RapaNet: A Web Tool for the Co-Expression Analysis of *Brassica rapa* Genes

Evolutionary Bioinformatics
Volume 13: 1–11

DOI: 10.1177/1176934317715421

Jiye Kim[1*], Kyong Mi Jun[2*], Joung Sug Kim[3], Songhwa Chae[3], Yoon-Mok Pahk[2], Tae-Ho Lee[4], Soo-In Sohn[4], Soo In Lee[4], Myung-Ho Lim[4], Chang-Kug Kim[4], Yoonkang Hur[5], Baek Hie Nahm[2,3] and Yeon-Ki Kim[3]

[1]Insilicogen, Inc., Suwon-si, Republic of Korea. [2]GreenGene Biotech Inc., Yongin, Republic of Korea. [3]Division of Biosciences and Bioinformatics, Myongji University, Yongin, Republic of Korea. [4]Department of Agricultural Biotechnology, National Institute of Agricultural Sciences, Jeonju, Republic of Korea. [5]Department of Biology, College of Biological Sciences and Biotechnology, Chungnam National University, Daejeon, Republic of Korea.

ABSTRACT: Accumulated microarray data are used for assessing gene function by providing statistical values for co-expressed genes; however, only a limited number of Web tools are available for analyzing the co-expression of genes of *Brassica rapa*. We have developed a Web tool called RapaNet (http://bioinfo.mju.ac.kr/arraynet/brassica300k/query/), which is based on a data set of 143 *B rapa* microarrays compiled from various organs and at different developmental stages during exposure to biotic or abiotic stress. RapaNet visualizes correlated gene expression information via correlational networks and phylogenetic trees using Pearson correlation coefficient (*r*). In addition, RapaNet provides hierarchical clustering diagrams, scatterplots of log ratio intensities, related pathway maps, and *cis*-element lists of promoter regions. To ascertain the functionality of RapaNet, the correlated genes encoding ribosomal protein (L7Ae), photosystem II protein D1 (psbA), and cytochrome P450 monooxygenase in glucosinolate biosynthesis (CYP79F1) were retrieved from RapaNet and compared with their *Arabidopsis* homologues. An analysis of the co-expressed genes revealed their shared and unique features.

KEYWORDS: *Brassica rapa*, microarray, network, L7Ae, psbA, CYP79F1

RECEIVED: March 3, 2017. **ACCEPTED:** May 24, 2017.

PEER REVIEW: Six peer reviewers contributed to the peer review report. Reviewers' reports totaled 1279 words, excluding any confidential comments to the academic editor.

TYPE: Original Research

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by grants from the Next-Generation BioGreen 21 Program (Y-KK, grant no. PJ011074012015; BHN, grant no. PJ011057032015), RDA, the Republic of Korea.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Yeon-Ki Kim, Division of Biosciences and Bioinformatics, Myongji University, Yongin 17058, Gyeonggi-do, Republic of Korea.
Email: kim750a11@gmail.com

Introduction

Brassica rapa comprises a variety of vegetables, including Chinese cabbage (*B rapa* ssp pekinensis), which is one of the most highly consumed vegetables throughout Eastern Asian countries, such as China, Japan, and Korea. The draft genome sequence of *B rapa* accession Chiifu-401-42, a Chinese cabbage, contains 41 174 protein-coding genes.[1] *Brassica rapa* is a member of the family Brassicaceae and considered a model plant for evolutionary research on genome polyploidy. *Brassica* species contain 3 diploids in the classical triangle of U, namely, *B rapa* (AA), *Brassica nigra* (BB), and *Brassica oleracea* (CC). The allotetraploids include *Brassica juncea* (AABB), *Brassica napus* (AACC), and *Brassica carinata* (BBCC).[2,3] The genome of *B oleracea*, which belongs to the same Brassicaceae family, has revealed detailed genome triplication because of its split from *Arabidopsis* approximately 20 million years ago (Mya).[4] The reshuffling of these triplicated genomic blocks was accompanied by massive gene loss during speciation, differential retention of the genes, and, presumably, changes in the regulation of gene expression.

Microarray analysis, one of the most commonly used methods for simultaneously measuring the expression levels of large numbers of genes, has helped biologists understand gene function on a genomic scale. Gene expression microarray data repositories, such as the Gene Expression Omnibus (GEO)[5] and ArrayExpress,[6] are publically available. In *B rapa* research, several microarrays have been used to study abiotic stress and Ogura cytoplasmic male sterility) on a genome-wide scale.[7,8] A microarray containing 90K expressed sequence tag consensus sequences from *B napus*, *B rapa*, and *B oleracea* was recently used to test genome-specific gene expression in *Brassica* species.[9] The recent accumulation of data obtained using next-generation sequencing technology provides insightful information at the nucleotide sequence level and many research opportunities.[10,11]

Many studies have attempted to obtain additional information from microarray data and to determine expression relationships between large numbers of genes. Several analysis tools and databases for analyzing co-expressed genes have been developed for individual model plants, such as *Arabidopsis* and rice, and for plants in general.[12-20] A co-expression analysis of the transcript abundance of developing seeds from 2 diverse *B rapa* morphotypes, specifically pak choi (leafy-type) and yellow sarson (oil-type), with 2 of their doubled haploid progenies, has been conducted using an Agilent array representing 42 000 genes.[21]

* J.K. and K.M.J. contributed equally to this work.

However, a Web tool for analyzing gene co-expression based on microarray data was not available at the time of the study.

In this report, we present the *B rapa* Array Network (RapaNet; http://bioinfo.mju.ac.kr/arraynet/brassica300k/query/), which was constructed using 143 *Brassica* microarrays based on the *Brassica rapa* 300k Microarray v2.0.[7] In RapaNet, the correlation levels between gene pairs are evaluated using Pearson correlation coefficient (r)[22] and are represented by a correlational network and phylogenetic tree. RapaNet provides hierarchical clustering diagrams, scatterplots of the log ratio intensities between 2 genes of interest, pathway maps, and *cis*-element lists of promoter regions. We examined the distributions of the r values between 47 548 unigenes on the microarray and found that the distributions followed a normal distribution and that 95% of the r values were within the range of −0.50 to 0.50. The co-expression of 3 representative genes was extracted from RapaNet and used to construct a gene regulatory profile: Brapa_ESTC000925 (L7Ae) encodes a chromosomally located ribosomal protein, Brapa_ESTC027946 (psbA) encodes photosystem II protein D1 and is located on the plastid chromosome, and Brapa_ESTC006895 (CYP79F1) encodes cytochrome P450 monooxygenase during the initial step of core glucosinolate (GS) biosynthesis. We also compared these co-expressed genes with their *Arabidopsis* homologues retrieved from CressExpress (cressexpress.org),[23] and the results show the common and unique features of the co-expressed genes and their homologous genes in the Brassicaceae family.

