Published online 11 October 2007 *Nucleic Acids Research, 2008, Vol. 36, Database issue* ***D77–D82***
doi:10.1093/nar/gkm840

COXPRESdb: a database of coexpressed gene networks in mammals

Takeshi Obayashi[1], Shinpei Hayashi[2], Masayuki Shibaoka[2], Motoshi Saeki[2], Hiroyuki Ohta[3] and Kengo Kinoshita[1,4,*]

[1]Human Genome Center, Institute of Medical Science, The University of Tokyo, 4-6-1 Shirokane-dai, Minato-ku, Tokyo 108-8639, [2]Graduate School of Information Science and Engineering, Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro-ku, Tokyo 152-8550, [3]Center of Biological Resources and Informatics, Tokyo Institute of Technology, 4259-B65, Nagatsuta-cho Midori-ku, Yokohama 26-8501 and [4]Structure and Function of Biomolecules, SORST, JST, 4-1-8 Honcho, Kawaguchi, Saitama 332-0012, Japan

Received August 14, 2007; Revised September 23, 2007; Accepted September 25, 2007

ABSTRACT

A database of coexpressed gene sets can provide valuable information for a wide variety of experimental designs, such as targeting of genes for functional identification, gene regulation and/or protein–protein interactions. Coexpressed gene databases derived from publicly available GeneChip data are widely used in Arabidopsis research, but platforms that examine coexpression for higher mammals are rather limited. Therefore, we have constructed a new database, COXPRESdb (coexpressed gene database) (http://coxpresdb.hgc.jp), for coexpressed gene lists and networks in human and mouse. Coexpression data could be calculated for 19 777 and 21 036 genes in human and mouse, respectively, by using the GeneChip data in NCBI GEO. COXPRESdb enables analysis of the four types of coexpression networks: (i) highly coexpressed genes for every gene, (ii) genes with the same GO annotation, (iii) genes expressed in the same tissue and (iv) user-defined gene sets. When the networks became too big for the static picture on the web in GO networks or in tissue networks, we used Google Maps API to visualize them interactively. COXPRESdb also provides a view to compare the human and mouse coexpression patterns to estimate the conservation between the two species.

INTRODUCTION

Gene coexpression provides key information to understand living systems because coexpressed genes are often involved in the same or related biological pathways (1). Coexpression data are now used for a wide variety of experimental designs, such as gene targeting, regulatory investigations and/or identification of potential partners in protein–protein interactions (PPIs) (2,3). Large-scale gene expression data are required to obtain reliable coexpression information. DNA microarray data represent one of the most abundant sources of gene expression data, which are now stored in public gene expression repositories (4–6). Therefore, it would be an appropriate time to establish a secondary database for coexpressed genes, using the large amount of publicly available DNA microarray data.

In the Arabidopsis field, several coexpression databases have been constructed and are widely used by researchers (7–12). On the other hand, mammalian coexpression information is rather limited because there is no established mammalian database. For example, genevestigator (for mouse and rat) (12) provides a sophisticated interface to check gene expression patterns using Java technology, but the coexpression data can only be obtained for the query pair of genes, i.e. information on coexpressed gene networks is not available. SymAtlas (for human, mouse and rat) (10) uses an advanced search interface and yields sets of coexpressed genes, but it does not provide a quantitative measure of the coexpression strength.

Here we report a new database, COXPRESdb (coexpressed gene database), which provides the coexpressed gene networks and the coexpressed gene lists *ordered by* the strength of coexpression for human and mouse. COXPRESdb provides four types of coexpressed gene networks: (i) highly coexpressed genes, (ii) genes with the same GO annotation, (iii) genes expressed in the same tissue and (iv) user-defined gene sets. COXPRESdb also prepares a cross-species view to compare the coexpression networks in human and mouse because conserved coexpression patterns may enhance the reliability of the coexpressed network and can be used

*To whom correspondence should be addressed. Tel: +81 3 5449 5131; Fax: +81 3 5449 5133; Email: kino@ims.u-tokyo.ac.jp

to identify possible PPIs more effectively (13). Notably, the known PPIs in HPRD (14) are also shown in the coexpression networks.

