been introduced as one of the useful tools for gene-functional discovery in animals. However, this approach has been poorly utilized in the plant sciences. In the present study, we have exploited ICA combined with pathway enrichment analysis to address the statistical challenges associated with genome-wide analysis in plant system. To generate an *Arabidopsis* metabolic platform, we collected 4,373 Affymetrix ATH1 microarray datasets. Out of the 3,232 metabolic genes and transcription factors, 99.47% of these genes were identified in at least one component, indicating the coverage of most of the metabolic pathways by the components. During the metabolic pathway enrichment analysis, we found components that indicate an independent regulation between the isoprenoid biosynthesis pathways. We also utilized this analysis tool to investigate some transcription factors involved in secondary cell wall biogenesis. This approach has identified remarkably more transcription factors compared to previously reported analysis tools. A website providing user-friendly searching and downloading of the entire dataset analyzed by ICA is available at http://kimjy.gnu.ac.kr/ICA.files/slide0002.htm. ICA combined with pathway enrichment analysis might provide a powerful approach for the extraction of the components responsible for a biological process of interest in plant systems.

INTRODUCTION

The completion of *Arabidopsis thaliana* genome sequencing has made it possible to identify gene families at the whole genome scale. In addition, the rapid accumulation of genome-wide gene expression data has facilitated the construction of a co-expressed gene network. In this model plant, co-expression databases are currently widely used for the prediction of unknown gene function, gene targeting, and potential interaction partners, and further for our understanding of plant cellular systems (Fukushima et al., 2009; Mao et al., 2009; Obayashi and Kinoshita, 2009). These co-expression networks are based on the concept that genes with a similar function should have similar expression patterns under a large number of experimental conditions (Obayashi and Kinoshita, 2010). The Pearson's correlation coefficient (r) between two pairs of genes is the most widely used computational method to identity networks of co-expressed genes involved in specific processes. This method has previously been utilized by two groups to investigate the *Arabidopsis* genes required for cellulose synthesis and secondary cell wall formation using public microarray datasets (Brown et al., 2005; Persson et al., 2005). Although the Pearson's correlation coefficient is widely used for the detection of co-expression, it has only been applied to relatively small-scale collections of gene expression data because of its extremely high calculation cost (Obayashi and Kinoshita, 2009). In addition, the Pearson's correlation coefficient method also has limitations in determining relationships between transcription factors and their target genes. For example, a set of transcription factors involved in the regulation of secondary cell wall formation has been identified using genetic approaches (Zhong et al., 2008). However, the majority of these transcription factors were not found in previous reports that utilized the Pearson's correlation coefficient method (Brown et al., 2005; Persson et al., 2005). This difference in findings between co-expression and genetic approaches might be due to the highly specific regulation of transcription factors and/or due to the limitations of database construction. An alternative method, the graphical Gaussian model (GGM), uses partial correlations to determine the degree of correlation remaining after removing the effects of other genes. Gene network analyses using GGM have been performed with *Arabidopsis thaliana* but displayed a limited power in dissecting regulatory networks (Ma et al., 2007). Thus, the further development of powerful methods to balance a high performance level with a certain ease of calculation remains a challenge in bioinformatics.

To circumvent data deficiencies, various statistical methods, such as clustering algorithms (Ihmels et al., 2005), topology

[1]Division of Applied Life Science (Brain Korea 21-World Class University Program), Plant Molecular Biology and Biotechnology Research Center, Gyeongsang National University, Jinju 660-701, Korea, [2]Institute of Mitochondrial Biology and Medicine, The Key Laboratory of Biomedical Information Engineering of Ministry of Education, Xi'an Jiaotong University School of Life Science and Technology, Xi'an, China, [3]The authors contributed equally to this work.
*Correspondence: kimjy@gnu.ac.kr.

Received April 18, 2012; revised July 7, 2012; accepted July 9, 2012; published online September 7, 2012

Keywords: independent component analysis (ICA), isoprenoid biosynthesis pathway, lignin biosynthesis pathway, secondary cell wall biogenesis, transcription factor (TF)

networks (Wang et al., 2008) and matrix decomposition (Li et al., 2007; Zhang et al., 2010), have been proposed. In addition, the MapMan tool is commonly employed for the identification of multiple pathways. Within MapMan, *A. thaliana* genes are grouped into more than 200 hierarchical categories organized by gene annotation and functional category (Thimm et al., 2004). However, many genes, including unknown genes, could not be categorized (Pitzschke and Hirt, 2010). Therefore, this tool has limited utility for analyzing genes with functions that cannot be categorized (Pitzschke and Hirt, 2010; Yuan et al., 2008). In contrast, ICA is a useful tool for the identification of candidate genes that are involved in a particular pathway *via* pathway enrichment, which is based on each gene's expression profile (Chiappetta et al., 2004; Frigyesi et al., 2006; Kong et al., 2009; Saidi et al., 2004; Yonekura-Sakakibara et al., 2012).

These candidate genes include potential upstream regulators as well as pathway enzymes. In addition, ICA can propose the possibility of cross-talks between different metabolic pathways. Although ICA interpretation has to integrate profiling data with preexisting biological functions to support statistical analyses, the independent components established in this study provide a meaningful module to explore metabolic regulation in *Arabidopsis*.

Independent Component Analysis (ICA) statistically splits an input microarray dataset into independent components that correspond to putative biological processes. Under the assumption that noise represents Gaussian distributions, ICA identifies non-Gaussian, typically super-Gaussian components, in a sample space (Kong et al., 2008). ICA was first applied to microarray data obtained from yeast cell cycling and B-cell lymphomas (Liebermeister, 2002), and this method has also been exploited to elucidate altered transcriptional programs (Teschendorff et al., 2007) and genetic relationships (Meda et al., 2012) to extract biologically significant features (Yang et al., 2011) and, more recently, to explore a whole genome to locate a specific factor of interest from a transcriptomic dataset (Liu et al., 2012). In comparison with principal component analysis (PCA) and clustering-based methods, ICA outperforms these other leading methods in the analysis of biological associations and phenotype-pathway relationships (Teschendorff et al., 2007). This result indicates that ICA, in the gene expression context, is a useful tool for gene-functional discovery. Earlier implementations of ICA have been limited to the animal sciences only. However, the application of ICA to plant expression data has recently been successfully introduced to delineate the regulation of anthocyanin modification (Yonekura-Sakakibara et al., 2012).