Materials and Methods

Microarray data normalization

To build the database, we used 143 microarrays generated using the *Brassica rapa* 300k Microarray v2.0.[7] Information on the microarray is available at NCBI GEO Platform GPL17248. Briefly, 10 probes (60 base pairs [bp] each) were extensively designed to cover the 3′ region of the coding sequence and the 3′ untranslated region. The probes were spaced 10 bp apart. In this manner, 507 495 oligonucleotides were designed from 47 548 consensus sequences. A BlastT analysis showed the 42 004 sequences matched to the 25 363 messenger RNAs (mRNAs) of *B rapa* genome v1.5 (brassicadb.org).

Microarray data sets typically have different levels of signal intensities and background noise depending on the conditions used in each experiment. To normalize the microarray data, the R statistical language and environment were used. First, XYS files were generated from Nimblegen pair files, and an annotation package for a Nimblegen microarray was built with the "pdInfoBuilder" package prior to normalization (bioconductor.org). The XYS files were subsequently loaded using the "read.xysfiles" function in the "oligo" package. The loaded microarray data were then normalized with the Robust Multi-Chip Analysis (RMA) function[24] in the Bioconductor package,[25] and the normalized data were exported into a database for further analysis using the DBI and RSQLite packages.

Evaluation of the expression relationships between genes

To determine the degree of correlation between 2 genes, RapaNet calculates Pearson correlation coefficient (r), the value of which ranges from +1 to −1. Specifically, r values of +1 and −1 indicate perfect positive and negative linear relationships between 2 genes, respectively, whereas an r value of 0 indicates the absence of a linear relationship between 2 genes. To assess the significance of an r value calculated for a gene pair, a P value is calculated numerically using the Student t test. In addition to the P value of the r value, a z score based on the actual distribution of r values is calculated. To obtain the z score of the r value for each gene in a gene pair, the r values between the first gene and the other genes is first calculated, and the z score is then calculated based on the distribution of the r values. Because the sample size is suitably large, most of the distributions of the r values calculated between one gene and other genes showed a normal distribution. The z score, which is based on the actual r value distribution, is calculated using the following numerical formula:

$$z = \frac{r - \mu}{\sigma}$$

where μ is the mean of the population, and σ is the standard deviation of the population. The closer the absolute z score is to 0, the lower the significance.

Development environment

RapaNet was constructed using Django (http://www.djangoproject.com/), a framework for Web application development that adopts an Model-View-Template (MVT) pattern in the Python programming language (http://www.python.org/). In the pattern, the model contains the essential fields and behaviors of the database based on class definitions. Each class definition is translated by Django via a single database table into an equivalent SQL. The view consists of a Python function that responds to HTTP requests for specific URLs. The Django template manages the display of information and provides various built-in tags and filters. The MVT supports multiple templates for the same model.

The Database Management System (DBMS) PostgreSQL (http://www.postgresql.org/), a popular open-source DBMS available for various operating systems, such as Linux, Mac, and Windows, was employed. PostgreSQL uses a multiple-row data storage strategy, which makes it extremely responsive in high-volume environments. In addition, mod_python, which integrates the Python programming language into the

Apache server, was used as the Apache HTTP server module (http://www.modpython.org/). To handle the huge number of calculations required in this study, the server programs for statistical analysis were written in the C programming language, which is 10 to 100 times faster than the Python language and uses less computer memory. The server program calculates the *r, P*, and *z* scores.

RapaNet presents the correlated gene information via the Internet. Users can easily search for genes of interest by keywords, sequences, gene IDs, or Gene Ontology (GO) terms and identify their relationships on the entry page. The relationships can be represented by correlational networks, phylogenetic trees, and/or hierarchical clustering diagrams. In addition, detailed scatterplots of the normalized log intensities, the *r* value distributions, related pathway maps, and the *cis* elements of promoter regions for gene pairs are provided by RapaNet.

Visualization of gene relationships

RapaNet provides phylogenetic trees, correlational networks, and hierarchical clustering diagrams for visualizing correlated gene information. To construct a phylogenetic tree, RapaNet calculates the evolutionary distance using a neighbor-joining method[26] and presents it via ATV (A Tree Viewer; http://www.phylosoft.org/archaeopteryx/) on the Internet. The Matplotlib package in Python is used to draw the correlational networks by RapaNet. The node and edge positions in the network are determined by the correlation values calculated by the C module, and the color of the edges varies depending on the *r* value. In addition, hierarchical clustering is performed on the normalized intensity data using the PyCluster package in Python, and the clustering diagrams are displayed by JAVA TreeView (http://jtreeview.sourceforge.net/).

In the correlational network, the edges represent the correlation between genes. A red edge indicates a positive correlation, whereas a green edge represents a negative correlation. As the color of an edge deepens and the absolute *r* value, |*r* value|, of the gene pair approaches 1, the closeness between the genes increases. The phylogenetic tree also shows the relationships among the correlated genes. The correlated genes are divided into subgroups based on their distance in the tree. In addition to correlational networks and phylogenetic trees, hierarchical clustering diagrams are provided by RapaNet, enabling users to examine gene expression patterns at a glance.

Comparison of commonly co-expressed genes

The genes co-expressed with *L7Ae*, Brapa_ESTC000925, and AT5G20160 were retrieved from the RapaNet and CressExpress Web sites, respectively. Because CressExpress gives r^2 values and RapaNet provides *r* values, the *r* values obtained using RapaNet are squared prior to comparison with the values found for *Arabidopsis* homologues with CressExpress. Analysis of the top 5% *r* values yielded 2372 among 47 548 unigenes from RapaNet and 1141 among 21 810 unigenes from CressExpress, and these unigenes were compared with yield the list of 336 *Arabidopsis* homologous genes in Supplementary 3.

Plant material

Brassica rapa L. (*B rapa* L. ssp *pekinensis* "Jangkang" AA, 2 m = 20) seeds were grown in 6.5 cm × 9 cm pots in a controlled culture room at 22°C under long-day conditions (16 hours light/8 hours dark) with white light (150 mol m^{-2} s^{-1}). The 20-day-old seedlings were transferred into growth chambers at 4°C for vernalization. After 70 days, the plants were transferred into larger pots (20 cm × 16 cm) and grown in a greenhouse with 16 hours of light (sunlight + halogen lamps) at 27 ± 3°C followed by 8 hours of darkness at 20 ± 4°C. The plants were regularly watered and fertilized. Approximately 200 seeds from each individual plant (or all the seeds if less than 200 were available) were sown in separate trays in the greenhouse.

RNA isolation and reverse transcription polymerase chain reaction

The total RNA from representative organs, such as the leaf, root, and flower, was extracted using the TRI Reagent (Molecular Research Center, www.mrcgene.com) and cleaned with the Qiagen RNeasy mini kit (Qiagen, www.qiagen.com). The complementary DNA (cDNA) templates were synthesized using the RevertAid H Minus M-MulV reverse transcriptase (Fermentas, www.fermentas.com). The semiquantitative reverse transcription polymerase chain reaction (RT-PCR) amplifications were performed in 20-mL volumes with the following protocol: 1 cycle of 95°C for 2 minutes and 25 to 30 cycles of 94°C for 30 seconds, 60°C for 30 seconds, and 72°C for 30 seconds.