DATABASE CONTENTS

COXPRESdb contains the coexpressed gene networks for 40 813 genes (19 777 for human and 21 036 for mouse), 1820 GO terms and 63 human tissues. All expression data for this database are based on the Affymetrix GeneChip (Human Genome U133 Plus 2.0 Array and Mouse Genome 430 2.0 Array), information on which has been released by NCBI GEO (4).

COXPRESdb mainly consists of three types of pages: *gene page*, *GO network page* and *tissue-specific network page*. Representative usages of these pages are described in Examples 1 and 2 below.

The *gene page* is the central page, which is constructed for every gene defined by NCBI Entrez (15) regardless of the existence of expression data. The *gene page* is composed of three sections: (i) functional annotation, (ii) gene coexpression and (iii) gene expression. The names of each part are highlighted by green bars on the *gene page* (Figure 1C). Each *gene page* has the URL, such as http://coxpresdb.hgc.jp/data/gene/(EntrezGeneID).html, and thus the external database can be directly linked to this page.

(i) The functional annotation section provides functional gene annotations obtained from NCBI (15), GO (16) and KEGG (17), as well as protein subcellular localization predicted by WoLF PSORT (18). (ii) The gene coexpression section provides the coexpressed gene network(s) relevant to this gene. The *gene page* lists the 20 most highly coexpressed genes based on expression pattern similarity (Figure 1C). Only the top 20 genes are shown because the expression similarity rapidly decreases after the top 20 genes, on average (data not shown), and a large number of elements in a single network are difficult to see on a single web page. A list of the top 300 coexpressed genes is also available to find any other related genes, and thus the user can draw the network containing more genes with a network drawing tool provided by COXPRESdb. The details of the coexpression can be seen by following the links to the 'coexpression detail' in Figure 1C. To focus on the coexpression network for each tissue, *tissue-specific network pages* are available (see Example 1 and Figure 1C and D). (iii) The gene expression section shows the gene expression pattern(s) of the corresponding probeset(s). The tissue-specific gene expression pattern is also shown.

To compare the gene information for orthologous gene sets, *ortholog pages* are prepared in which the information in the *gene page* for human and that from the corresponding mouse *gene page* are presented in parallel. To identify the homologous gene set, we used HomoloGene (15) data, in which 16 981 gene sets were defined.

A *GO network page* is constructed for each GO term (16). The 30 most highly coexpressed genes are selected, and their networks are drawn in parallel views for human and mouse (Figure 2C). The GO networks as well as tissue-specific networks are constructed based on the same coexpression data as presented in the *gene page*, and the difference is in the selection of genes to be drawn. There are 6623 GO terms, and the networks are depicted as 1820 GO terms. The other terms are not considered because they have no highly coexpressed gene pairs with *mutual rank* (MR) <50 (see DATA SOURCES AND CALCULATION for MR).

The *tissue-specific network page* is constructed for 63 human tissues using the annotation of gene expression in HPRD (14) (Figure 1D). The global picture of a coexpressed gene network in a tissue is too large to be visualized in a single picture on a static page, and therefore we employed Google Maps API (Application Programming Interface, http://www.google.com/apis/maps/) to interactively navigate the huge coexpression networks.

Example 1: functional estimation of the gene, *ACTR3*

The human gene *ACTR3* served as a model to describe the functionalities of COXPRESdb. This protein is one of the seven subunits of the Arp2/3 complex that regulates F-actin formation at lamellipodial protrusions (19). (The following steps are also shown as a tutorial in COXPRESdb at http://coxpresdb.hgc.jp/help/tutorial/1.html.)

(1) Search 'actin related protein 3' as a keyword, using the search form to the right of the title logo on the top page (Figure 1A). In the initial setting, the human annotations are searched, but this can be changed by the toggle switch under the search box. The BLAST search against COXPRESdb is also provided as a search page.
(2) As a result, *ACTR3* is found in the list, and the user can see its *gene page* by clicking the 'symbol' of *ACTR3* (Figure 1B).
(3) In this example, *ACTR3* is surrounded by genes for actin regulation (Figure 1C). *ACTR2*, *ARPC2*, *ARPC3* and *ARPC5* are other components of the Arp2/3 complex, which are supported by PPIs (red edges), and *TMOD3* and *CAPZA1* encode capping proteins at the pointed-end and the barbed-end of F-actin with supports from homologs (orange edges). In short, this coexpression network clearly reflects the direct functional partners of *ACTR3*.