Here, we used the ICA method to locate regulation modules for *Arabidopsis* based on the data from 4,373 Affymetrix ATH1 microarray experiments, which were submitted to the NASC database (Craigon et al., 2004). Based on the analysis of transcription factors clustering in the components and metabolic pathway enrichment, we have identified components functioning in the secondary cell wall biogenesis and isoprenoid biosynthesis pathways. In addition, the ICA displayed a crosstalk between stress-related pathways and the lignin biosynthesis pathway, revealing new candidate transcription factors involved in these pathways.

MATERIALS AND METHODS

Arabidopsis Affymetrix microarray data

Arabidopsis expression datasets (4,373) based on Affymetrix ATH1 were downloaded from the NASCArrays website (http://affymetrix.arabidopsis.info/narrays/help/usefulfiles.html). The expression profiles for 3,232 genes (1,705 genes from 278 metabolism pathways and 1,527 genes for transcription factors) were used for the independent component analysis (Supplementary Table S1).

Independent component analysis

To construct the ICA platform, the FastICA package was downloaded (http://research.ics.tkk.fi/ica/newindex.shtml). The ICA algorithm and its applicability to expression data have been described previously (Liebermeister, 2002). After normalization, the microarray data were changed to a proportion of the maximum expression from the 4,373 datasets. A matrix indicates the gene expression data, with the columns corresponding to genes and the rows corresponding to the samples, and determined the relative activity of each component in the expression profile. To calculate the ICA, we used the FastICA-algorithm (Hyvarinen and Oja, 2000). Based on the fixed-point iteration scheme, the algorithm maximizes non-Gaussianity as a determination of statistical independence, and the Newton iteration was then applied for this algorithm. The algorithm performs data whitening to generate new components unrelated by linearly transforming, such as the eigenvalue decomposition of the covariance matrix of the data. Finally, an iterative algorithm was performed with the function of Gaussian non-linearity using the symmetric approach, which estimates independent components in parallel.

Identification of significant metabolic pathways in the components

To identify metabolic genes from each component, a threshold with an absolute value of 2 was established and p-values were calculated using Fisher's exact test for 278 metabolic pathways. This test determined if the analysis resulted in the differential expression of genes encoding for enzymes that are present at a significant number out of the total enzymes in the pathway. Therefore, if the p-value was less than 0.05, the result was considered significant.

The gene ontology enrichment analysis was performed using the web-based database AmiGO (http://amigo.geneontology.org/cgi-bin/amigo/go.cgi). All 3,232 genes were included as a background set with each component gene cluster.

RESULTS AND DISCUSSION

Construction of the *Arabidopsis* metabolic platform using ICA

The ICA method to analyze transcriptomes is based on the hypothesis that a mixture of independent biological processes determines the gene expression profiles from the individual microarray experiments (Lee and Batzoglou, 2003). Although this method has been suggested to be a powerful tool for separating mixed independent signals in animal microarray data, it has been poorly employed in plant systems. Therefore, we collected the 4,373 Affymetrix *Arabidopsis* ATH1 microarray datasets deposited in the NASC database, which contain gene expression profiles from a variety of mutants, under different environmental conditions and at different developmental stages. To analyze *Arabidopsis* metabolic genes regulation, 1,705 metabolic genes from 278 metabolism pathways and 1,527 transcriptions factors were selected as reference genes (Fig. 1). The dimension of the data was reduced to 150 to maintain a variance of more than 95%. The algorithms are described in the

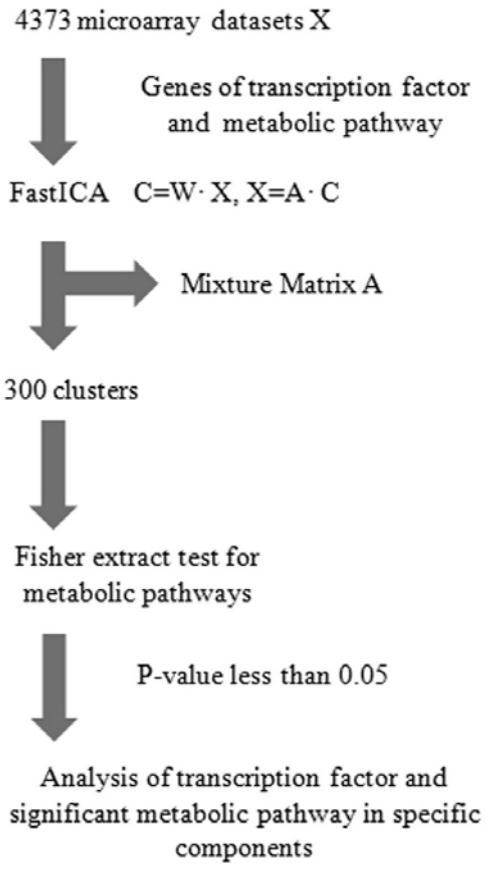

Fig. 1. The flow chart for the independent component analysis (ICA). From 4,373 Affymetrix ATH1 microarray datasets, the expression profiles of the 3,232 genes encoding enzymes and transcription factors were collected for ICA. The genes were clustered, and the significance of the metabolic pathways in each component was calculated based on Fisher's exact test.

Materials and Methods. The genes with absolute values greater than 2 on a given component were then retrieved to identify genes from independent components. The retrieved genes were subsequently considered to be a member of the given component. The negative loading and positive loading, which might act oppositely, were divided into two distinct groups (Supplementary Tables S2 and S3). Out of the 3,232 reference genes, 3,215 genes (99.47%) were identified from at least one component, suggesting that independent components covered the majority of the metabolic genes and transcription factors. The numbers of genes per component ranged from 0 to 267, with an average of 83 gene members. Several genes were assigned into at least two components. We also developed a website providing user-friendly searching and downloading for the whole dataset that was analyzed by ICA (http://kimjy.gnu.ac.kr/ICA.files/ slide0002.htm).

As was expected, the mixture matrix showed the state of each component among the 4,373 datasets (Fig. 2). Some of the components have limited activity among the 4,373 datasets, indicating that these components are specific for certain experimental condition(s). To identify the components active under a specific experimental condition, we set a threshold and analyzed our dataset using 5%, 10% and 25% cut-off values for maximum activity. Remarkably, with both 5% and 10% cut-off values, 5 components were active for less than 10 of the microarray datasets that were analyzed for gene expression profiles in heart and torpedo stage *Arabidopsis* embryos (Supplementary Table S4), whereas 24 components with 25% cut-off value were analyzed for gene expression profiles in torpedo stage embryos, roots, and under different stress conditions (Supplementary Table S5). Based on this comparison between cut-off values, we found that at the expense of the specificity, the number of independent components showing specific biological processes increases when higher cutoff-values are used.