A semiquantitative RT-PCR analysis was conducted using real-time PCR Pre-mix with EvaGreen (SolGent, www.solgent.com) in accordance with the manufacturer's recommended protocol. The *B rapa* actin gene (Brapa_ESTC001256) was used as an endogenous control. All the primer pairs are listed in Supplementary 12.

Results

Visualization of gene relationships in RapaNet

To construct the RapaNet database, we used 143 microarray data sets generated using the *B rapa* 300k microarray (NCBI GEO Platform GPL17248) as described in the "Materials and Methods" section. The sample information can be found in Supplementary 1. To determine the degree of correlation among the expression of 2 genes, RapaNet calculates Pearson correlation coefficient (*r*), which is widely used for measuring the correlation between 2 genes using normalized intensity values, as previously described.[17] To assess the level of significance, *r*, *P*, and *z* values were calculated numerically using Student *t* test. RapaNet was constructed with Django (http://

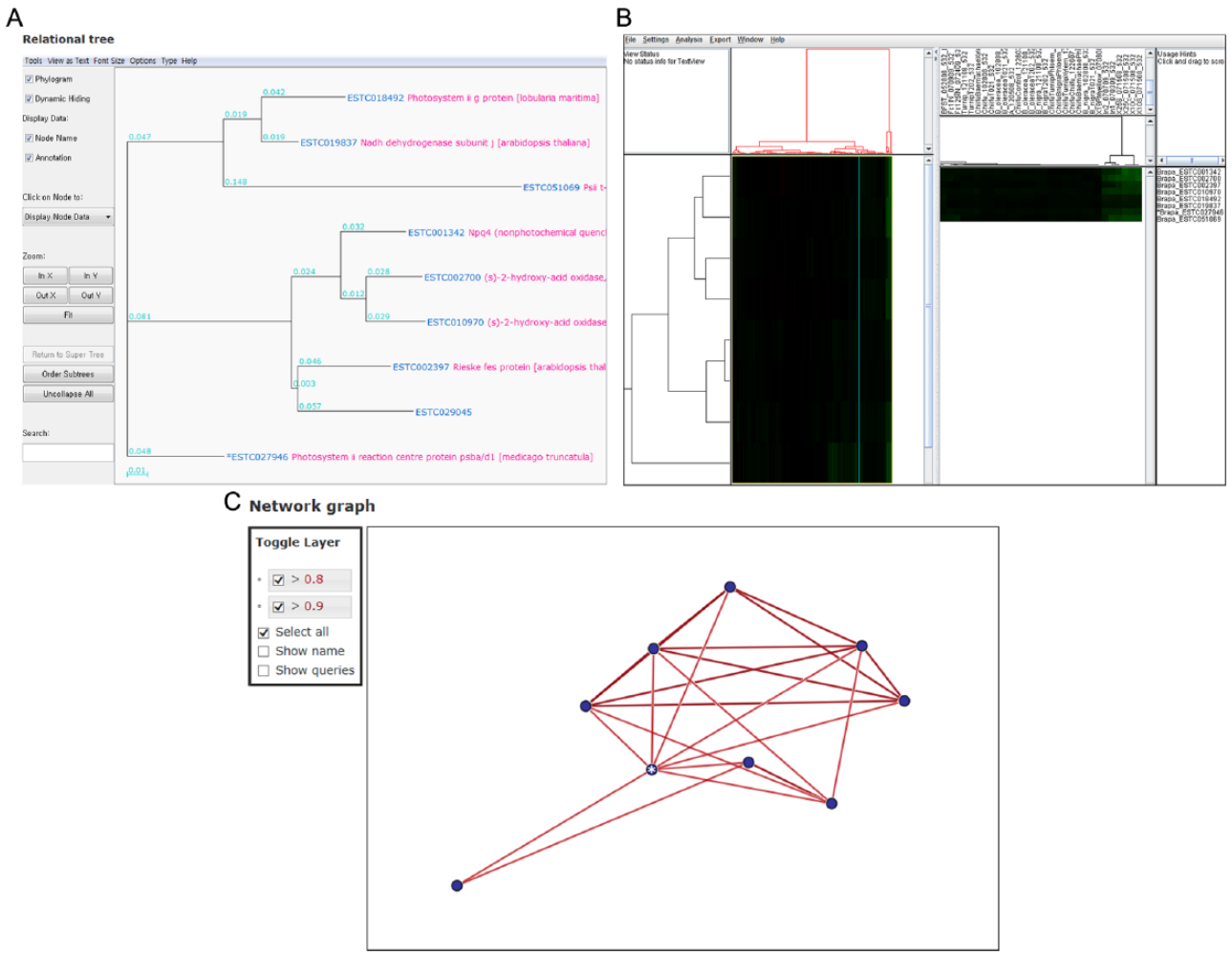

Figure 1. Visualization of the genes co-expressed with Brapa_ESTC027946, the *psbA* gene encoding for photosystem II protein D1. (A) The phylogenetic tree was drawn based on 9 *psbA*-related genes by setting |*r* value| ⩾ 0.8 and depth = 1 in RapaNet. The depth represents the degree of the direct relationship of the related gene to a query or a "seed" gene,[17] and a "0" indicates a direct relation to the seed gene. The query gene, *psbA*, is marked with an asterisk. The (B) hierarchical clustering diagram and (C) correlational network were drawn using highly correlated genes (8 genes), which were identified by increasing the *r* values to 0.8.

www.djangoproject.com/), a framework for Web application development in the Python programming language (http://www.python.org/, Supplementary 2). PostgreSQL (http://www.postgresql.org/) was used as the DBMS. The statistical analysis modules were written in the C programming language to handle the enormous number of required calculations. The gene expression correlations are represented in 3 different ways: correlational trees, networks, and hierarchical clustering. As an example, Brapa_ESTC027946, the *psbA* gene encoding photosystem II protein D1, was selected to analyze the gene regulatory network found by RapaNet (Figure 1A). In total, 9 genes were selected by setting the following criteria: absolute *r* value ⩾ 0.8 and depth = 1. A total of 1054 genes were selected by lowering the value to 0.5. The depth represents the degree of directness with the seeded gene. The correlation network is drawn with edges and circles based on the correlation levels. Each circle in the network corresponds to a gene, and the query gene is indicated with an asterisk. To avoid complicated visualization due to a prohibitive number of genes, the network and hierarchical clustering representations are derived from less than 200 genes, which are identified by increasing the *r* values. For example, for the images shown in Figure 1B and C, the *r* values were increased, and the network diagrams show the genes with *r* values greater than 0.8.

Distribution of r values and correlation between P values and z scores

Because potential bias can be associated with specific resource tools,[27] we tested the distributions of the *r* values of all the possible gene pairs of the 47 548 *B* genes in the RapaNet data set. The distribution of the *r* values follows a normal distribution with a mean (μ) *r* value of 0.003 and a standard deviation (σ) of 0.26 (Figure 2A). Furthermore, 95% of the *r* values were within the range of −0.50 to 0.50.