In the expression section, the tissue-specific gene expression pattern indicates that this gene is expressed in immune system organs, veins and oral tissues, as supported by four of the five GeneChip probesets (with the exception of 239170_at probe). To consult the coexpressed gene networks in these expressed tissues, the coexpressed gene networks for foetus and neutrophil are provided. Click 'Neutrophil' as a representative of the immune system organs.

(4) On the *tissue-specific network page*, *ACTR3* is automatically placed at the centre of the window and highlighted with a yellow symbol (Figure 1D). In the neutrophil, the weak edges on the *gene page*

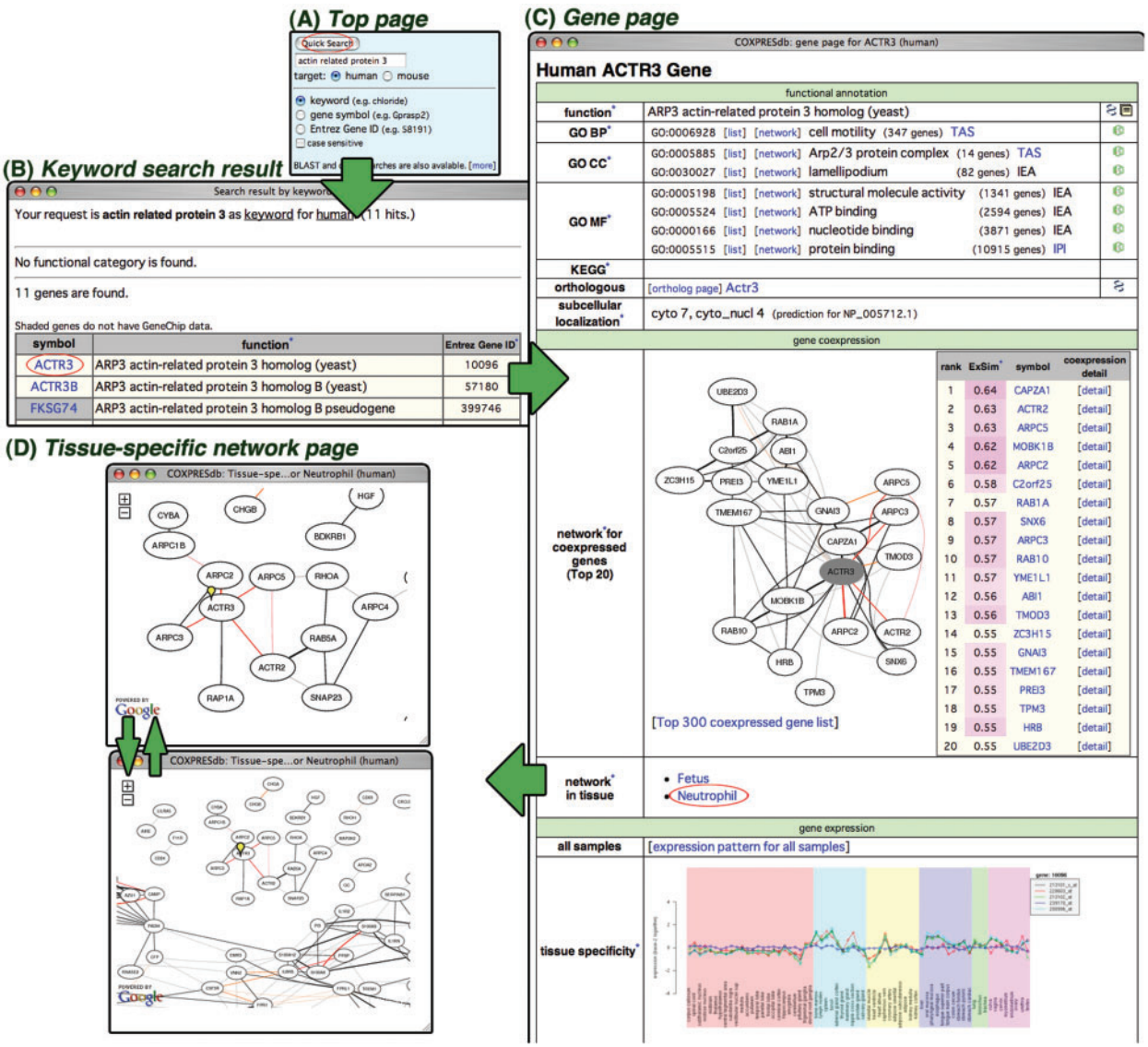

Figure 1. An example of the estimation of gene functions using COXPRESdb. The thin red circles in (A), (B) and (C) indicate the link to the next step indicated by green arrows. (**A**) Top page. (**B**) Keyword search result. (**C**) *Gene page* for *ACTR3*. In the functional annotation section, the 'list' is linked to the list of genes, including the GO term and all GO subcategory terms. The 'network' is linked to *GO network page* (see Figure 2C). When a GO annotation is linked to any publication, the GO evidence codes (TAS and IPI in this example) of the GO annotation is linked to a corresponding PubMed page. The 'ortholog page' is linked to the page for parallel view of the *gene page* for this gene and its mouse ortholog. Small icons on the right side of the page are external links. Subcellular localizations are predicted by WoLF PSORT software, the score for which is scaled from 0 (less reliable) to 10 (most reliable), in integers. In the coexpression section, a coexpression network for the 20 highest coexpressed genes to the