Enrichment of metabolic pathways in the components

To evaluate the significance of ICA clustering, we performed a Fisher's exact test on all the components for the 278 metabolic pathways. The p-value for the over-represented metabolic pathways was recorded for each component. A pathway with a lower p-value indicated that its genes were not randomly clustered into each component, suggesting that these types of pathways are controlled by the same factors. Out of the 278 metabolic only 8 pathways are not significantly enriched. This result indicates that the inferred components were mapped to known pathways. Interestingly, our independent components covered almost all the metabolic pathways with significance. Among these pathways, 24 were co-regulated with more than half of their gene members (Table 1, Supplementary Table S6). For example, we found that all six genes involved in 'light reactions in photosynthesis' were clustered specifically into component P2. This finding suggests that plants design a regulation module to control basic metabolic processes.

Because co-expression network analyses only determine one linkage between genes, it is difficult to define complicated gene regulation networks in a genome. However, the ICA approach has overcome this problem, as it allows for the clustering of genes into more than one component. For example, NAD/NADH phosphorylation and dephosphorylation, which is an important biological process to generate the reducing power required for a variety of cellular processes, was enriched in 27 components co-regulated with other metabolic pathways (Supplementary Table S7). In addition, glycolysis I, glycolysis II, starch degradation, gluconeogenesis and aerobic respiration, which regulate the energy production pathways in plants, were also located in more than 18 components. *S*-adenosylmethionine (SAM) plays an important role in the methylation of DNA, RNA and proteins and acts as a cofactor for a wide range of biosynthetic processes (Palmieri et al., 2006). Thus, the SAM pathway is expected to be co-regulated with different metabolic pathways. Indeed, we found that the SAM pathway belongs to various components, supporting as a multiple correlation with other metabolic pathways (Supplementary Table S7). Furthermore, we found that the glutathione redox reactions were enriched in 20 components. This result indicates that plants have a delicate mechanism to maintain ROS homeostasis or to remove the harmful byproducts produced by many physiological reactions by co-regulating glutathione redox reactions. Taken together, these findings strongly suggest that ICA is a useful approach to address the multiple connections between metabolic pathways.

Comparison of ICA with the graphical Gaussian model (GGM) through the analysis of the isoprenoid biosynthesis pathway

To evaluate the independence of each component resulting from ICA, we compared ICA with the GGM method, one of the

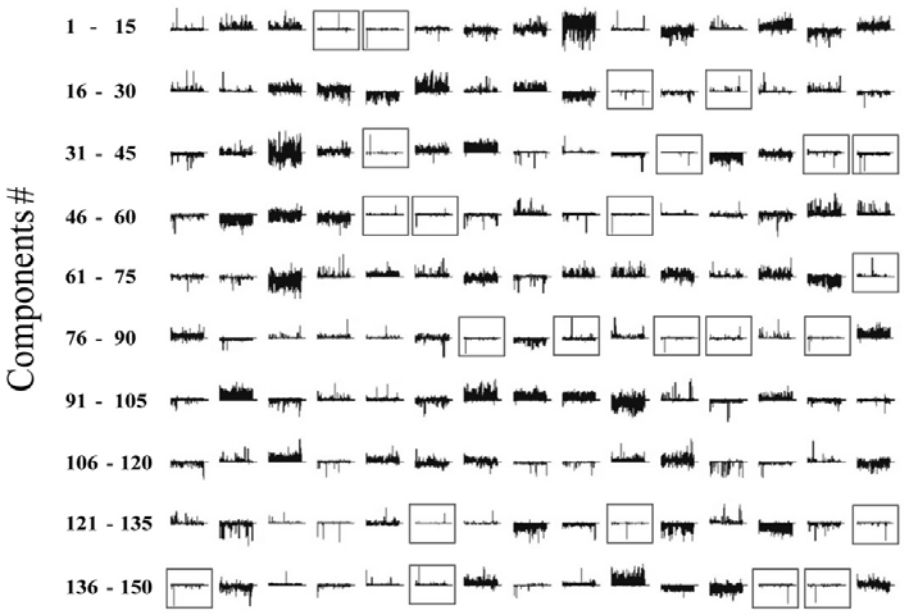

Fig. 2. Activities of the components in each microarray dataset. The subplots represent the activities of the 150 components in the 4,373 microarray datasets. The components that displayed specificity are highlighted with a red frame.

most powerful correlation analyzers. In higher plants, isoprenoids are synthesized through two distinct pathways from plastids and cytosols for the formation of isopentenyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP) (Laule et al., 2003). Cytoplasmic isoprenoids are synthesized from mevalonate (MVA)-derived IPP, whereas the conversion of methylerythritol phosphate (MEP)-derived IPP and DMAPP to isoprenoids occurs in plastids. Cytoplasmic IPP and DMAPP, which are synthesized *via* the MVA pathway, are the building blocks for the biosynthesis of sterols, sesquiterpenes and the ubiquinone side chain. Isoprenes, carotenoids and the side chains of chlorophyll and plastoquinone are produced from plastidal IPP/DMAPP *via* the MEP pathway. Although both pathways operate independently in plants, a co-regulation between the MVA and MEP pathway genes has been suggested earlier using the GGM approach (Wille et al., 2004). In a genetic regulatory network based on GGM, *HMGR1* was found to be negatively correlated to the module of connected genes in the MEP pathway (those encoding DXR, CMS, CMK, MCS and HDS) (Wille et al., 2004). However, the enhanced expression of *HMGR1S* did not affect the accumulation of carotenoids or chlorophylls (Munoz-Bertomeu et al., 2007).

When we applied ICA to provide insight into these pathways, we detected several components representing two independent pathways for isoprenoid biosynthesis with a p-value < 0.05. Series scatter graphs revealed the independence of these components' activities in the 4237 microarray datasets (Supplementary Fig. S1). Components P15, P21, P43, P64, N11 and N47 displayed significant enrichment in the MEP pathway (Fig. 3). In contrast, component N47 contained most of the MEP pathway genes, including *DXR, MCT, CMK, MECPS* and *HDR*. In addition, component P15 only controlled three downstream genes in the MEP pathway, *MECPS*, *HDS* and *HDR*, whereas upstream genes, such as *DXPS2* and *DXR*, are located in component P43. These differential distributions of genes indicated a modular regulation between up- and downstream genes in the MEP pathway. The MVA pathway includes branch point steps with unique enzymes and parallel steps with two isoenzymes. In contrast to the MEP pathway, the components separately control the branch point and parallel steps in the MVA pathway instead of the up- and down-stream modules (Fig. 3). Taken together, these findings suggest that gene regulation in the MVA pathway might be different compared to MEP pathway, indicating that these pathways operate independently.

