Similarity of gene expression profiles provides important clues for understanding the biological functions of genes, biological processes and metabolic pathways related to genes. A gene expression network (GEN) is an ideal choice to grasp such expression profile similarities among genes simultaneously. For GEN construction, the Pearson correlation coefficient (PCC) has been widely used as an index to evaluate the similarities of expression profiles for gene pairs. However, calculation of PCCs for all gene pairs requires large amounts of both time and computer resources. Based on correspondence analysis, we developed a new method for GEN construction, which takes minimal time even for large-scale expression data with general computational circumstances. Moreover, our method requires no prior parameters to remove sample redundancies in the data set. Using the new method, we constructed rice GENs from large-scale microarray data stored in a public database. We then collected and integrated various principal rice omics annotations in public and distinct databases. The integrated information contains annotations of genome, transcriptome and metabolic pathways. We thus developed the integrated database OryzaExpress for browsing GENs with an interactive and graphical viewer and principal omics annotations (http://riceball.lab.nig.ac.jp/oryzaexpress/). With integration of Arabidopsis GEN data from ATTED-II, OryzaExpress also allows us to compare GENs between rice and Arabidopsis. Thus, OryzaExpress is a comprehensive rice database that exploits powerful omics approaches from all perspectives in plant science and leads to systems biology.

Keywords: Correspondence analysis • Database • Gene expression network • Microarray • Oryza sativa • Systems biology.

Abbreviations: CA, correspondence analysis; DCA, distances obtained from CA; GEN, gene expression network; GO, gene ontology; MR, mutual rank; PAC, partial correlation coefficient; PCC, Pearson correlation coefficient; RMA, robust multiarray average.

Introduction

To maintain biological activity, an appropriate gene set essential for cells, tissues, organs and the individual level is selected from the genome and enhanced/suppressed through proper regulatory mechanisms of gene expression. An essential gene set controlled by a biological or physiological process frequently shows similar temporal and/or spatial expression profiles (Gibson et al. 2004, Al-Ghazi et al. 2009, Matsumoto et al. 2009, Swanson-Wagner et al. 2009). Therefore, similarity of gene expression profiles provides important clues for understanding the biological functions of genes, biological processes and metabolic pathways related to genes (Chen et al. 2008, Endo et al. 2009, Yamagishi et al. 2009, Hsu et al. 2010). Microarray technology is a powerful and effective tool for genome-wide gene expression analysis within species (Matsuura et al. 2010, Sakuraba et al. 2010) and between species (Tsaparas 2006, Miller et al. 2010), and is used to identify gene sets showing similar expression profiles among various biological conditions (samples). Recently, large-scale microarray
data have been accumulated in publicly available databases such as NCBI GEO (Barrett et al. 2009), EBI ArrayExpress (Parkinson et al. 2009) and DDBJ CIBEX (Ikeo et al. 2003). In addition, combined methods with laser microdissection and microarray have also been applied to isolating specific cells from complicated plant tissues and separate transcriptomes (Hobo et al. 2008, Suwabe et al. 2008, Watanabe 2008). With a wealth of microarray data and a high resolution method for the transcriptome, a comprehensive classification of gene expression profiles is possible, allowing us to elucidate gene functions and families (homologs).

A gene expression network (GEN) is an ideal technique for grasping similarities of expression profiles among genes simultaneously. For GEN construction, the Pearson correlation coefficient (PCC) has been widely used as an index to evaluate similarities of expression profiles for gene pairs (Aoki et al. 2007, Fuji et al. 2010, Matsuura et al. 2010, Soeno et al. 2010). Obayashi and Kinoshita (2009) suggested that PCCs are significantly overestimated when many replicates of a sample (sample redundancies) are contained in a microarray data set. Thus, instead of the PCC, mutual rank (MR) based on rank transformations of the weighted PCC (wPCC) has been used as a more sensitive index for sample redundancies. However, this algorithm needs a cut-off threshold value in estimating sample redundancy. The cut-off threshold value should be statistically tested to evaluate similarities of gene expression profiles appropriately. On the other hand, correspondence analysis (CA) (Greenacre 2007), which is a multivariate analysis method for profile data, permits concise interpretation of the correspondence between genes and samples in microarray analysis (Yano et al. 2006). CA for microarray data summarizes an originally high dimensional data matrix [rows (genes) and columns (samples)] into a low dimensional projection (space). Scores (coordinates) in the low dimensional space are given to each gene and sample. With the coordinates, genes and samples can be plotted into a two- or three-dimensional subspace. The distance between plots (genes) in a low dimensional space, which is calculated from all or statistically significant dimensions, depends on the degree of similarity of gene expression profiles: a short distance means similar gene expression profiles and a long distance means different expression profiles. Thus, distances can be used as an index for similarity of gene expression profiles. In addition, CA does not require any prior parameters to evaluate similarity, because it only calculates distances between plots. Moreover, the effect of sample redundancies in a data set can be mathematically eliminated by reducing dimensions with the CA algorithm. Moreover, CA takes minimal time even for a large-scale microarray data set (within 30 min for approximately 50,000 probes × 600 samples with a personal computer, such as a MacBook3,1 OS X 10.6.5 with Intel Core 2 Duo 2.2 GHz and 4 Gb memory). The new index presented here, distances obtained from CA (DCA), is suitable for appropriate and quick evaluation of the similarities of gene expression profiles and for construction of GENs.