query gene is provided. The network consists of nodes and edges. Each node represents a coexpressed gene, which works as a link to the corresponding *gene page*, and each edge indicates the *strength of coexpression* by its thickness (see DATA SOURCES AND CALCULATIONS for details). The shaded node highlights the current gene, and the red edges, if any, indicate existence of reported PPIs according to the HPRD annotation (14). The orange edges indicate the existence of strong coexpression between the corresponding orthologs in mouse. The gene list is located in a table next to the network. The value of each expression pattern similarity (ExSim) in the table is highlighted by red colour corresponding to the edge thicknesses. The 'detail' in the table is linked to the *coexpression viewer*, which illustrates the contribution of each sample to the expression pattern similarity. See the text for the explanation of the other section shown. (**D**) Tissue-specific network page for neutrophils. *ACTR3* is marked with a yellow symbol.

have disappeared, suggesting that those weakly coexpressed genes are coexpressed in other tissues or samples. In the zoom-out view, this coexpression group, including *ACTR3*, is composed of distinct gene groups for actin regulation, which includes all seven subunits of the Arp2/3 complex and other genes involved in actin regulation (*RHOA*, *MAP2K2* and *RAP1A*).

Example 2: obtaining coexpressed genes for a particular function

In the previous example, we introduced the central *gene page* and the *tissue-specific network page*. Here, we introduce the *GO network page* using the human viral defence system as an example. Humans have a complicated and well-developed system to counter viral challenge.