Furthermore, our analysis indicated a crosstalk between the IPP pathways and other metabolic pathways. We found that a number of chlorophyll biosynthetic process genes, *CHLI2*, *CHLM*, *PCB2*, *CHLI1*, *G4*, *NTRC*, *ALB1*, *GUN5*, *PORB*, *ISPF*, *HEMC* and *CH1*, were significantly enriched in component N47 (Table 2). In addition, this component contained extra pigment metabolic genes, such as *CYP97A3, PDS3*, *ACD2*, *BETA-OHASE 1* and *LUT1*, indicating that the MEP pathway is co-regulated with the chlorophyllide a and carotenoid biosynthesis pathways. In fact, it has been shown that inhibition of the MEP pathway results in a reduction of carotenoid and chlorophyll levels in seedlings (Bick and Lange, 2003; Laule et al., 2003). Thus, co-regulation with the MEP pathways could be beneficial in coordinating the control of chlorophyll synthesis. Enhancing the MVA pathway through the overexpression of *Brassica juncea HMGS* up-regulates the genes involved in sterol biosynthesis (Wang et al., 2012), suggesting that the sterol biosynthesis pathway is downstream of the MVA pathway. In this context, a co-regulation between MVA and sterol synthesis was found in component P108, which contains sterol synthesis genes, such as *SMO2-2*, *BR6OX1*, *3BETAHSD/D1*, *SMT2*, *CYP51*, *G1*, *DWF1*, *FK*, *CPI1*, *HYD1* and *SMT3* (Table 2).

Identification of transcription factors as modulators of secondary cell wall biogenesis

Currently, cell wall synthesis and its regulatory genes, including those involved in cellulose and lignin biosynthesis, have been targeted for genetic modifications to increase biomass yield or to reduce the amount of lignin for improving enzymatic saccharification (Lee et al., 2010; Vanholme et al., 2008). Therefore, the identification of cell wall synthesis and regulatory genes and the understanding of these genes' regulations are necessary to

Table 1. Significant metabolic pathways in the components

Metabolic pathways	Component	p-value[a]	No. of metabolic pathway genes found / No. of all metabolic genes in the components[b]	Pathway genes in the *Arabidopsis* genome
Light reaction in photosynthesis	P2	0.00E + 00	6/82	6
Aerobic respiration	N74	0.00E + 00	24/74	27
Aerobic respiration - alternative oxidase pathway	N137	0.00E + 00	5/73	6
Cytokinins-O-glucoside biosynthesis	N15	0.00E + 00	4/48	5
tRNA charging pathway	N42	0.00E + 00	6/109	8
Cytokinins 7-N-glucoside biosynthesis	N44	0.00E + 00	6/50	8
Cytokinins 9-N-glucoside biosynthesis	P13	0.00E + 00	6/75	8
Glycolysis II (plant plastids)	N137	0.00E + 00	4/73	6
NAD/NADH phosphorylation and dephosphorylation	N68	0.00E + 00	4/71	6
Sucrose degradation to ethanol and lactate (anaerobic)	P65	0.00E + 00	4/99	6
TCA cycle	N42	0.00E + 00	6/109	9
Sucrose degradation	P78	0.00E + 00	8/63	12
Starch degradation	N31	0.00E + 00	5/66	8
Glycolysis I (plant cytosol)	N47	0.00E + 00	5/120	8
Methylglyoxal degradation	P140	0.00E + 00	8/102	13
UDP-glucose biosynthesis (from glucose 6-phosphate)	N42	0.00E + 00	7/109	12
Photorespiration	P60	0.00E + 00	4/91	7
Sterol biosynthesis	P60	0.00E + 00	4/91	7
Chorismate biosynthesis	N20	0.00E + 00	5/95	9
Chlorophyllide a biosynthesis	N42	0.00E + 00	5/109	9
Calvin cycle	P18	0.00E + 00	5/60	9
Glyoxylate cycle	P66	0.00E + 00	5/68	9
Lysine biosynthesis	P145	0.00E + 00	6/47	11
de novo biosynthesis of purine nucleotides	P2	0.00E + 00	7/82	13
Leucine degradation	N15	0.00E + 00	4/48	8
Fatty acid beta-oxidation I (saturated)	N47	0.00E + 00	5/120	10
Jasmonic acid biosynthesis	P80	0.00E + 00	7/103	14

[a]The p-value was calculated based on Fisher's exact test.
[b]The number of metabolic pathway genes found in the components was compared with the number of all the metabolic genes in the components. There were a total 1,705 metabolic genes in our analysis. The metabolic pathways in the list contain more than 4 genes.

facilitate functional studies for improving the production of biofuel from lignocellulosic biomass. In *Arabidopsis thaliana*, a regression method based on co-expression analysis has been used for the identification of genes that are potentially involved in the secondary cell wall synthesis pathway (Persson et al., 2005). Using a genetic approach, it has been shown that NST1 and NST3, which are NAC transcription factors, are key regulators for the formation of secondary walls in woody tissues (Mitsuda et al., 2007). Furthermore, SND1-regulated transcription factors (TFs), namely SND2, SND3, MYB103, MYB85, MYB52, MYB54 and KNAT7, are required for normal secondary wall biosynthesis in *Arabidopsis* (Zhong et al., 2008). Although these TFs are key factors for improving secondary wall biosynthesis, they were not found in a gene list generated using a large-scale co-expression approach (Ruprecht et al., 2011).