Rice omics data including genome annotations are available from major public databases. Current genome annotations have been provided by the Michigan State University (MSU) Rice Genome Annotation Project (Ouyang et al. 2007) and the Rice Annotation Project (RAP-DB) (Rice Annotation Project 2008). Since their bioinformatics approaches to genome assembly and annotation are different from each other, the predicted sequences and identifiers (IDs) of loci and mRNA are consequently different between MSU and RAP-DB. For example, a locus Os06g0103700 (Os06t0103700-01) in RAP-DB is named LOC_Os06g01410 in MSU. Other omics databases in rice also employ IDs of either MSU or RAP-DB. Information on metabolic pathways in RiceCyc (Li et al. 2008) is described by locus IDs of MSU, while the KEGG PATHWAY database (Okuda et al. 2008) provides information on metabolic pathways with locus IDs of RAP-DB. For annotations of microarray probes, IDs of MSU and RAP-DB are used for Affymetrix and Agilent platforms, respectively. Due to the inconsistent IDs among databases, users cannot directly compare omics information from distinct databases, such as RiceCyc and KEGG PATHWAY, as is the case for rice GENs. Information on rice GENs is available from public databases such as ATTED-II (Obayashi et al. 2007), RiceArrayNet (Lee et al. 2009), GeneCAT (Mutwil et al. 2008) and Rice Array Database (Jung et al. 2008). ATTED-II and GeneCAT provide information on GENs constructed from expression data of Affymetrix GeneChip Rice Genome Array. GENs in RiceArrayNet are constructed from the Rice60k Microarray. Rice Array Database provides only a list of genes pairs with PCCs, and does not provide a GEN viewer to grasp similarities among multiple genes simultaneously. Differences in microarray platforms and GEN information formats prevent analyses and comparisons of information on GENs in distinct databases. Moreover, speculation regarding biological features hidden in GENs requires various annotations from omics databases. To overcome such issues, we constructed rice GENs by CA and integrated principal omics information. The information is available from our database OryzaExpress (http://riceball.lab.nig.ac.jp/oryzaexpress/). OryzaExpress enables us to trace gene IDs from different databases/projects, browse GENs and refer to principal omics data stored in public databases. GENs and annotation data integrated in OryzaExpress thus provide more detailed and comprehensive information.

**Gene expression data for GEN analysis**

A total of 624 sample data sets in 37 experimental series (CEL files from Affymetrix GeneChip Rice Genome Array, GPL2025) were collected from NCBI GEO (Barrett et al. 2009). The 37 experimental series included gene expression data along with a variety of biological and experimental conditions, such as time courses, stress treatments, growth stages, organs, transformed plants and mutant lines (Supplementary Table S1; see also the download page in OryzaExpress). The collected data were normalized in logarithmic scale by the robust multiarray

---

**Plant Cell Physiol. 52(2): 220–229 (2011) doi:10.1093/pcp/pcq195 © The Author 2010. 221**
average (RMA) method using the programs in R/Bioconductor (Gentleman et al. 2004). Since the expression data normalized with the RMA method have no negative values, the data set can be used to perform CA calculations. To detect outliers, the average and variance of gene expression levels among the 624 samples were also calculated for each probe.

CA for GEN

CA (Yano et al. 2006) was conducted against the normalized gene expression data by the statistical package R (library `ca`) (Nenadíc and Greenacre 2007) and our developed software. The calculation was performed in a Linux machine [RedHat 5, 64 bit operating system with Intel(R) Core(TM) 2 Duo 2.33 GHz and 4 Gb memory] (Supplementary Fig. S2). For each probe (gene), the coordinates in the low dimensional space were obtained. To evaluate similarities of gene expression profiles for each probe pair, a DCA (Euclidean distance) between two probes in the low dimensional space was calculated. As the DCA value is close to zero, the two probes have similar expression profiles (Supplementary Fig. S1). We also used PCCs (PCC_CAs) for the coordinates of each gene pair to attempt to identify genes showing reciprocal (inverse) expression profiles. Reciprocal expression profiles are sometimes effective in searching repressor or downstream genes (e.g. Zeng et al. 2010). When the expression profiles for a gene pair show a largely reciprocal profile in the majority of samples, the PCC_CAs value becomes negative (Supplementary Fig. S1). In the calculation of DCAs and PCC_CAs, we used the first 15 dimensions whose explained percentages are ≥1%. The cumulative explained percentage of 15 dimensions is 71.9%.

Calculations of other indices for gene expression similarities

As additional indices for gene expression similarities, PCCs, MRs and partial correlation coefficients (PACs) were calculated. The PCC between probes \( x \) and \( y \) was obtained by the following equation:

\[
PCC(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}},
\]

where \( n \) is the total number of samples, \( x_i \) and \( y_i \) are expression levels of \( x \) and \( y \) in the \( i \)th sample, respectively, and \( \bar{x} \) and \( \bar{y} \) are means of expression levels among samples, respectively. MR and PAC for each probe pair were calculated when the PCC value was ≥0.4. Although MRs should be obtained from the rank transformations of wPCCs (Obayashi et al. 2009), calculations of MRs were based on the rank transformations of PCC for simplicity in our analysis. PAC between probes \( x \) and \( y \) given probe \( z \) [the first-order PAC \( (x, y | z) \)] provides the strength of a direct association between \( x \) and \( y \) by eliminating the effect of \( z \) (Snedecor and Cochran 1989). For example, it is assumed that genes \( x \) and \( y \) are controlled and up-regulated by the expression of gene \( z \). The mechanisms involved in the regulation of expression could be suggested between \( x \) and \( z \) by the indices DCA, PCC and MR.