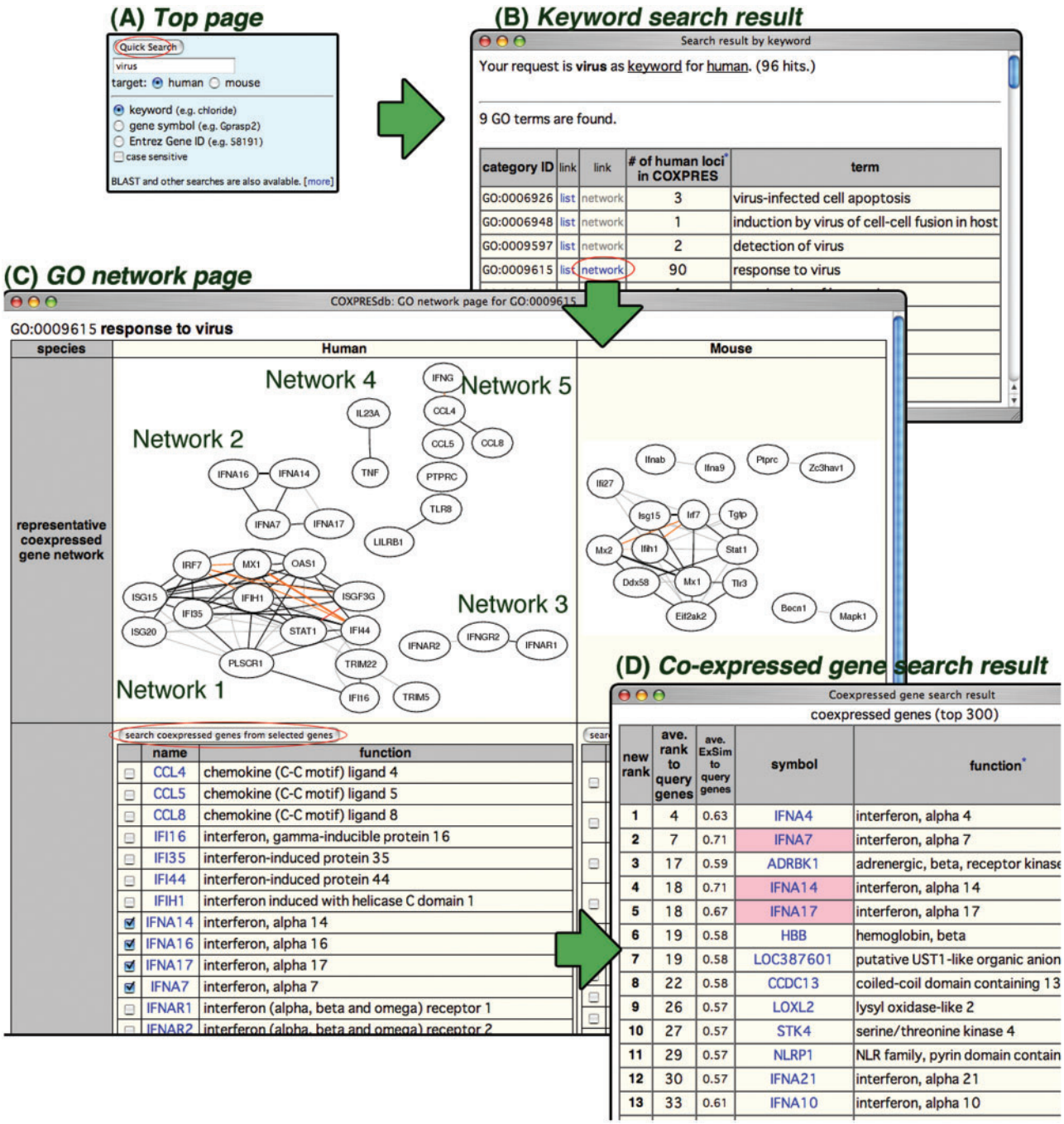

Figure 2. An example of the listing for genes related to virus-response using COXPRESdb. The thin red circles in (A), (B) and (C) indicate the link to the next step indicated by green arrows. (**A**) Top page. (**B**) Keyword search result. The 'list' is linked to the list of the genes including the GO term and all GO subcategory terms. The 'network' is linked to the *GO network page*. (**C**) *GO network page*. See text for details. (**D**) Results of the coexpressed gene search. Gene symbols in the pink box indicate query genes from the previous step.

Identification of the genes in a system is the first step for deeper understanding of the system. For gene identification, it is efficient to list and classify the genes of co-functional candidates using coexpressed genes. For this purpose, COXPRESdb could be used as follows (Figure 2). (Also shown at http://coxpresdb.hgc.jp/help/tutorial/2.html.)

(1) Search 'virus' as a keyword using the search form on the top page (Figure 2A), and follow the 'response to virus' link in the search result table (Figure 2B).
(2) The *GO network page* provides the coexpressed gene networks of 30 highly coexpressed genes in the query GO term (Figure 2C). The table under the network contains more detailed gene descriptions.

Five networks for human and four for mouse can be found for the GO term (Figure 2C). To deduce the biological meaning of each network, the information in the external links on each *gene page* is useful in addition to the information presented in COXPRESdb. As a result of careful inspection of each *gene page*, the five gene networks seem to correspond to a biological function in response to a virus, as follows: (i) interferon-responsive gene network, (ii) interferon α gene network, (iii) interferon receptor network, (iv) interleukin and tumour necrosis factor network and (v) three chemokines and other networks. Two additional networks, corresponding to networks 1 and 2, are also found in mouse.