In RapaNet, 2 different indicators, *z* score and *P* value, are used to determine the statistical significance of the *r* values. The *P* value can be rapidly calculated because it is a type of

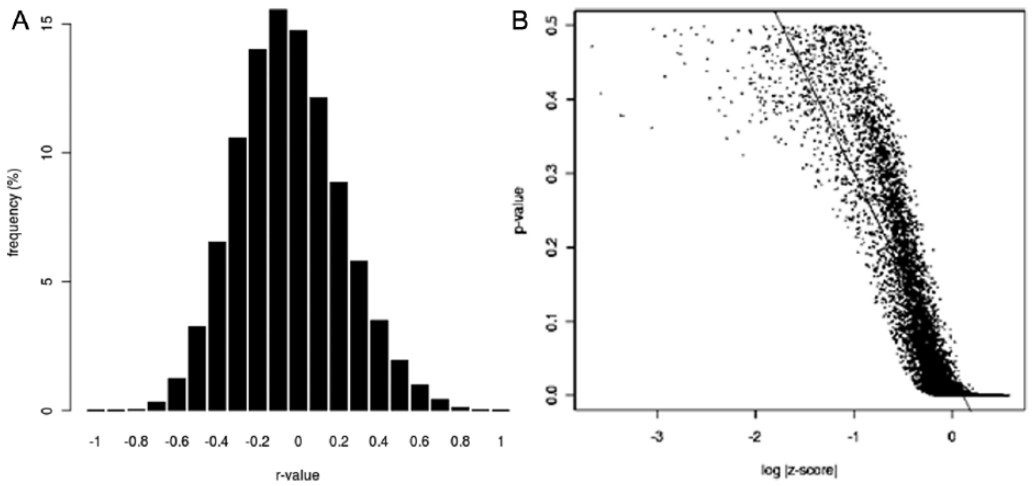

Figure 2. Distribution of the statistical values of the genes of the RapaNet microarray. (A) The distribution of the *r* values of all the genes in the microarray follows a normal distribution. The mean (μ) of the *r* values is 0.003, and the standard deviation (σ) is 0.26. (B) Scatterplots of the *P* values and *z* scores of 10 000 randomly selected gene pairs are shown.

statistical prediction, whereas the *z* score is calculated with the mean and standard deviation of the actual *r* value distributions and thus reflects the genuine character of the *r* values. As the absolute *z* score, |*z* score|, increases and the *P* value decreases, the significance of the *r* value increases. To compare the *P* value and absolute *z* score indicators, 10 000 gene pairs were randomly selected and drawn as scatterplots (Figure 2B). The scatterplots show that the log of the absolute *z* score, log |*z* score|, is inversely related to the *P* value. The linear model gives a slope of −0.90, indicating that the predicted *r* value distribution is similar to the actual *r* value distribution. Thus, the data show an unbiased distribution of the *r* values.

Analysis of the co-expressed genes encoding ribosomal protein L7Ae and photosystem II protein psbA

After evolving from their ancestor, *Arabidopsis*, several events occurred in the genomes of the Brassicaceae family, such as the triplication of their genomes and the reshuffling of genomic blocks. These events resulted in differential retention of genes of the Brassicaceae family, and the regulation of gene expression might have undergone changes during speciation. To test how genes in the Brassicaceae family are co-expressed with other genes, we selected several genes and compared them with their homologues in *Arabidopsis*. First, the ribosomal protein L7Ae, Brapa_ESTC000925, was selected, and its co-expressed genes were identified using RapaNet. The r^2 value of the genes was found to equal 0.11 (SD: 0.10, Figure 3A). To confirm the reliability of the regulatory network of L7Ae, we searched for the co-expressed genes of its *Arabidopsis thaliana* gene homologue, AT5G20160, in the CressExpress database (cressexpress.org).[23] The r^2 values were obtained with the options "1779 arrays processed with RMA normalization" using AT5G20160 (probe set id: 246070_at). The mean was found to equal 0.075 (SD: 0.10, the box plot on the right in Figure 3A [left]). Because each data set contains different number of probe sets, the number of co-expressed genes that ranked within 5% of L7Ae from *Arabidopsis* (1141) and *B rapa* (2372) were selected and compared with test the number of shared co-expressed genes in *Arabidopsis*. The results showed that 336 *Arabidopsis* genes were commonly co-expressed (Figure 3B and Supplementary 3). In contrast, Brapa_ESTC001256, an actin used as a control, has few genes in common (257) with any gene co-expressed with AT5G20160.

The *psbA* gene, which encodes photosystem II protein D1 and is located on the plastid chromosome, was selected for analysis of its gene regulatory network. The entire set of co-expressed genes was searched using Brapa_ESTC027946. The r^2 value of the genes was found to equal 0.048 (SD: 0.079, Figure 3C). The co-expressed genes of an *A thaliana* homologue, ATCG00020, were also retrieved from the CressExpress database. The mean of the r^2 values was found to equal 0.074 (SD: 0.11, box plot in Figure 3C). The comparison of the co-expressed genes that ranked within 5% of *psbA* from *Arabidopsis* (1140) and *B rapa* (2371) revealed that 578 *Arabidopsis* genes were shared with those of *B rapa* (Figure 3D and Supplementary 4). More than half of the genes co-expressed with *Arabidopsis* ATCG00020 were found in those co-expressed with the *B rapa* homologue, which suggests that the critically conserved genes involved in photosynthesis exhibit very similar profiles of co-expressed genes over the speciation process. Interestingly, among these 578 genes, only 13 genes consist of plastid DNA, whereas the others are located on nuclear chromosomes, which implies that the regulation of the genes in chromosomes and chloroplasts is tightly regulated. In addition, the genes retrieved with Brapa_ESTC001256, an actin, did not include any of the 578 genes co-expressed with *psbA* in *Arabidopsis* and *B rapa* (Figure 3D).

We also tested the functional categories of the genes co-expressed with *psbA* by GO term enrichment. RapaNet retrieved 1504 genes using a maximum absolute *r* value of 0.5,