Like the relationship between \( x \) and \( z \), the similarity between \( y \) and \( z \) could be also implied by indices. However, if expression profiles between \( x \) and \( y \) are significantly similar according to indices, the similarity is indirectly caused by the expression profile of \( z \). To remove such a false positive between \( x \) and \( y \), the effect of the expression profile of gene \( z \) should be correctly eliminated in order to evaluate the similarity between \( x \) and \( y \). The association could be given by the equation

\[
PAC(x, y | z) = \frac{PCC(x, y) - PCC(x, z) PCC(y, z)}{\sqrt{1 - PCC(x, z)^2 \sqrt{1 - PCC(y, z)^2}}}
\]

In this study, we calculated first-order PACs \((x, y | z_i)\), where \( i = 1 \ldots n \) and \( n \) is the total number of probes except for \( x \) and \( y \). Among first-order PACs \((x, y | z_i)\), the minimum amount of PAC \((x, y | z_i)\) (PACmin) was used as a similar index which suggests the lowest association between genes \( x \) and \( y \). When the PACmin value is ≥0.13152 (the significant probability of 0.1%), the association between genes \( x \) and \( y \) is considered significant. On the other hand, when the PACmin value is less than the threshold value, the association between \( x \) and \( y \) is considered a false positive. Expression profiles of genes detected by PCC and MR and the numbers of false positives predicted by PACmin are shown in Supplementary Fig. S1. The calculations for PAC, MR and PACmin were performed on a Linux server (CentOS5.5 with Xeon 7560 2.26G 32core and 1Tb memory) to obtain the results in a relatively short time (Supplementary Fig. S2). The calculations were conducted separately with the 30 cores in parallel.

Construction of web interfaces for GENs

For visual inspection of similarities of expression profiles among multiple genes, web interfaces for GENs were developed using the graph (network) visualization tool ‘Graphviz’ (Gansner and North 2000). In the network graph as shown in Fig. 1, nodes indicate genes and edges across nodes show the strength of the associations (similarities of gene expression profiles). DCAs, PCCs, MRs and PACmin were used as the indices for the similarities of gene expression profiles. PAC_CAs, MRs and PCCs were used as the indices for reciprocal gene expression profiles. The statistics of gene pairs detected by DCA, PAC_CAs and PCC are provided in OryzaExpress.

Integration of the Arabidopsis GEN

Fundamental biological systems in gene expression are conserved over all species (Mochida and Shinozaki 2010, Shikata et al. 2010). Comparison of GENs among different species facilitates identification of conserved and species-specific gene expression mechanisms. To assist this, data of the Arabidopsis GEN were collected from the ATTED-II and integrated into OryzaExpress. Arabidopsis genes were mapped in rice GENs according to information on orthologs between rice and Arabidopsis. From the InParanoid7 (Ostlund et al. 2010), we collected 15,743 orthologous genes (10,637 groups) between Arabidopsis and rice.
Among them, 12,481 predicted orthologous genes (gene pairs) of rice and Arabidopsis have corresponding microarray probes on the Affymetrix GeneChip Rice Genome Array and Affymetrix GeneChip Arabidopsis ATH1 Genome Array, respectively. Based on the information on the orthologs and microarray probes, the Arabidopsis orthologs were mapped in the rice GENs in OryzaExpress. Using the developed interface of GEN viewers in OryzaExpress, the additional information on similarities of expression profiles between Arabidopsis orthologs could be shown.

We compared a GEN between rice and Arabidopsis to assess the potential usefulness of the integrated GENs from different species. We used a GEN which contained transcription factors for flower development reported to be conserved among Arabidopsis, tobacco and lily (Chang et al. 2009, Hsu et al. 2010). As well as Arabidopsis genes in flower development, these rice orthologs comprise a GEN. In the Arabidopsis GEN, positive correlations (PCC = 0.41–0.84) among the genes AP1 (At1g69120), AP3 (At3g54340), LFY (At5g61850), AG (At4g18960), PI (At5g20240) and SEP (At5g15800) were shown (Supplementary Table S2). Among their rice orthologs, positive correlations were also observed (PCC = 0.52–0.86). In addition, whereas the PCC between Arabidopsis genes FT (At1g65480) and AP1 (At1g69120) was too low to detect the relationship significantly (PCC = 0.10), the rice orthologs showed a high PCC value (PCC = 0.87).

### Integration of omics information

To browse omics information with GEN, principal omics data including rice genome annotations were collected and stored in OryzaExpress (Table 1). The relationships of locus IDs between RAP-DB and MSU were downloaded from RAP-DB. MSU loci lacking counterparts in RAP-DB, which were not included in the above downloaded data, were appended to the relationships by our perl scripts. Using the relationship of locus ID between RAP-DB and MSU, omics data in other databases were integrated. The integrated information is as follows: protein data (UniProt), metabolic pathway data (KEGG and RiceCyc), gene expression data (GEO and ‘Rice MPSS’) and annotations of microarray probes in Agilent and Affymetrix platforms.

### Database functions

OryzaExpress provides integrated information on GENs and annotations in rice. From the page for GEN, information from integrated annotation data is accessible, and vice versa. Detailed information in the public (external) databases is also accessible from hyperlinks in OryzaExpress. The database functions and usage are also described in the help menu in OryzaExpress.
Search function on GEN

Information on GENs for each probe (gene) was prepared and stored in OryzaExpress. The GEN viewer is accessible from the page for information on each probe. The information on the probe (query probe) can be searched by probe IDs or annotation keywords of Affymetrix microarray probes (Fig. 2A). A GEN for a query probe is shown by the hyperlinks in the retrieved page. The GEN page for a query probe contains annotations of the query probe, a GEN viewer, a list and brief annotations of gene (probe) pairs showing similar/reciprocal expression profiles, gene expression profiles collected from GEO, Arabidopsis orthologs and metabolic pathway names (Fig. 2B).