(3) The GO network page provides the network of the genes that are already annotated as 'response to virus'. To find other components lacking the annotation but relating to virus response, it would be promising to search for coexpressed genes from these groups. This can be done easily in COXPRESdb by activating the check box on the left side of the table and pushing the button 'search coexpressed genes from selected genes' (Figure 2D). This will provide a list of the top 300 highly coexpressed genes. The list contains putative functional and unknown genes in addition to known interferon α genes. In the same way, the user can obtain coexpressed genes for other gene groups. Finally, entire gene lists containing putative virus-responsive genes can be obtained.

DATA SOURCES AND CALCULATIONS

Calculation of gene coexpression

To define reliable coexpressed genes, we constructed gene expression profiles using as many genes and samples as possible. Toward this end, GPL570 (Human Genome U133 Plus 2.0 Array: 54 614 probesets) and GPL1261 (Mouse Genome 430 2.0 Array: 45 037 probesets) were selected from NCBI GEO (4). Some samples were omitted due to different GeneChip usage, e.g. ChIP-on-chip or heterohybridization of close species. As a result, we used 3749 human and 2226 mouse samples as the raw data (CEL files). The correspondence between probes and genes is based on the NCBI annotations. Only the probes mapped to a single gene were used, resulting in 44 793 and 40 083 probes mapped to 19 777 and 21 036 genes for human and mouse, respectively.

We used the Robust Multi-array Average (RMA) method (20) for GeneChip normalization and weighted Pearson's correlation coefficients based on sample redundancies to measure probe-to-probe expression pattern similarity (8). The sample redundancy is calculated as the number of similar samples in the data set, and the sample similarity is measured by the correlation between samples. (See the help page at http://coxpresdb.hgc.jp/help/coex_cal.html for details.) Since most of the genes have multiple probes in the mammalian GeneChip, gene-to-gene expression pattern similarities (ExSims) are evaluated as the maximum value of corresponding probe-to-probe correlations from the corresponding probe-to-probe ExSims. All coexpression values can be downloaded from COXPRESdb in tab-delimitated text files.

Strength of coexpression for network and edge thickness

The ExSims were converted to *mutual rank (MR)* to evaluate the strength of coexpression. For any given pair, gene A and gene B, the MR is calculated as an average of the rank of gene B in the coexpressed genes to gene A (ordered by ExSims) and the average of the rank of gene A to gene B. For our coexpression data, the correlation rank and MR were a better measure of similarity than the correlation value to determine related genes (Obayashi *et al.*, unpublished results). This is partly because even the gene pair with a low ExSim can work together if no other genes are highly coexpressed, as in the example of human histone cluster—where one gene is highly coexpressed according to the MRs, although ExSims are lower than 0.5 (see http://coxpresdb.hgc.jp/help/mr.html). To draw the coexpression network, we used three thresholds to determine the thickness of edges: bold edges (MR <5), normal edges ($5\leq$ MR <30) and thin edges ($30\leq$ MR <50). The MR is also used to select genes in GO networks and tissue-specific networks, where genes are selected from highly coexpressed pairs up to a defined number (30 genes for GO network and 1000 genes for tissue-specific network).

Tissue-specific gene expression data

Data from GSE3526 for human (21) and GSE1986 for mouse in NCBI GEO were used for the tissue-specific gene expression graph on the *gene page*. After RMA normalization, the probe intensities were averaged for each tissue. These tissues were manually ordered and grouped from the viewpoints of tissue function and gene expression similarity. For the construction of the *tissue-specific network page*, HPRD data (release version 7.0) were used. The coexpressed gene networks were constructed for 63 tissues, in which more than 50 genes are highly coexpressed (MR <50).

ACKNOWLEDGEMENTS

This work was partially supported by a Grant-in Aid for Scientific Research on the Priority Area 'Transportsome' from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan to KK. Computation time was provided by the Super Computer System, Human Genome Center, Institute of Medical Science, The University of Tokyo. Funding to pay the Open Access publication charges for this article was provided by MEXT of Japan.