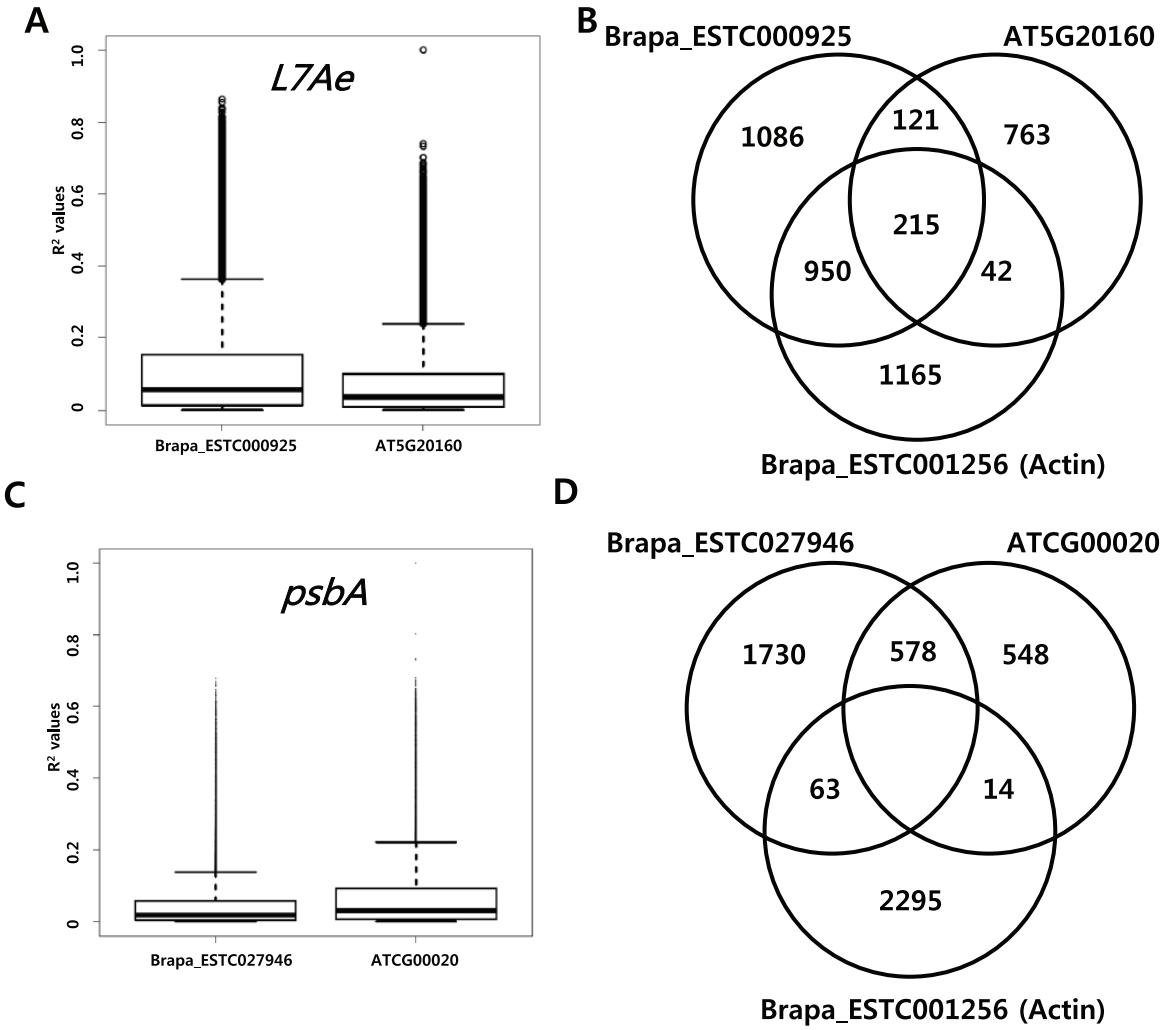

Figure 3. Distribution of the squared Pearson correlation coefficients of the co-expressed genes. (A) The r^2 values (leaf box plot) of the total genes (47 538) associated with Brapa_ESTC000925, an *L7Ae* gene encoding a ribosomal protein, were obtained by RapaNet. The mean is 0.11 (SD: 0.10). The r^2 values (right box plot) of the correlations of AT5G20160 with all of the genes (21 810) with a homologous gene in *Arabidopsis* were obtained using the 1779 RMA-normalized arrays included in the CressExpress database (cressexpress.org) with the default options (Probe Set ID: 246070_at). The mean is 0.075 (SD: 0.10) and the number of co-expressed genes among the top 5% of the genes. (B) The common co-expressed genes with the top 5% *r* values with *L7Ae* in *Brassica rapa* (2372) identified using RapaNet and *Arabidopsis* (1141) retrieved from CressExpress are shown in a Venn diagram (A). In total, 336 genes of the *Arabidopsis* homologues are shared between the species. In contrast, the genes co-expressed with Brapa_ESTC001256, an actin used as a control, have only 257 genes in common with those co-expressed with AT5G20160. The distribution of the squared Pearson correlation coefficients of the co-expressed genes is shown. (C) The genes were obtained from RapaNet and CressExpress using *psbA*, photosystem II protein D1. The mean is 0.048 (SD: 0.079) for the genes co-expressed with Brapa_ESTC027946 obtained from RapaNet, and the corresponding value for *Arabidopsis* homologue ATCG00020 (probe set id; 245047_at) obtained from CressExpress is 0.074 (SD: 0.11). (D) The 578 co-expressed genes that are ranked within 5% of *psbA* are shown. In contrast, Brapa_ESTC001256, an actin used as a control, does not share any genes with Brapa_ESTC027946 and ATCG00020 (578).

and these were analyzed using the Singular Enrichment Analysis (SEA) tool in agriGO.[28] The hierarchical tree graphs of the GO terms generated using the SEA tool showed that significant biological process GO terms were associated with *psbA*-related genes, as shown in Supplementary 5. Most of the positively related genes are associated with light reactivity in photosynthesis, as expected. In addition, the negatively *psbA*-related genes were not associated with any significant biological process GO term. Thus, the gene network obtained by RapaNet reflects the actual biological features of gene expression and can be used for constructing expression networks.

Unique profiling of a GS synthesis gene in B rapa and Arabidopsis

In addition, we profiled the co-expressed genes of the aliphatic GS biosynthesis pathway. Glucosinolates, a group of sulfur-rich secondary metabolites, play an important role in plant defense against herbivores and microbes. The 3 major GS groups are aliphatic, indolyl, and aromatic GS.[29] In total, 91 GS biosynthesis genes (Supplementary 6) and 11 transcription factors have been identified in the *B rapa* genome.[30,31] Among the GS biosynthesis–related genes, 59 genes were unambiguously identified in the microarray used in this study, and 11

genes were involved in the synthesis of short aliphatic GS using methionine as the substrate (Supplementary 7). The genes involved in methionine chain elongation include *BCAT4*, *MAM*, and *BCAT3*. The initial step of core GS biosynthesis is catalyzed by cytochrome P450 monooxygenase (CYP79F1) and forms aldoxime. Subsequently, aldoxime is used for conjunction, C-S cleavage, glucosylation and sulfation by monooxygenase (CYP83A1), lyase (C-S lyase) and UGT74B1, respectively. The side chains are then modified by oxidation, elimination, alkylation, or esterification through the action of FMOGS-OX15, AOP, GSL-OH, and BZO1, respectively.

Several aliphatic GS biosynthesis pathway genes were selected (11 genes), and a network was drawn by setting the absolute *r* value to greater than 0.71 and the depth of degree to 1 in RapaNet (Supplementary 8). The *r* values between the genes in the pathway are presented in Table 1. In total, 77 genes clustered into 4 groups (Supplementary 9). Cluster 1 contains the genes (66) in the pathway, including Brapa_ESTC010094, Brapa_ESTC006895, Brapa_ESTC026266, Brapa_ESTC006716, Brapa_ESTC002247, Brapa_ESTC011946, Brapa_ESTC022657, and Brapa_ESTC029776. The other 11 genes seeded with Brapa_ESTC034451, Brapa_ESTC007581, and Brapa_ESTC021288 were excluded from the main cluster (cluster 1) and 3 different clusters, suggesting that the regulation of gene expression diverged over time.