With an interactive and graphical viewer, a GEN is shown with nodes (probes) and edges (associations) (Fig. 1). This viewer allows rotation, zooming in and out, and panning of the GEN image. A query probe is depicted in the center of the GEN image by a yellow node. Red and blue edges imply similarities of expression profiles and reciprocal expression profiles between two probes, respectively. Graphs for expression profiles (levels) of the query and associated probes are also shown in the GEN page (Fig. 2B). These expression profiles help us to assess the reliabilities of the GEN. For the query probe, average and variance of expression levels among samples are also displayed by histograms in the GEN page. The histograms for average and show frequency distributions for all probes. The bars containing the query probe are highlighted in orange. Using the histograms, an outlier probe, such as an extraordinarily high average or variance, can be easily distinguished.

Annotations for each probe are shown in the GEN viewer and page. A brief annotation of each probe pops up by scrolling the mouse cursor over a node. Detailed annotations are accessible with internal and external links in the GEN page (Fig. 2B, D). Further information on metabolic pathways (KEGG and RiceCyc), Arabidopsis genes (TAIR), Arabidopsis GENs (ATTED-II) and microarray experiment data (GEO) are also available through the external links.

Other GENs can be depicted by the setting options. Users can select the maximum number of edges for a node. In the current version, the maximum number of edges for a node can be selected from one to six. Although the DCA index is used as a default setting, other indices for GEN construction can also be selected. In addition, GO terms (Gene Ontology Consortium 2000), metabolic pathway names and Arabidopsis ortholog names for each probe can also be shown in a node as optional settings. The similarities of expression profiles between Arabidopsis orthologs (ATTED-II) are shown as black edges in the GEN.

OryzaExpress has an optional function to eliminate potential false positives. In the current version of OryzaExpress, false positives are detected by the PACmin index with a significant probability of 0.1%. According to this threshold, 89,653,195 probe pairs, which make up 79.8% of the total probe pairs, are regarded as false positives. To assess reliability, we checked a GEN containing the TDR (tapetum degeneration retardation) gene in pollen development. TDR is expressed in tapetum at an early stage of pollen development and controls lipid metabolism for pollen outer wall formation. In a tdr-defective mutant, expression of >200 genes is affected (Zhang et al. 2008). Among these genes, two enzyme genes, LOC_Os01g65590 (galactosyl transferase family protein) and LOC_Os05g9830 (lipase family protein), show a significant association from the PCC (data not shown). However, the two genes have no actual association since they are related to

### Table 1 Omics data collected from public databases

| Information       | Database/microarray platform | URL                                      | Notes                                  |
|-------------------|------------------------------|------------------------------------------|----------------------------------------|
| Genome annotation | RAP-DB                       | http://rapdb.dna.affrc.go.jp/            | IRGSP/IRAP build 5                     |
|                   | MSU Rice Genome Annotation   | http://rice.plantbiology.msu.edu/        | Release 6.1                            |
| Functional annotations | The Gene Ontology       | http://www.geneontology.org/            |                                       |
|                   | UniProtKB/Swiss-Prot         | http://www.ebi.ac.uk/uniprot/index.html  |                                       |
| Metabolic pathways | KEGG PATHWAY                | http://www.genome.jp/kegg/pathway.html  | Version 3.0                            |
|                   | RiceCyc                      | http://www.gramene.org/pathway/          | Download from Agilent eArray           |
| Microarray platforms | Agilent Rice Oligo Microarray (22 k) | https://earray.chem.agilent.com/earray/ | Rice Annotations, CSV format, Release 30 |
|                   | Agilent Rice Oligo Microarray (44 k) | https://earray.chem.agilent.com/earray/ |                                       |
|                   | Affymetrix Rice Genome Array | http://www.affymetrix.com/estore/browse/products.js?navMode=34000&productId=131497&navAction=jump&aid=productsNav |                                       |
| Gene expression data | NCBI GEO                      | http://www.ncbi.nlm.nih.gov/geo/         |                                       |
|                   | Rice MPSS                    | http://mpss.udel.edu/rice/               |                                       |
different metabolic pathways. The false positive relationship could be removed by PACmin. Although the PACmin used here is a minimum value of all first-order PACs for a gene pair, the highest order of PAC should be used to detect false positives. We applied PACmin as a simple measurement to reduce calculation time for the highest order of PACs of all probe pairs.

Search function for omics data

In OryzaExpress, search functions of integrated annotation data are available by IDs or annotation keywords (Fig. 2C). Gene IDs of RAP and MSU, accession numbers of GenBank and probe IDs of three microarray platforms can be used to search annotations from IDs. The detailed information page for each gene contains information on the locus ID and annotations (RAP-DB and MSU), accession numbers (GenBank), protein IDs and annotations (UniProt), metabolism pathway names and Enzyme Commission (EC) numbers (KEGG and RiceCyc), probe IDs on three microarray platforms and the link to gene expression data (Rice MPSS database) (Fig. 2D). The information on integrated annotations is shown together with internal and external links. The integrated annotation page also contains the hyperlink to a GEN information page (Fig. 2B).