Conflict of interest statement. None declared.

REFERENCES

1. Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.
2. 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.
3. Shoemaker,B.A. and Panchenko,A.R. (2007) Deciphering protein-protein interactions. Part I. Experimental techniques and databases. *PLoS Comput. Biol.*, **3**, e42.
4. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Rudnev,D., Evangelista,C., Kim,I.F., Soboleva,A., Tomashevsky,M. *et al.* (2007) NCBI GEO: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Res.*, **35**, D760–D765.
5. Parkinson,H., Sarkans,U., Shojatalab,M., Abeygunawardena,N., Contrino,S., Coulson,R., Farne,A., Lara,G.G., Holloway,E. *et al.* (2005) ArrayExpress–a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.*, **33**, D553–D555.
6. 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.
7. Manfield,I.W., Jen,C.H., Pinney,J.W., Michalopoulos,I., Bradford,J.R., Gilmartin,P.M. and Westhead,D.R. (2006) Arabidopsis Co-expression Tool (ACT): web server tools for microarray-based gene expression analysis. *Nucleic Acids Res.*, **34**, W504–W509.
8. Obayashi,T., Kinoshita,K., Nakai,K., Shibaoka,M., Hayashi,S., Saeki,M., Shibata,D., Saito,K. and Ohta,H. (2007) ATTED-II: a database of co-expressed genes and cis elements for identifying co-regulated gene groups in Arabidopsis. *Nucleic Acids Res.*, **35**, D863–D869.
9. Steinhauser,D., Usadel,B., Luedemann,A., Thimm,O. and Kopka,J. (2004) CSB.DB: a comprehensive systems-biology database. *Bioinformatics*, **20**, 3647–3651.
10. Su,A.I., Cooke,M.P., Ching,K.A., Hakak,Y., Walker,J.R., Wiltshire,T., Orth,A.P., Vega,R.G., Sapinoso,L.M. *et al.* (2002) Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl Acad. Sci. U S A*, **99**, 4465–4470.
11. Toufighi,K., Brady,S.M., Austin,R., Ly,E. and Provart,N.J. (2005) The Botany Array Resource: e-Northerns, Expression Angling, and promoter analyses. *Plant J.*, **43**, 153–163.
12. Zimmermann,P., Hennig,L. and Gruissem,W. (2005) Gene-expression analysis and network discovery using Genevestigator. *Trends Plant Sci.*, **10**, 407–409.
13. Bhardwaj,N. and Lu,H. (2005) Correlation between gene expression profiles and protein-protein interactions within and across genomes. *Bioinformatics*, **21**, 2730–2738.
14. Peri,S., Navarro,J.D., Amanchy,R., Kristiansen,T.Z., Jonnalagadda,C.K., Surendranath,V., Niranjan,V., Muthusamy,B., Gandhi,T.K. *et al.* (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res.*, **13**, 2363–2371.
15. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R. *et al.* (2007) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **35**, D5–D12.
16. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
17. Kanehisa,M., Goto,S., Hattori,M., Aoki-Kinoshita,K.F., Itoh,M., Kawashima,S., Katayama,T., Araki,M. and Hirakawa,M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, **34**, D354–D357.
18. Horton,P., Park,K.J., Obayashi,T., Fujita,N., Harada,H., Adams-Collier,C.J. and Nakai,K. (2007) WoLF PSORT: protein localization predictor. *Nucleic Acids Res.*, **35**, W585–W587.
19. Welch,M.D., DePace,A.H., Verma,S., Iwamatsu,A. and Mitchison,T.J. (1997) The human Arp2/3 complex is composed of evolutionarily conserved subunits and is localized to cellular regions of dynamic actin filament assembly. *J. Cell Biol.*, **138**, 375–384.
20. Irizarry,R.A., Hobbs,B., Collin,F., Beazer-Barclay,Y.D., Antonellis,K.J., Scherf,U. and Speed,T.P. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.
21. Roth,R.B., Hevezi,P., Lee,J., Willhite,D., Lechner,S.M., Foster,A.C. and Zlotnik,A. (2006) Gene expression analyses reveal molecular relationships among 20 regions of the human CNS. *Neurogenetics*, **7**, 67–80.