To test the effectiveness of ICA, we first focused on the phenylpropanoid pathway, which is most prominent in lignin synthesis. Two components, N31 and N105, were identified in the phenylpropanoid synthesis pathway and had the lowest p-values by Fisher's exact test. Component N105 contains phenylpropanoid pathway genes such as *4CL2*, *CYP98A3*, *PAL1*, *CCR1* and *ATCAD4*. However, component N31 contains more phenylpropanoid pathway genes than component N105 (*CAD5*, *CYP98A3*, *UGT76C2*, *PAL2*, *OMT1*, *C4H*, *PAL1*, *CCR1*, *HCT* *4CL1*, *UGT72E1*, *FAH1* and *4CL5*). In addition, we found that component N68 is significantly enriched in genes involved in the phenylpropanoid and cellulose synthesis pathways, such as the cellulose synthase (CESA) genes, e.g., *CESA2*, *CESA4*, *CESA5*, *CESA7* (*IRX3*) and *CESA8* (*IRX1*) (Table 3). In *Arabidopsis thaliana*, the CESA genes are placed into two subgroups. The *CESA1*, *CESA2*, *CESA3*, *CESA5* and *CESA6* genes belong to group I and are required for primary cell wall synthesis, whereas the *CESA4*, *CESA7* and *CESA8* genes are in group II and play essential roles in secondary cell wall synthesis (Persson

Table 2. Metabolic pathways and gene ontology terms in the components that are involved in the isoprenoid biosynthesis pathways

Component	Metabolic pathway	p-value	GO term	p-value
P108	Sterol biosynthesis	2.95E-06	GO:0006694 steroid biosynthetic process* (*SMO2-2*, *BR6OX1*, *3BETAHSD/D1*, *SMT2*, *CYP51*, *G1*, *DWF1*, *FK*, *CPI1*, *HYD1*, and *SMT3*)	7.76E-05
	GDP-L-fucose biosynthesis I	1.17E-03		
	UDP-D-xylose biosynthesis	3.10E-03		
	Cholesterol biosynthesis	4.62E-03		
	Fatty acid biosynthesis - initial steps	5.28E-03		
	Aerobic respiration	6.70E-03		
N47	Chlorophyllide a biosynthesis	4.30E-10	GO:0015995 chlorophyll biosynthetic* (*CHLI2*, *CHLM*, *PCB2*, *CHLI1*, *G4*, *NTRC*, *ALB1*, *GUN5*, *PORB*, *ISPF*, *HEMC*, and *CH1*)	2.98E-07
	Histidine biosynthesis	7.51E-05		
	Lysine biosynthesis	8.44E-04		
	tRNA charging pathway	1.95E-03	GO:0046148 pigment biosynthetic* (*CYP97A3*, *CHLI2*, *CHLM*, *PCB2*, *CHLI1*, *PDS3*, *G4*, *BETA-OHASE 1*, *NTRC*, *ALB1*, *GUN5*, *PORB*, LUT1, *ISPF*, *HEMC*, and *CH1*)	1.13E-05
	Chlorophyll cycle	5.04E-03		
	Carotenoid biosynthesis	5.23E-03		
	Calvin cycle	6.58E-03		
P21	Photosynthesis, light reaction	1.61E-03	GO:0035304 regulation of protein dephosphorylation (*PSAD-1, GLK2, ZFP8, PIF4, AT1G76110, COL4, GPRI1, PSAD-2, HPR, PSAH2, PSBO2, PMDH2,* and *PSBY*)	8.56E-07
	Methylerythritol phosphate pathway	4.83E-03		
	Carotenoid biosynthesis	4.83E-03		
	Histidine biosynthesis	2.70E-02		
	Monolignol glucoside biosynthesis	3.40E-02		
	Glutathione redox reactions	4.98E-02		
	Glucosinolate biosynthesis from homomethionine	4.98E-02		
P64	Tyrosine degradation	2.71E-03	GO:0045087 innate immune response (*AT5G38710, TRP1, TAT3, ALDH2C4, SUS1, UGT85A1, ELI3-1, PAD3, CYP71A13, AT5G24210, ACX1, AT3G48080, NAC042, NAP, NAC083, NAC019, PARVUS, ACD1, ADT4, PGDH, UGT87A2, BGL2, JMT, PMZ, EFE, HDS* and *OBP2*)	1.30E-05
	Ethylene biosynthesis from methionine	7.58E-03		
	Camalexin biosynthesis	1.29E-02		
	Phaseic acid biosynthesis	2.47E-02		
	IAA biosynthesis I	2.53E-02		
	Methylerythritol phosphate pathway	3.31E-02		
	Vitamin E biosynthesis	3.94E-02		
	Homogalacturonan degradation	4.25E-02		
P43	Methylerythritol phosphate pathway	5.59E-03		
	Phaseic acid biosynthesis	6.98E-03		
	Homogalacturonan degradation	1.23E-02		
	Starch degradation	1.70E-02		
	Nitrate assimilation pathway	3.74E-02		
	Homogalacturonan biosynthesis	4.96E-02		
P15	Methylglyoxal degradation	1.31E-03		
	Glucosinolate biosynthesis from phenylalanine	1.49E-03		
	Glucosinolate biosynthesis from tryptophan	1.49E-03		
	Molybdenum cofactor biosynthesis	2.52E-03		
	Homogalacturonan degradation	5.65E-03		
	Methylerythritol phosphate pathway	1.05E-02		
	Chlorophyll degradation	1.08E-02		
	Non-oxidative branch of the pentose phosphate pathway	1.35E-02		
	Monolignol glucosides biosynthesis	1.38E-02		
	Ketoglutarate dehydrogenase complex	2.55E-02		

(continued)