Conclusion and Discussion

What are the applications of OryzaExpress to different kinds of research?

To overcome the current decentralized omics data and databases in rice, we constructed the web-based and integrated database OryzaExpress. OryzaExpress provides an overview of rice GENs and various kinds of omics information including genome annotations, metabolic pathways and gene expression. It is the first database to provide information on both GENs and integrated omics annotations among rice databases. This information should help us to grasp gene features, characteristics of gene expression profiles and their expression control modules.
from a variety of perspectives. More detailed information from public databases is also accessible through hyperlinks on the pages of OryzaExpress. Researchers can obtain principal omics data specific to their needs, quickly, using OryzaExpress as a starting point. This database allows researchers to maximize productivity by providing the vast amounts of omics data currently available and should be a powerful database especially for research on rice by omics approaches in plant science.

OryzaExpress provides the GEN using Affymetrix microarray data of >600 samples. Notably, it provides a visual network diagram that can judge the relevance of gene expression profiles. In addition, OryzaExpress provides GENs with many kinds of omics information such as GO, metabolic pathways and comparison of GENs between rice and Arabidopsis. Most omics information is also available by hyperlinks in OryzaExpress. However, to exploit the information fully, it still requires additional information on protein–protein interactions. Although GEN is key for retrieving genes and their expression control modules, it is based only on mRNA levels. This information is insufficient when the target gene is not influenced at the protein level. Information on protein–protein interactions is also important for understanding biological events at the protein level. Therefore, we are focusing our future work on launching such protein–protein interaction information in OryzaExpress.

CA method for GEN construction

This is the first report of GEN construction by CA. CA is a reasonable method for evaluating similarities of expression profiles, since it has been developed to analyze profile data (Yano et al. 2006, Greenacre 2007). To detect a gene set related to the same biological process, the up-/down-regulation patterns across samples, namely expression profiles, are important. DCAs theoretically reflect the similarity of gene expression profiles. Genes with the same expression profiles are located at the same position in the low dimensional space (Yano et al. 2006). Genes with similar expression profiles are closely located to each other. DCA is an appropriate index to evaluate the similarities of gene expression profiles directly without any prior parameters. In addition, CA provides the relationships between genes and samples (bi-plot), and the information provided by a bi-plot can facilitate the detection of novel genes related to the various biological and environmental conditions. Thus the bi-plot data will be integrated into OryzaExpress.

In addition, CA is preferable for large-scale omics data analysis, as CA calculations for large-scale data sets is both time-effective and has a minimal requirement for computational resources (Supplementary Fig. S2). This method is indispensable in current research that involves large-scale omics data. Data-mining and analyses are generally repeated using the same large-scale data set to acquire new biological findings. A time-consuming method with a high-performing computational system is thus not practical. Although calculation of DCAs between plots (genes) requires some time, calculation of DCA for a probe pair could be omitted when the coordinates of two probes in one dimension (axis) are remarkably different. That is, such a remarkable difference in the coordinates immediately indicates that the distance between the two genes is great in space. A vast majority of gene pairs have remarkably different rather than very close coordinates, due to the limited number of genes with similar expression profiles, compared with the total number of genes in the genome. On the other hand, using PCCs for similarity evaluations between two genes, no step could be skipped, as the amount of a PCC is not known until the calculation is complete.

The PCC has been widely used as an index for the similarity of gene expression profiles. A high PCC indicates that two genes have very similar expression profiles. In some cases, even genes with biologically meaningful relationships have a low PCC (Obayashi et al. 2009). These results may be caused by the characteristics of the PCC as it is not a statistical method for profile data analysis. PCC calculation is based on the sum of products of deviations from the means. The amount of deviation from the mean, therefore, has a considerable effect on the value of the PCC. However, it is not clear whether the deviations of two gene expression levels can always reflect the biological expression profiles. On the other hand, CA calculation uses expression data directly as the matrix [rows (genes) and columns (samples)]. The gene list detected by CA shows exactly the considerable similarity of up-/down-regulation profiles across samples (Supplementary Fig. S1). Although probe pairs may show high PCCs of around 0.8–0.9, the similarities of gene expression profiles are not always statistically significant from DCA. In some cases, gene pairs are given high PCCs unless the up-/down-regulation patterns across samples (expression profiles) are actually the same (Supplementary Fig. S3). The differences between PCCs and DCAs should be examined to evaluate the applicability of the new index DCA in detail. However, the current omics data from public databases and experiments are rapidly expanding, and PCC calculations for large-scale omics data are notoriously difficult, due to the long calculation time and large computer memory requirements. We thus have to develop novel indices such as DCA for future large-scale omics analyses.

In our assessment, we found an example where DCA was effective in detecting biologically meaningful gene pairs. With a DCA of 0.08 as a low threshold value, which enabled us to avoid many false positives (data not shown), we detected many gene pairs in the same reaction of RiceCyc metabolic pathways. Surprisingly, some gene pairs on >10 metabolic pathways could not be detected by even the low PCC threshold value of 0.4 that statistically means a weak positive correlation relationship. For example, on the pathway ‘aerobic respiration–electron donors reaction list’, 28 gene pairs were detected only by DCA, and 15 gene pairs only by PCC (Supplementary Table S3), whereas 21 gene pairs were detected by both methods. This result suggests that DCA and PCC would compensate each other to discover biologically related gene pairs.
To grasp the similarities of expression profiles among many genes simultaneously with DCAs (distances), a 3D image viewer of plots (genes) in the subspace rather than a GEN viewer helps us to detect genes (plots) located close to each other. However, the current version of OryzaExpress has no facility to view a 3D image. We have developed GUI software using Java3D to perform CA easily and view plots (genes) in the low dimensional subspace. With the data sets of DCAs obtained from OryzaExpress, a 3D image can be visualized in the software which allows rotation, zooming in and out, and panning of the image. Annotation data obtained from OryzaExpress could also be imported into the software. Users can search plots (genes) in the 3D image by annotation keywords. In the low dimensional space, plots (genes) of unknown function around a gene of known function would be candidates for further analysis. The software (beta version) is freely available for academic research by e-mail request.