Brapa_ESTC006895 (CYP79F1) catalyzes the oxidation of amino acids to aldoximes, the initial step in GS biosynthesis. We compared the co-expressed genes of Brapa_ESTC006895 with those of an *Arabidopsis* homologue, AT1G16410. The genes were retrieved with CressExpress as described previously for the *psbA* gene. The mean *r* values for Brapa_ESTC006895 and AT1G16410 were found to equal 0.037 (SD: 0.05) and 0.038 (SD: 0.053), respectively (Figure 4A). The top 5% of the ranked genes within each library share 215 genes (Figure 4B). A GO analysis suggested that 389 terms related to biological process are enriched in this set (data not shown), including GO:0010439_regulation of GS biosynthetic process, GO:0019760_GS metabolic process, and GO:0019761_GS biosynthetic process. In addition, we tested all r^2 values of the GS pathway genes (59) with Brapa_ESTC006895 in *B rapa* and directly compared them with those of the corresponding homologues of CYP79F1 in *Arabidopsis* (Supplementary 10 and Supplementary 11). The resulting linear model produces the line $y = 0.58x + 0.20$. The *P* value for the slope is $2.25*10^{-7}$, which indicates high significance. These data suggest that the regulation of the genes co-expressed with CYP79F1 was likely conserved during the speciation of the Brassicaceae family.

The *r* values for the co-expression of the genes involved in the synthesis of the short aliphatic GS that were retrieved with Brapa_ESTC006895 (CYP79F1) ranged from −0.16 with Brapa_ESTC034451 to 0.83 with Brapa_ESTC026266 (Table 1). We validated the 11 genes involved in the pathway by semi-quantitative RT-PCR using the total RNA from representative organs, such as the leaf, root, and flower (Figure 5). Similar to the results obtained for Brapa_ESTC006895, their expression appears to be more consistent when the *r* values are high and less consistent when the *r* values are low. The *r* values associated with Brapa_ESTC010094 and Brapa_ESTC026266 are 0.82 and 0.83, respectively, and their expression patterns are consistent across all of the organs. In contrast, the *r* values associated with Brapa_ESTC034451 and Brapa_ESTC007581 were found to equal −0.16 and 0.32, respectively, and their expression patterns differed significantly from those of Brapa_ESTC006895.

Discussion

The accumulation of genome-wide gene expression data from microarrays and next-generation sequencing provides an opportunity to derive an understanding of the relationships between the genes involved in the various biological systems of an organism.[10,11,14,23] Given the wide spectrum of databases available, specific types of gene regulation and function can potentially be ascertained through the correlation of gene expression profiles. Although RNA-Seq technology offers more advantages regarding this than microarrays; co-expression analysis is in its infancy and requires further refinement. In this article, we present RapaNet, a tool for analyzing the co-expression of *B rapa* genes based on 143 data sets. RapaNet uses Pearson correlation coefficients for calculating edge weights. Multiple seeded genes can be selected, and the edges can be further extended by setting the number of degrees to range from 0 to 2, resulting in clusters. The clustered genes can then be visualized via network-like, tree-like, and cluster-shaped modules (Figure 1).

We used the RapaNet database to study the co-expression patterns of *L7Ae*, *psbA*, and *CYP79F1*. L7Ae is a ribosomal protein, and psbA and CYP79F1 are involved in primary and secondary metabolism, respectively. A comparison of *L7Ae* with an *Arabidopsis* homologue showed that approximately 27% (303/1141) of the *Arabidopsis* genes are included in the list of *L7Ae* co-expressed genes, and this number is comparable with that found for rice, a monocot plant.[17] This number is greater than the 225 genes identified from a comparison with the genes co-expressed with an actin (Brapa_ESTC001256) and is markedly higher than those found from a randomized list (30) of these genes (data not shown), suggesting that the co-expressed genes of homologous genes are comparable throughout the plant kingdom (Figure 3). This finding was further confirmed using a gene, *psbA*, involved in primary metabolism. Of the genes co-expressed with *psbA* (top 5%) in *Arabidopsis*, more than half (578/1141) are shared with the genes co-expressed with *psbA* in *B rapa*. Interestingly, although the *psbA* gene is located on the chloroplast genome, 13 of the 578 genes are located on the chloroplast genome, and the remaining genes are located on the chromosome (data not shown). It has been proposed that anterograde signals

Table 1. Correlation coefficients (*r* values) between the genes involved in the synthesis of short aliphatic glucosinolate.

	BRAPA_ ESTC010094	BRAPA_ ESTC021288	BRAPA_ ESTC006895	BRAPA_ ESTC026266	BRAPA_ ESTC006716	BRAPA_ ESTC002247	BRAPA_ ESTC011946	BRAPA_ ESTC034451	BRAPA_ ESTC022657	BRAPA_ ESTC007581	BRAPA_ ESTC029776	BRAPA_ ESTC001256 (ACTIN)
Brapa_ ESTC010094		−0.19	0.82	0.71	0.65	0.42	0.39	−0.01	0.47	0.28	0.80	−0.06
Brapa_ ESTC021288			0.06	−0.05	−0.13	−0.07	0.23	−0.05	0.004	−0.09	−0.19	0.16
Brapa_ ESTC006895				0.83	0.77	0.49	0.60	−0.16	0.50	0.32	0.80	0.10
Brapa_ ESTC026266					0.82	0.65	0.56	0.001	0.60	0.20	0.81	0.16
Brapa_ ESTC006716						0.73	0.51	−0.35	0.44	0.24	0.78	0.19
Brapa_ ESTC002247							0.60	−0.06	0.62	0.09	0.53	0.20
Brapa_ ESTC011946								0.06	0.74	0.16	0.43	0.01
Brapa_ ESTC034451									0.27	−0.17	−0.18	−0.09
Brapa_ ESTC022657										0.11	0.47	−0.08
Brapa_ ESTC007581											0.34	−0.01
Brapa_ ESTC029776												0.00

Among the GS biosynthesis–related genes, 59 genes were unambiguously identified in the microarray, and 11 genes identified in the synthesis of the short aliphatic glucosinolate and their gene correlations are shown. Brapa_ ESTC001256 (an actin), a gene that is not involved in the pathway, is included for comparison purposes.