Component	Metabolic pathway	p-value	GO term	p-value
N11	Methylglyoxal degradation	1.31E-03		
	Glucosinolate biosynthesis from phenylalanine	1.49E-03		
	Glucosinolate biosynthesis from tryptophan	1.49E-03		
	Molybdenum cofactor biosynthesis	2.52E-03		
	Homogalacturonan degradation	5.65E-03		
	Methylerythritol phosphate pathway	1.05E-02		
	Chlorophyll degradation	1.08E-02		
	Non-oxidative branch of the pentose phosphate pathway	1.35E-02		
	Monolignol glucosides biosynthesis	1.38E-02		
	Ketoglutarate dehydrogenase complex	2.55E-02		
P59	Mevalonate pathway	2.18E-04	GO:0009833 primary cell wall biogenesis (*CESA1, CESA6, CESA2* and *CEV1*)	9.00E-04
	Cellulose biosynthesis	3.89E-04		
	Coniferin metabolism	3.34E-03		
	Camalexin biosynthesis	3.34E-03		
	Ureide degradation	3.34E-03		
	Cytokinins 7-N-glucoside biosynthesis	1.36E-02		
	Cytokinins 9-N-glucoside biosynthesis	1.36E-02		
	Cytokinins-O-glucoside biosynthesis	1.36E-02		
	Sterol biosynthesis	1.53E-02		
	Glucosinolate biosynthesis from tryptophan	1.56E-02		
	Glutamine biosynthesis	2.14E-02		
	Lysine degradation II	2.79E-02		
	Nitrate assimilation pathway	3.51E-02		
	IAA biosynthesis I	4.30E-02		
	Ammonia assimilation cycle	4.30E-02		
P28	Mevalonate pathway	1.81E-03	GO:0048437 floral organ development (*MYB24, PI, TCP24, RGL1, BPEp, PDF2, AP2, CPD, RGL2, SEP3, AP3, TCP3, AP1, BEL1, TCP4, MYB106* and *ARF8*)	6.11E-04
	Acetyl-CoA biosynthesis (from pyruvate)	7.57E-03		
	Fatty acid biosynthesis - initial steps	9.55E-03		
	Ketoglutarate dehydrogenase complex	9.55E-03		
	Keto acid dehydrogenase complex	2.18E-02		
	Epicuticular wax biosynthesis	2.67E-02		
	& alpha;-amyrin biosynthesis	2.67E-02		
P127	Glutamate degradation II	4.30E-04		
	4-aminobutyrate degradation I	2.38E-03		
	Acetyl-CoA biosynthesis (from pyruvate)	6.07E-03		
	Nitrate assimilation pathway	7.55E-03		
	Mevalonate pathway	1.04E-02		
	Cytokinins 7-N-glucoside biosynthesis	1.06E-02		
	Cytokinins 9-N-glucoside biosynthesis	1.06E-02		
	Cytokinins-O-glucoside biosynthesis	1.06E-02		
	Putrescine biosynthesis *via* carbamoylputrescine	1.32E-02		
	Chorismate biosynthesis	2.23E-02		

The p-value was calculated using Fisher's exact test for metabolic pathways and hypergeometric probability for the GO term.
*Component genes enriched in these biological processes following the GO term.

et al., 2005). *CESA4*, *CESA7*, *CESA8* and the genes of the phenylpropanoid biosynthesis pathway were assembled in component N68, suggesting that this component is one of the well identified modules within the network involved in secondary cell wall formation. Furthermore, the GO enrichment analysis indicated that this component contains many additional genes for polysaccharide metabolic processes, such as galacturonosyltransferase 12 (GAUT12) and galacturonosyltransferase-like

Table 3. Gene ontology terms in the components of the lignin pathway

Component	GO Term	p-value	Sample	Background	Genes
N68	GO:2000652 regulation of secondary cell wall biogenesis	4.34E-10	10/113 (8.8%)	13/3220 (0.4%)	*AtMYB103*, *KNAT7*, *NAC010*, *NAC073*, *MYB20*, *MYB43*, *MYB52*, *MYB58*, *MYB63*, *MYB85*
N68	GO:0042546 cell wall biogenesis	1.51E-18	23/113 (20.4%)	50/3221 (1.6%)	*PARVUS*, *CESA5*,*CEV1*, *CESA2*, *CESA4*, *CESA7*, *CESA8*, *MYB52*, *MYB43*, *KNAT7*, *MYB58*, *MYB20*, *NST1*, *NAC073*, *NAC066*, *MYB46*, *MYB85*, *ERF38*,*GAUT12*, *NAC010*, *MYB63*, *NAC012*, *MYB103*
N31	GO:0009607 response to biotic Stimulus	8.78E-05	20/120 (16.7%)	158/3674 (4.3%)	*WRKY27*, *EFE*, *CYP83B1*, *NHO1*, *WRKY48*, *AOS*, *AT5G61890*, *WRKY11*, *TGA3*, *4CL1*, *TTR1*, *WRKY53*, *DHS1*, *LACS2*, *NodGS*, *WRKY60*, *OBP2*, WRKY17, *TGA4*, *FAAH*

Table 4. Comparison of ICA with other methods using transcription factors involved in secondary cell wall synthesis

Method	Number of TFs	Matched TFs[a]	p-value
Persson	3	3/18	1.38e-6
GGM network	7	4/18	4.62e-7
Markov chain graph	10	6/18	2.17e-10
ICA	37	14/18	4.34e-21

The p-value was calculated based on hypergeometric probability.
[a] Number of transcription factors found by each method/number of transcription factors found in the experiments.

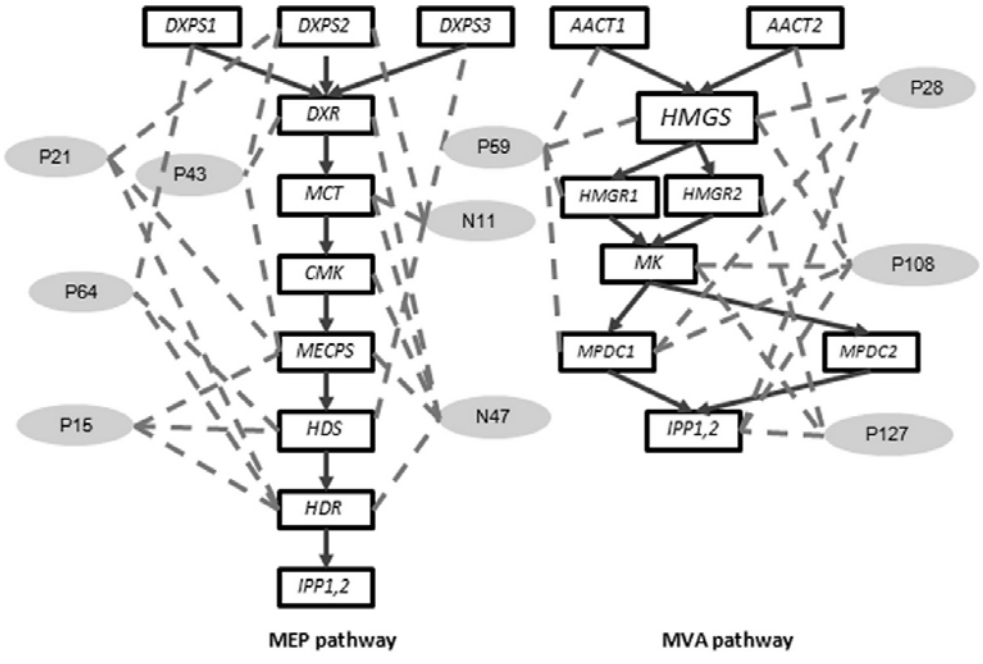

Fig. 3. Genes involved in the MEP and MVA pathways are enriched in different components.