**Mining of genes with reciprocal expression profiles**

OryzaExpress can be adopted to search genes with reciprocal expression profiles, although most GEN databases have offered only positive correlations. A negative PCC has sometimes been found to be effective in evaluation (Zeng et al. 2010). In a potato plant overexpressing the sucrose synthase gene, a negative PCC in gene expression was observed between sucrose synthase and acid invertase (Whittaker et al. 2010). However, whether a negative PCC is effective for searching their repressor or downstream genes is still unclear. Yano et al. (2006) reported CA detects genes and samples with reciprocal expression profiles. Genes with reciprocal expression profiles separate into positive and negative coordinates of the axes in the low dimensional space. This suggests that PCCs for coordinates obtained by CA (PCC_CAs) could classify genes with reciprocal expression profiles. We tested the new PCC_CA index to mine genes with reciprocal expression profiles. As expected, the results show a negative PCC_CA, allowing detection of genes with reciprocal expression profiles (Supplementary Fig. S1). Although a long calculation time is needed for PCC_CAs, like PCCs, the combined analysis of CA and PCC_CA could yield a highly desirable tool.

**Detection of false positives in GENs**

We applied first-order PACs to detect false positives among predicted genes with similar expression profiles. This index provides strength of association of expression profiles between a gene pair by removing the effect of other genes. PAC has been proposed as an index for expression similarities (Usadel et al. 2009), and a false positive can be greatly reduced using first-order PACs to construct GEN (Roessner-Tunali et al. 2003, de la Fuente et al. 2004, Han et al. 2008, Sawada et al. 2009a, Sawada et al. 2009b, Shinozaki and Sakakibara 2009). In fact, we found approximately 80% false positives among total gene pairs in the rice GEN.

The highest order of PAC is theoretically desirable to assess false positives in GENs. In the evaluation of similarities between genes x and y with the highest order of PAC, the effects of gene expression profiles of all proves except for genes x and y could be simultaneously removed. Alternatively, the calculation of the highest order of PACs for all proves would be difficult even with large computer resources. We thus need to develop other efficient indices to mine false positives in GENs.

**Integration of omics annotations**

We collected and integrated principal omics annotations of rice into OryzaExpress, allowing quick access to various annotations from distinct databases. In particular, genome annotations and IDs from different genome annotation projects can be easily accessed in OryzaExpress. The annotations in GENs contain various types of information such as metabolic pathways, GO terms and comparisons of GENs between rice and Arabidopsis. The integrated annotation data are updated into OryzaExpress depending on the main public databases (once or twice a year). It could lead us to promote comprehensive systems biology approaches. Recent progress on genome sequencing via next-generation sequencers should lead to even more genomic annotations in many model plants. However, the problem of different IDs and annotations from individual projects still remains an issue. Construction and maintenance of an integrated database such as OryzaExpress are also points for further discussion for maximizing the knowledge gained from experimental data and individual public databases.

In conclusion, we developed the comprehensive rice database OryzaExpress. This database provides both GENs and various types of omics information from public and distinct databases. It also allows us to apply powerful omics approaches from all perspectives to plant science and leads to systems biology.

**Supplementary data**

Supplementary data are available at PCP online.

**Funding**

This work was supported by the Ministry of Education, Culture, Sports, Science and Technology of Japan (MEXT) [Grants-in-Aid for Special Research on Priority Areas (Nos. 18075003 and 18075012) to M.W., 18075006 to M.M., 18075009, 18075011 and 18075012 to N.K., and 19043015 and 21024010 to K.Y.]; the Japan Society for Promotion of Science (JSPS) [Grants-in-Aid for Exploratory Research (19651084) to K.Y., Scientific Research (B) (20380022) to K.Y. and Young Scientists (Start-up) (21880022) to K.S.]; the Japan Science and Technology Agency (JST) [Grants-in-Aid to K.Y.].
Acknowledgments

We thank Drs. Hirohisa Kishino and Hiroyoshi Iwata (Tokyo University) for their valuable comments on statistical methods.

References

Al-Ghazi, Y., Bourrot, S., Arioli, T., Dennis, E.S. and Llewellyn, D.J. (2009) Transcript profiling during fiber development identifies pathways in secondary metabolism and cell wall structure that may contribute to cotton fiber quality. Plant Cell Physiol. 50: 1364–1381.

Aoki, K., Ogata, Y. and Shibata, D. (2007) Approaches for extracting practical information from gene co-expression networks in plant biology. Plant Cell Physiol. 48: 381–390.

Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Rudnev, D., Evangelista, C. et al. (2009) NCBI GEO: archive for high-throughput functional genomic data. Nucleic Acids Res. 37: D885–D890.

Chang, Y.Y., Chiu, Y.F., Wu, J.W. and Yang, C.H. (2009) Functional analysis of three (Oncidium Gower Ramsey) API/AGL9-like MADS box genes show novel expression patterns and cause different effects on floral transition and formation in Arabidopsis thaliana. Plant Cell Physiol. 50: 1425–1438.