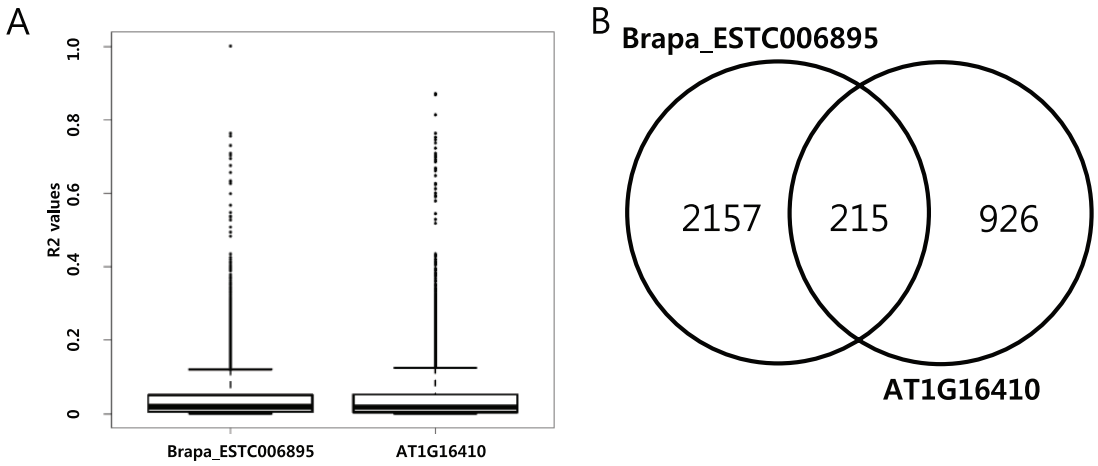

Figure 4. Distribution of the squared Pearson correlation coefficients of all of the genes and the shared genes. (A) The distribution of the squared Pearson correlation coefficients of all of the genes co-expressed with Brapa_ESTC006895 and AT1G16410 is box plotted. These genes are involved in catalyzing the oxidation of amino acids to aldoximes, the initial step in the short aliphatic glucosinolate pathway. The values were obtained as described above from RapaNet and CressExpress using Brapa_ESTC006895 and AT1G16410, respectively. The mean r^2 values for Brapa_ESTC006895 and AT1G16410 equal 0.037 (SD: 0.05) and 0.038 (SD: 0.05), respectively. (B) The top 5% of the ranked genes within each library were retrieved, and 215 genes were shared between Brapa_ESTC006895 and AT1G16410.

Figure 5. Expression of the 11 genes involved in the synthesis of the short aliphatic glucosinolate in representative organs. The total RNA from the leaf, root, and flower was analyzed by semiquantitative reverse transcription polymerase chain reaction. The digits represent Pearson correlation coefficients for the correlations of Brapa_ESTC006895, which catalyzes the oxidation of amino acids to aldoximes, with the genes involved in the synthesis of short aliphatic glucosinolate. L indicates leaf; F, flower after pollination; FB, flower before pollination; R, root.

originating from the nucleus and retrograde signals from the chloroplast are used to achieve coordinated gene expression.[32] As a part of photosystem II, the expression of *psbA* might be regulated in both the nucleus and chloroplast to ensure the correct stoichiometric subunit composition of these complexes and to efficiently regulate the genes involved in the synthesis of other metabolites. The *psbA* data described in this article suggest that the regulation of gene expression is tightly controlled through anterograde and retrograde signals. In contrast, none of the genes retrieved with Brapa_ESTC001256, an actin, are shared with the genes co-expressed with *psbA* in *Arabidopsis* under the same conditions. These findings indicate that the co-expression analysis performed using RapaNet is consistent with the existing GO database. The genes co-expressed with *psbA* were retrieved using RapaNet and subjected to SEA using agriGO, and most of the positively related genes were found to play relevant roles in the light reaction in photosynthesis, as expected (Supplementary 4).

We also applied RapaNet to analyze the genes co-expressed with *CYP79F1*, which is involved in secondary metabolism. CYP79F1 catalyzes the oxidation of amino acids to aldoximes, the initial step in the formation of the core structure in the synthesis of short-chain methionine-derived GS.[33] The 1141 genes co-expressed with *CYP79F1* of *Arabidopsis* have 215 counterparts in the list of genes co-expressed with *CYP79F1* of *B rapa* (Supplementary 11). This number appears to be markedly lower than that found for *psbA*, a gene involved in primary metabolism, but is comparable with that found for ribosomal protein *L7Ae*. In addition, a GO analysis showed that these 215 genes were highly enriched for GS biosynthetic process terms, such as GO:0010439, GO:0019760, and GO:0019761. Moreover, the r^2 values obtained for the *B rapa* GS pathway genes (59) with Brapa_ESTC006895 (*CYP79F1*) and the corresponding

Arabidopsis CYP79F1 homologues show a significant linear relationship (Supplementary 10). Tight coupling in the initial steps of the synthesis of methionine-derived GS[34] might be important for achieving efficient control from the perspective of global regulation. Thus, the values found for the correlation of *CYP79F1* with *BACT4* and *BAT5* were quite high, in the range of 0.7 to 0.8, in both *B rapa* and *Arabidopsis*. The co-regulation of *BCAT4* in the cytosol or of *BACT3* in the chloroplast might be essential because they immediately follow amino acid chain elongation. In addition, chain-elongated amino acids in the chloroplast must be exported to the cytosol for further modification, and *BAT5* functions as a translocator of 2-keto acids between the chloroplasts and cytosol. The data also suggest that this pathway is highly comparable and conserved in *B rapa* and *Arabidopsis*.

The evolutionary split between the Brassicaceae family and *A thaliana* occurred ~20 Mya.[1,2] Recent Brassicaceae spp genome sequencing has revealed that these genomes underwent segment inversions and translocations as well as fusions and fissions, resulting in Brassicaceae lineage–specific whole-genome triplication.[4] Variation in the number of members of the gene families in the genome might have contributed to the remarkable morphological plasticity and presumably functional specification of these genes. A comparison of the co-expression profiles of *B rapa* and *A thaliana* genes suggests that the genes involved in primary pathways, such as photosynthesis, have been relatively strictly conserved throughout evolution. In contrast, the co-expressed genes of cell constituents or secondary metabolites, such as *L7Ae* and *CYP79F1*, have largely diverged throughout evolution. However, most of the key enzymes that play central roles in pathway regulation are strictly co-regulated, as demonstrated with *CYP79F1*.

Conclusions

RapaNet was designed to delineate the co-expressed genes in *B rapa*. The Web pages and programs associated with RapaNet can be accessed using any browser and operating system. The analysis of the genes co-expressed with a ribosomal protein, a photosystem II protein, and a protein in the GS synthesis pathway retrieved using RapaNet in this study highlight the functional characteristics of these genes. RapaNet can help researchers identify gene relationships and analyze gene functions. In addition, RapaNet was designed in a species-independent manner and can thus be expanded to construct similar Web-based tools for other organisms.

Acknowledgements

The authors thank Dr Sang Choon Lee and Beom-Seok Park for the advice provided during the initial stage of this work.

Author Contributions

JK and KMJ designed the software architecture and wrote most of the manuscript. JK and T-HL built the database and implemented the software. JSK, SC, and Y-MP participated in the microarray analysis. S-IS, SIL, M-HL, C-KK, and YH provided the data. BHN assisted with project development. Y-KK inspired the overall work and revised the final manuscript. All the authors read and approved the final manuscript.

Supplementary Material

Supplementary 1. Sample information of microarray data in RapaNet.
Supplementary 2. The structure of RapaNet.
Supplementary 3. List of genes that are highly and commonly (336) co-expressed with L7Ae in B. rapa and Arabidopsis.
Supplementary 4. List of genes that are highly and commonly (1103) co-expressed with psbA in B. rapa and Arabidopsis.
Supplementary 5. Hierarchical tree graph of the significant GO terms of the genes positively related to psbA.
Supplementary 6. Metabolic pathway for hydroxyalkenyl glucosinolate biosynthesis from methionine.
Supplementary 7. Gene network of the metabolic pathway of hydroxyalkenyl glucosinolate synthesis from methionine.
Supplementary 8. The list of genes retrieved with the 11 genes involved in the synthesis of the short aliphatic glucosinolate from methionine.
Supplementary 9. r2-values of the genes involved in glucosinolate pathways with CYP79F1.
Supplementary 10. A linear model of the r2-values of the genes involved with CYP79F1 in glucosinolate pathways.
Supplementary 11. List of genes that are highly and commonly (215) co-expressed with CYP79F1 in B. rapa and Arabidopsis.
Supplementary 12. Primer pairs used for semi-quantitative RT-PCR analysis.