1 (PARVUS) (Table 3), which are also involved in plant cell wall synthesis. A scatter plot displays the independence of these components in all the microarray datasets, and we found at least 3 different regulation points on the lignin synthesis pathway (Supplementary Fig. S2). The GO enrichment analysis showed that component N31 contains WRKY TFs, *WRKY27*, *WRKY48*, *WRKY11*, *WRKY53*, *WRKY60*, *WRKY17*, and other genes that respond to biotic stimuli (Supplementary Fig. S3 and Table 3), indicating that these genes might be involved in pathogen-induced lignin biosynthesis. Currently, 18 genes have

been identified as TFs involved in secondary cell wall biogenesis (Bhargava et al., 2010; Brown et al., 2005; McCarthy et al., 2009; Mitsuda and Ohme-Takagi, 2008; Mitsuda et al., 2007; Zhong et al., 2006; 2008). Remarkably, 14 out of the 18 TFs have been identified in component N68 (Supplementary Fig. S4 and Table 4), whereas the co-expression method, GGM network and Markov chain graph method were only able to indentify 3 to 6 TFs (Ma et al., 2007; Mentzen and Wurtele, 2008; Persson et al., 2005). These findings strongly suggest that the ICA approach, which had the lowest p-value, is a powerful tool to locate specific regulators of metabolic pathways.

In this study, we designed an ICA approach, which had the lowest p-value of all the tested methods, for the extraction of component(s) corresponding to a particular interest. Comparative analyses with GGM indicated that the ICA approach could efficiently separate linked pathways such as the MEP-chlorophyll and MVA-sterol pathways. In addition, we have shown that the ICA approach is a useful tool not only to analyze co-regulated metabolic pathways but also to identify TFs controlling metabolic pathways.

Note: Supplementary information is available on the Molecules and Cells website (www.molcells.org).

ACKNOWLEDGMENTS

This work was supported by the World Class University program (grant R33-10002) through the National Research Foundation of Korea, which is funded by the Ministry of Education, Science and Technology, and by a grant from the Next-Generation BioGreen 21 Program (SSAC, grant PJ008109), Rural Development Administration, Republic of Korea.

REFERENCES

Bhargava, A., Mansfield, S.D., Hall, H.C., Douglas, C.J., and Ellis, B.E. (2010). MYB75 functions in regulation of secondary cell wall formation in the Arabidopsis inflorescence stem. Plant Physiol. *154*, 1428-1438.

Bick, J.A., and Lange, B.M. (2003). Metabolic cross talk between cytosolic and plastidial pathways of isoprenoid biosynthesis: unidirectional transport of intermediates across the chloroplast envelope membrane. Arch. Biochem. Biophys. *415*, 146-154.

Brown, D.M., Zeef, L.A., Ellis, J., Goodacre, R., and Turner, S.R. (2005). Identification of novel genes in Arabidopsis involved in secondary cell wall formation using expression profiling and reverse genetics. Plant Cell *17*, 2281-2295.

Chiappetta, P., Roubaud, M.C., and Torresani, B. (2004). Blind source separation and the analysis of microarray data. J. Comput. Biol. *11*, 1090-1109.

Craigon, D.J., James, N., Okyere, J., Higgins, J., Jotham, J., and May, S. (2004). NASCArrays: a repository for microarray data generated by NASC's transcriptomics service. Nucleic Acids Res. *32*, D575-D577.

Frigyesi, A., Veerla, S., Lindgren, D., and Hoglund, M. (2006). Independent component analysis reveals new and biologically significant structures in micro array data. BMC Bioinformatics *7*, 290.

Fukushima, A., Kusano, M., Redestig, H., Arita, M., and Saito, K. (2009). Integrated omics approaches in plant systems biology. Curr. Opin. Chem. Biol. *13*, 532-538.

Hyvarinen, A., and Oja, E. (2000). Independent component analysis: algorithms and applications. Neural Netw. *13*, 411-430.

Ihmels, J., Bergmann, S., Berman, J., and Barkai, N. (2005). Comparative gene expression analysis by differential clustering approach: application to the *Candida albicans* transcription program. PLoS Genet. *1*, e39.

Kong, W., Vanderburg, C., Gunshin, H., Rogers, J., and Huang, X. (2008). A review of independent component analysis application to microarray gene expression data. Biotechniques *45*, 501-520.

Kong, W., Mou, X., Liu, Q., Chen, Z., Vanderburg, C.R., Rogers, J. T., and Huang, X. (2009). Independent component analysis of Alzheimer's DNA microarray gene expression data. Mol. Neurodegener. *4*, 5.

Laule, O., Furholz, A., Chang, H.S., Zhu, T., Wang, X., Heifetz, P.B., Gruissem, W., and Lange, M. (2003). Crosstalk between cytosolic and plastidial pathways of isoprenoid biosynthesis in *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. USA *100*, 6866-6871.

Lee, S.I., and Batzoglou, S. (2003). Application of independent component analysis to microarrays. Genome Biol. *4*, R76.

Lee, S.H., Teramoto, Y., and Endo, T. (2010). Enhancement of enzymatic accessibility by fibrillation of woody biomass using batch-type kneader with twin-screw elements. Bioresour. Technol. *101*, 769-774.

Li, H., Sun, Y., and Zhan, M. (2007). The discovery of transcriptional modules by a two-stage matrix decomposition approach. Bioinformatics *23*, 473-479.

Liebermeister, W. (2002). Linear modes of gene expression determined by independent component analysis. Bioinformatics *18*, 51-60.

Liu, J., Ghassemi, M.M., Michael, A.M., Boutte, D., Wells, W., Perrone-Bizzozero, N., Macciardi, F., Mathalon, D.H., Ford, J.M., Potkin, S.G., et al. (2012). An ICA with reference approach in identification of genetic variation and associated brain networks. Front. Hum. Neurosci. *6*, 21.

Ma, S., Gong, Q., and Bohnert, H.J. (2007). An Arabidopsis gene network based on the graphical Gaussian model. Genome Res. *17*, 1614-1625.

Mao, L., Van Hemert, J.L., Dash, S., and Dickerson, J.A. (2009). Arabidopsis gene co-expression network and its functional modules. BMC Bioinformatics *10*, 346.

McCarthy, R.L., Zhong, R., and Ye, Z.H. (2009) MYB83 is a direct target of SND1 and acts redundantly with MYB46 in the regulation of secondary cell wall biosynthesis in Arabidopsis. Plant Cell Physiol. *50*, 1950-1964.