Chen, M.K., Lin, I.C. and Yang, C.H. (2008) Various spatiotemporal expression profiles of anther-expressed genes in rice. Plant Cell Physiol. 49: 704–717.

de la Fuente, A., Bing, N., Hoeschele, I. and Mendes, P. (2004) Discovery of meaningful associations in genomic data using partial correlation coefficients. Bioinformatics 20: 3565–3574.

Endo, M., Tsuchiya, T., Hamada, K., Kawamura, S., Yano, K., Ohshima, M. et al. (2009) High temperatures cause male sterility in rice plants with transcriptional alterations during pollen development. Plant Cell Physiol. 50: 1911–1922.

Fujii, S., Yamada, M., Fujita, M., Itabashi, E., Hamada, K., Yano, K. et al. (2010) Cytoplasmic–nuclear genomic barriers in rice pollen development revealed by comparison of global gene expression profiles among five independent cytoplasmic male sterile lines. Plant Cell Physiol. 51: 610–620.

Gansner, E.R. and North, S.C. (2000) An open graph visualization system and its applications to software engineering. Software Pract. Exper. 30: 1203–1233.

Gene Ontology Consortium. (2000) Gene ontology: tool for the unification of biology. Nat. Genet. 25: 25–29.

Gentelman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S. et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 5: R80.

Gibson, G., Riley-Berger, R., Harshman, L., Kopp, A., Vacha, S., Nuzhdin, S. et al. (2004) Extensive sex-specific nonadditivity of gene expression in Drosophila melanogaster. Genetics 167: 1791–1799.

Greenacre, M.J. (2007) In Correspondence Analysis in Practice, 2nd edn. Chapman and Hall, London.

Han, L. and Zhu, J. (2008) Using matrix of thresholding partial correlation coefficients to infer regulatory network. Biosystems 91: 158–165.

Hobo, T., Suwabe, K., Aya, K., Suzuki, G., Yano, K., Ishimizu, T. et al. (2008) Various spatiotemporal expression profiles of anther-expressed genes in rice. Plant Cell Physiol. 49: 1417–1428.

Hsu, H.F., Hsieh, W.P., Chen, M.K., Chang, Y.Y. and Yang, C.H. (2010) C/D class MADS box genes from two monocots, orchid (Oncidium Gower Ramsey) and lily (Lilium longiflorum), exhibit different effects on floral transition and formation in Arabidopsis thaliana. Plant Cell Physiol. 51: 1029–1045.

Ikeo, K., Ishi-i, J., Tamura, T., Gojobori, T. and Tateno, Y. (2003) CIBEX: center for information biology gene expression database. C. R. Biol. 326: 1079–1082.

Jung, K.H., Dardick, C., Bartley, L.E., Cao, P., Phetsom, J., Canlas, P. et al. (2008) Refinement of light-responsive transcript lists using rice oligonucleotide arrays: evaluation of gene-redundancy. PLoS ONE 3: e3337.

Lee, T.H., Kim, Y.K., Pham, T.T., Song, S.J., Kim, J.K., Kang, K.Y. et al. (2009) RiceArrayNet: a database for correlating gene expression from transcriptome profiling, and its application to the analysis of coexpressed genes in rice. Plant Physiol. 151: 16–33.

Liang, C., Jaiswal, P., Hebbard, C., Avraham, S., Buckler, E.S., Casstevens, T. et al. (2008) Gramene: a growing plant comparative genomics resource. Nucleic Acids Res. 36: D947–D953.

Matsuurra, H., Ishibashi, Y., Shinmyo, A., Kanaya, S. and Kato, K. (2010) Genome-wide analyses of early translational responses to elevated temperature and high salinity in Arabidopsis thaliana. Plant Cell Physiol. 51: 448–462.

Matsumoto, T., Lian, H.L., Su, W.A., Tanaka, D., Liu, C., Iwasaki, I. et al. (2009) Role of the aquaporin PIP1 subfamily in the chilling tolerance of rice. Plant Cell Physiol. 50: 216–229.

Miller, J.A., Horvath, S. and Geschwind, D.H. (2010) Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways. Proc. Natl Acad. Sci. USA 107: 12698–12703.

Morishita, H. and Shinozaki, K. (2010) Genomics and bioinformatics resources for crop improvement. Plant Cell Physiol. 51: 497–523.

Nemati, M. and Greenacre, M. (2007) Correspondence analysis in R, with two- and three-dimensional graphics: the ca package. J. Stat. Software 20: 1–13.

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., Kinoshita, K., Nakai, K., Shibata, M., Hayashi, S., Saeki, M. et al. (2007) ATTED-II: a database of co-expressed genes and cis elements for identifying co-regulated gene groups in Arabidopsis. Nucleic Acids Res. 35: D636–D639.

Okuda, S., Yamada, T., Hamajima, I., Itoh, M., Katayama, T., Bork, P. et al. (2008) KEGG Atlas mapping for global analysis of metabolic pathways. Nucleic Acids Res. 36: W423–W426.

Ostlund, G., Schmit, T., Forslund, K., Köstler, T., Messina, D., Roopra, S. et al. (2010) InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Res. 38: D196–D203.

Ouyang, S., Zhu, W., Hamilton, J., Lin, H., Campbell, M., Childs, K. et al. (2007) The TIGR Rice Genome Annotation Resource: improvements and new features. Nucleic Acids Res. 35: D883–D887.