REFERENCES

1. Wang X, Wang H, Wang J, et al; The *Brassica rapa* Genome Sequencing Project Consortium. The genome of the mesopolyploid crop species *Brassica rapa*. *Nat Genet*. 2011;43:1035–1039.
2. Lysak MA, Koch MA, Pecinka A, Schubert I. Chromosome triplication found across the tribe *Brassiceae*. *Genome Res*. 2005;15:516–525.
3. Nagaharu U. Genome analysis in *Brassica* with special reference to the experimental formation of *B. napus* and peculiar mode of fertilization. *Japan J Botany*. 1935;7:389–452.
4. Liu S, Liu Y, Yang X, et al. The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes. *Nat Commun*. 2014;5:3930.
5. Barrett T, Edgar R. Gene expression omnibus (GEO): microarray data storage, submission, retrieval, and analysis. *Methods Enzymol*. 2006;411:352–369.
6. Kolesnikov N, Hastings E, Keays M, et al. ArrayExpress update—simplifying data submissions. *Nucleic Acids Res*. 2015;43:D1113–D1116.
7. Dong X, Kim WK, Lim YP, Kim YK, Hur Y. Ogura-CMS in Chinese cabbage (*Brassica rapa* ssp. *pekinensis*) causes delayed expression of many nuclear genes. *Plant Sci*. 2013;199-200:7–17.
8. Lee SC, Lim MH, Kim JA, et al. Transcriptome analysis in *Brassica rapa* under the abiotic stresses using *Brassica* 24K oligo microarray. *Mol Cells*. 2008;26:595–605.
9. Trick M, Cheung F, Drou N, et al. A newly-developed community microarray resource for transcriptome profiling in *Brassica* species enables the confirmation of *Brassica*-specific expressed sequences. *BMC Plant Biol*. 2009;9:50.
10. Ballouz S, Verleyen W, Gillis J. Guidance for RNA-seq co-expression network construction and analysis: safety in numbers. *Bioinformatics*. 2015;31:2123–2130.
11. Giorgi FM, Del Fabbro C, Licausi F. Comparative study of RNA-seq- and microarray-derived coexpression networks in *Arabidopsis thaliana*. *Bioinformatics*. 2013;29:717–724.
12. Cao P, Jung KH, Choi D, Hwang D, Zhu J, Ronald PC. The rice oligonucleotide array database: an atlas of rice gene expression. *Rice*. 2012;5:17.
13. De Bodt S, Carvajal D, Hollunder J, van den Cruyce J, Movahedi S, Inze D. CORNET: a user-friendly tool for data mining and integration. *Plant Physiol*. 2010;152:1167–1179.

14. Hamada K, Hongo K, Suwabe K, et al. OryzaExpress: an integrated database of gene expression networks and omics annotations in rice. *Plant Cell Physiol*. 2011;52:220–229.
15. Hokamp K, Roche FM, Acab M, et al. ArrayPipe: a flexible processing pipeline for microarray data. *Nucleic Acids Res*. 2004;32:W457–W459.
16. Katari MS, Nowicki SD, Aceituno FF, et al. VirtualPlant: a software platform to support systems biology research. *Plant Physiol*. 2010;152:500–515.
17. Lee TH, Kim YK, Pham TTM, et al. RiceArrayNet: a database for correlating gene expression from transcriptome profiling, and its application to the analysis of coexpressed genes in rice. *Plant Physiol*. 2009;151:16–33.
18. Lorenz WW, Alba R, Yu YS, Bordeaux JM, Simões M, Dean JF. Microarray analysis and scale-free gene networks identify candidate regulators in drought-stressed roots of loblolly pine (*P. taeda* L.). *BMC Genomics*. 2011;12:264.
19. Mutwil M, Klie S, Tohge T, et al. PlaNet: combined sequence and expression comparisons across plant networks derived from seven species. *Plant Cell*. 2011;23:895–910.
20. Obayashi T, Okamura Y, Ito S, et al. ATTED-II in 2014: evaluation of gene co-expression in agriculturally important plants. *Plant Cell Physiol*. 2014;55:e6.
21. Basnet RK, Moreno-Pachon N, Lin K, et al. Genome-wide analysis of coordinated transcript abundance during seed development in different *Brassica rapa* morphotypes. *BMC Genomics*. 2013;14:840.
22. Aoki K, Ogata Y, Shibata D. Approaches for extracting practical information from gene co-expression networks in plant biology. *Plant Cell Physiol*. 2007;48:381–390.
23. Srinivasasainagendra V, Page GP, Mehta T, Coulibaly I, Loraine AE. CressExpress: a tool for large-scale mining of expression data from *Arabidopsis*. *Plant Physiol*. 2008;147:1004–1016.
24. Irizarry RA, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*. 2003;4:249–264.
25. Gentleman RC, Carey VJ, Bates DM, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*. 2004;5:R80.
26. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*. 1987;4:406–425.
27. Di Salle P, Incerti G, Colantuono C, Chiusano ML. Gene co-expression analyses: an overview from microarray collections in *Arabidopsis thaliana*. *Brief Bioinform*. 2017;18:215–225.
28. Du Z, Zhou X, Ling Y, Zhang Z, Su Z. agriGO: a GO analysis toolkit for the agricultural community. *Nucleic Acids Res*. 2010;38:W64–W70.
29. Selmar D. Biosynthesis of cyanogenic glycosides, glucosinolates and non protein amino acids. In: Wink M, ed. *Biochemistry of Plant Secondary Metabolism*. Boca Raton, FL: CRC Press; 1999:79–150.
30. Wang H, Wu J, Sun S, et al. Glucosinolate biosynthetic genes in *Brassica rapa*. *Gene*. 2011;487:135–142.
31. Zang YX, Kim HU, Kim JA, et al. Genome-wide identification of glucosinolate synthesis genes in *Brassica rapa*. *FEBS J*. 2009;276:3559–3574.
32. Woodson JD, Chory J. Coordination of gene expression between organellar and nuclear genomes. *Nature Rev Genet*. 2008;9:383–395.
33. Reintanz B, Lehnen M, Reichelt M, et al. Bus, a bushy *Arabidopsis* CYP79F1 knockout mutant with abolished synthesis of short-chain aliphatic glucosinolates. *Plant Cell*. 2001;13:351–367.
34. Gigolashvili T, Yatusevich R, Rollwitz I, Humphry M, Gershenzon J, Flugge UI. The plastidic bile acid transporter 5 is required for the biosynthesis of methionine-derived glucosinolates in *Arabidopsis thaliana*. *Plant Cell*. 2009;21:1813–1829.