Meda, S.A., Narayanan, B., Liu, J., Perrone-Bizzozero, N.I., Stevens, M.C., Calhoun, V.D., Glahn, D.C., Shen, L., Risacher, S.L., Saykin, A.J., et al. (2012). A large scale multivariate parallel ICA method reveals novel imaging-genetic relationships for Alzheimer's disease in the ADNI cohort. Neuroimage *60*, 1608-1621.

Mentzen, W.I., and Wurtele, E.S. (2008). Regulon organization of Arabidopsis. BMC Plant Biol. *8*, 99.

Mitsuda, N., and Ohme-Takagi, M. (2008). NAC transcription factors NST1 and NST3 regulate pod shattering in a partially redundant manner by promoting secondary wall formation after the establishment of tissue identity. Plant J. *56*, 768-778.

Mitsuda, N., Iwase, A., Yamamoto, H., Yoshida, M., Seki, M., Shinozaki, K., and Ohme-Takagi, M. (2007). NAC transcription factors, NST1 and NST3, are key regulators of the formation of secondary walls in woody tissues of Arabidopsis. Plant Cell *19*, 270-280.

Munoz-Bertomeu, J., Sales, E., Ros, R., Arrillaga, I., and Segura, J. (2007). Up-regulation of an N-terminal truncated 3-hydroxy-3-methylglutaryl CoA reductase enhances production of essential oils and sterols in transgenic *Lavandula latifolia*. Plant Biotechnol. J. *5*, 746-758.

Obayashi, T., and Kinoshita, K. (2009). Rank of correlation coefficient as a comparable measure for biological significance of gene coexpression. DNA Res. *16*, 249-260.

Obayashi, T., and Kinoshita, K. (2010). Coexpression landscape in ATTED-II: usage of gene list and gene network for various types of pathways. J. Plant Res. *123*, 311-319.

Palmieri, L., Arrigoni, R., Blanco, E., Carrari, F., Zanor, M.I., Studart-Guimaraes, C., Fernie, A.R., and Palmieri, F. (2006). Molecular identification of an Arabidopsis S-adenosylmethionine transporter. Analysis of organ distribution, bacterial expression, reconstitution into liposomes, and functional characterization. Plant Physiol. *142*, 855-865.

Persson, S., Wei, H., Milne, J., Page, G.P., and Somerville, C.R. (2005). Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets. Proc. Natl. Acad. Sci. USA *102*, 8633-8638.

Pitzschke, A., and Hirt, H. (2010). Bioinformatic and systems biology tools to generate testable models of signaling pathways and their targets. Plant Physiol. *152*, 460-469.

Ruprecht, C., Mutwil, M., Saxe, F., Eder, M., Nikoloski, Z., and Persson, S. (2011). Large-scale co-expression approach to dissect secondary cell wall formation across plant species. Front. Plant

Sci. *2*, 23.
Saidi, S.A., Holland, C.M., Kreil, D.P., MacKay, D.J., Charnock-Jones, D.S., Print, C.G., and Smith, S.K. (2004). Independent component analysis of microarray data in the study of endometrial cancer. Oncogene *23*, 6677-6683.
Teschendorff, A.E., Journee, M., Absil, P.A., Sepulchre, R., and Caldas, C. (2007). Elucidating the altered transcriptional programs in breast cancer using independent component analysis. PLoS Comput. Biol. *3*, e161.
Thimm, O., Blasing, O., Gibon, Y., Nagel, A., Meyer, S., Kruger, P., Selbig, J., Muller, L.A., Rhee, S.Y., and Stitt, M. (2004). MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. Plant J. *37*, 914-939.
Vanholme, R., Morreel, K., Ralph, J., and Boerjan, W. (2008). Lignin engineering. Curr. Opin. Plant Biol. *11*, 278-285.
Wang, X., Dalkic, E., Wu, M., and Chan, C. (2008). Gene module level analysis: identification to networks and dynamics. Curr. Opin. Biotechnol. *19*, 482-491.
Wang, H., Nagegowda, D.A., Rawat, R., Bouvier-Nave, P., Guo, D., Bach, T.J., and Chye, M.L. (2012). Overexpression of *Brassica juncea* wild-type and mutant HMG-CoA synthase 1 in Arabidopsis up-regulates genes in sterol biosynthesis and enhances sterol production and stress tolerance. Plant Biotechnol. J. *10*, 31-42.
Wille, A., Zimmermann, P., Vranova, E., Furholz, A., Laule, O., Bleuler, S., Hennig, L., Prelic, A., von Rohr, P., Thiele, L., et al. (2004). Sparse graphical Gaussian modeling of the isoprenoid gene network in *Arabidopsis thaliana*. Genome Biol. *5*, R92.
Yang, L., Rau, M.H., Hoiby, N., Molin, S., and Jelsbak, L. (2011). Bacterial adaptation during chronic infection revealed by independent component analysis of transcriptomic data. BMC Microbiol. *11*, 184.
Yonekura-Sakakibara, K., Fukushima, A., Nakabayashi, R., Hanada, K., Matsuda, F., Sugawara, S., Inoue, E., Kuromori, T., Ito, T., Shinozaki, K., et al. (2012) Two glycosyltransferases involved in anthocyanin modification delineated by transcriptome independent component analysis in *Arabidopsis thaliana*. Plant J. *69*, 154-167.
Yuan, J., Zhu, M., Lightfoot, D.A., Iqbal, M.J., Yang, J.Y., and Meksem, K. (2008). In silico comparison of transcript abundances during *Arabidopsis thaliana* and *Glycine max* resistance to *Fusarium virguliforme*. BMC Genomics *9* (Suppl 2), S6.
Zhang, W., Edwards, A., Fan, W., Zhu, D., and Zhang, K. (2010). svdPPCS: an effective singular value decomposition-based method for conserved and divergent co-expression gene module identification. BMC Bioinformatics *11*, 338.
Zhong, R., Demura, T., and Ye, Z.H. (2006). SND1, a NAC domain transcription factor, is a key regulator of secondary wall synthesis in fibers of Arabidopsis. Plant Cell *18*, 3158-3170.
Zhong, R., Lee, C., Zhou, J., McCarthy, R.L., and Ye, Z.H. (2008) A battery of transcription factors involved in the regulation of secondary cell wall biosynthesis in Arabidopsis. Plant Cell *20*, 2763-2782.