Parkinson, H., Kapushesky, M., Kolesnikov, N., Rustici, G., Shojaatlab, M., Abeygunawardena, N. et al. (2009) ArrayExpress update—from an archive of functional genomics experiments to the atlas of gene expression. Nucleic Acids Res. 37: D1028–D1033.

Rice Annotation Project. (2008) The Rice Annotation Project Database (RAP-DB): 2008 update. Nucleic Acids Res. 36: D1028–D1033.

Roe, S. et al. (2008) Arabidopsis Data Integration. The Rice Annotation Project Database (RAP-DB): 2008 update. Nucleic Acids Res. 36: D1028–D1033.

Roessner-Tunali, U., Urbanczyk-Wochniak, E., Czechowski, T., Kolbe, A., Willmitzer, L. and Fernie, A.R. (2003) De novo amino acid analysis for their valuable comments on statistical methods.
acid biosynthesis in potato tubers is regulated by sucrose levels. *Plant Physiol.* 133: 683–692.

Sakuraba, Y., Yokono, M., Akimoto, S., Tanaka, R. and Tanaka, A. (2010) Deregulated chlorophyll b synthesis reduces the energy transfer rate between photosynthetic pigments and induces photodamage in *Arabidopsis thaliana*. *Plant Cell Physiol.* 51: 1055–1065.

Sawada, Y., Akiyama, K., Sakata, A., Kuwahara, A., Otsuki, H., Sakurai, T. et al. (2009a) Widely targeted metabolomics based on large-scale MS/MS data for elucidating metabolite accumulation patterns in plants. *Plant Cell Physiol.* 50: 37–47.

Sawada, Y., Toyooka, K., Kuwahara, A., Sakata, A., Nagano, M., Saito, K. et al. (2009b) Arabidopsis bile acid: sodium symporter family protein 5 is involved in methionine-derived glucosinolate biosynthesis. *Plant Cell Physiol.* 50: 1579–1586.

Shikata, M., Koyama, T., Mitsuda, N. and Ohme-Takagi, M. (2009) Arabidopsis SBP-box genes SPL10, SPL11 and SPL2 control morphological change in association with shoot maturation in the reproductive phase. *Plant Cell Physiol.* 50: 2133–2145.

Shinozaki, K. and Sakakibara, H. (2009) Omics and bioinformatics: an essential toolbox for systems analyses of plant functions beyond 2010. *Plant Cell Physiol.* 50: 1177–1180.

Snedecor, G.W. and Cochran, W.G. (1989) Partial correlation. In *Statistical Methods*, 8th edn pp. 363–373. Blackwell Publishing, Iowa.

Soeno, K., Goda, H., Ishii, T., Ogura, T., Tachikawa, T., Sasaki, E. et al. (2010) Auxin biosynthesis inhibitors, identified by a genomics-based approach, provide insights into auxin biosynthesis. *Plant Cell Physiol.* 51: 524–536.

Suwabe, K., Suzuki, G., Takahashi, H., Shiono, K., Endo, M., Yano, K. et al. (2008) Separated transcriptomes of male gametophyte and tapetum in rice: validity of a laser microdissection (LM) microarray. *Plant Cell Physiol.* 49: 1407–1416.

Swanson-Wagner, R.A., DeCook, R., Jia, Y., Bancroft, T., Ji, T., Zhao, X. et al. (2009) Paternal dominance of trans-eQTL influences gene expression patterns in maize hybrids. *Science*. 326: 1118–1120.

Tsaparas, P., Mariño-Ramirez, L., Bodenreider, O., Koonin, E.V. and Jordan, I.K. (2006) Global similarity and local divergence in human and mouse gene co-expression networks. *BMC Evol. Biol.* 6: 70.

Usadel, B., Poree, F., Nagel, A., Lohse, M., Czedik-Eysenberg, A. and Stitt, M. (2009) A guide to using MapMan to visualize and compare Omics data in plants: a case study in the crop species, maize. *Plant Cell Environ.* 32: 1211–1229.

Watanabe, M. (2008) Towards a comprehensive understanding of molecular mechanisms of sexual reproduction in higher plants. *Plant Cell Physiol.* 49: 1404–1406.

Whittaker, A., Marotti, I., Dinelli, G., Calamai, L., Romagnoli, S., Manzelli, M. et al. (2010) The influence of tuber mineral element composition as a function of geographical location on acrylamide formation in different Italian potato genotypes. *J. Sci. Food Agric.* 90: 1968–1976.

Yamagishi, K., Tatematsu, K., Yano, R., Preston, J., Kitamura, S., Takahashi, H. et al. (2009) CHOTTO1, a double AP2 domain protein of *Arabidopsis thaliana*, regulates germination and seedling growth under excess supply of glucose and nitrate. *Plant Cell Physiol.* 50: 330–340.

Yano, K., Imai, K., Shimizu, A. and Hanashita, T. (2006) A new method for gene discovery in large-scale microarray data. *Nucleic Acids Res.* 34: 1532–1539.

Zeng, T. and Li, J. (2010) Maximization of negative correlations in time-course gene expression data for enhancing understanding of molecular pathways. *Nucleic Acids Res.* 38: e1.

Zhang, D.S., Liang, W.Q., Yuan, Z., Li, N., Shi, J., Wang, J. et al. (2008) Tapetum degeneration retardation is critical for aliphatic metabolism and gene regulation during rice pollen development. *Mol. Plant.* 1: 599–610.