W320–W326 *Nucleic Acids Research, 2008, Vol. 36, Web Server issue* *Published online 14 May 2008*
doi:10.1093/nar/gkn292

GeneCAT—novel webtools that combine BLAST and co-expression analyses

Marek Mutwil[1,2,*], Jens Øbro[1], William G. T. Willats[1] and Staffan Persson[2]

[1]Department of Molecular Biology, University of Copenhagen, Ole Maaløes vej 5, 2200 Copenhagen N, Denmark and [2]Max-Planck-Institute for Molecular Plant Physiology, Am Muehlenberg 2, 14476 Potsdam, Germany

Received January 3, 2008; Revised April 14, 2008; Accepted April 29, 2008

ABSTRACT

The gene co-expression analysis toolbox (GeneCAT) introduces several novel microarray data analyzing tools. First, the multigene co-expression analysis, combined with co-expressed gene networks, provides a more powerful data mining technique than standard, single-gene co-expression analysis. Second, the high-throughput Map-O-Matic tool matches co-expression pattern of multiple query genes to genes present in user-defined subdatabases, and can therefore be used for gene mapping in forward genetic screens. Third, Rosetta combines co-expression analysis with BLAST and can be used to find 'true' gene orthologs in the plant model organisms *Arabidopsis thaliana* and *Hordeum vulgare* (Barley). GeneCAT is equipped with expression data for the model plant *A. thaliana*, and first to introduce co-expression mining tools for the monocot Barley. GeneCAT is available at http://genecat.mpg.de

INTRODUCTION

The ability to measure the activity of several thousands of genes simultaneously has revolutionized the way we currently view biological processes. Substantial amounts of such expression data that represent experiments from a variety of tissues, developmental stages and stimuli, are currently publicly available for different organisms. Widely used public microarray data repositories are ArrayExpress (1) and Gene Expression Omnibus (GEO). As each microarray experiment often generates large amounts of expression data, it is often difficult for researchers without background in bioinformatics to extract the information they seek. Several web-based tools that analyze collections of publicly available microarray data for the plant model organism *Arabidopsis thaliana* have therefore been developed including Genevestigator (2), Arabidopsis Co-expression Tool (3), Botany Array Resource (4), CSB.DB (5) and ATTED-II (6). These tools provide comparative gene analyses including *cis*-element prediction, expression profiling and co-expression analysis. In addition, a tool that combines co-expression and predicted protein–protein interactions has recently been developed (7). It therefore appears that future webtools will combine different types of data to facilitate a more complex and multidimensional view of organisms such as *Arabidopsis*.

Several studies exploit the fact that genes which are functionally related may be transcriptionally coordinated (8,9). Recent studies have shown that this is also the case in plants (10–13). Consequently, most of the current web-based tools are mainly focused on retrieving expression and/or co-expression patterns for individual genes. We have extended and refined this process and produced several new tools under the banner gene co-expression analysis toolbox (GeneCAT). This platform provides the user with both standard co-expression tools, such as gene clustering and expression profiling, and also includes tools that use multiple bait genes and makes functional inferences across different organisms by combining BLAST and co-expression. GeneCAT is pre-loaded with datasets for two plant model organisms, *Arabidopsis* and Barley, and dataset from other species can readily be added. To increase the accessibility to the tools we have made GeneCAT accessible via the web (http://genecat.mpg.de).

MATERIALS AND METHODS

Implementation and calculation

GeneCAT is running on Apache server using cgi to link html forms with Python scripts. PhyFi (14) and Graphviz (www.graphviz.org) are used for visualization of ExpressionTree and co-expressed gene network, respectively. Calculations are performed on the fly by Python scripts.

Microarray data sources and processing

Databases for *Arabidopsis* and Barley use Affymetrix ATH1 (22 810 probe sets) and Barley1 (22 840 probe sets) GeneChips, respectively. *Arabidopsis thaliana* microarray

*To whom correspondence should be addressed. Tel: +49 331 567 8149; Fax: +49 331 567 8408; Email: mutwil@mpimp-golm.mpg.de

datasets consisting of 351 RMA normalized ATH1 microarray data were obtained from TAIR (15). Separate *A. thaliana* tissue atlas dataset of 121 microarrays used for ExpressionProfiling was generated by the AtGenExpress project (16) and obtained from TAIR. For the Barley tissue, atlas 64 MAS5 normalized microarray datasets were obtained from the BarleyBase (17) and was created by (18).

RESULTS AND DISCUSSION

GeneCAT provides expression analyzing tools for two major model organisms in plant biology; *Arabidopsis* and Barley. To provide an easy introduction to the application of the GeneCAT tools, we present them individually and give biological example for how each tool may be used. A more detailed description of the different tools can be found on http://genecat.mpg.de FAQ section.

Expression profiling and tree view—cellulose synthases

The ExpressionProfiling tool generates line plots of expression profiles across different tissues for specified genes in *Arabidopsis* and Barley. The ExpressionTree tool uses these data to generate dendrograms corresponding to the tightness of co-expression for the same set of genes. Since tools similar to the ExpressionProfiling tool are also present at other co-expression databases we chose to exemplify only the ExpressionTree tool using the cellulose synthase (*CESA*) genes in *Arabidopsis* and Barley. There are 10 and at least 8 members of the CESA families in *Arabidopsis* and Barley, respectively. The current model for cellulose synthesis proposes that at least three different CESA proteins are assembled into a functional complex (19). Mutant analyses have shown that *AtCESA1*, *3* and *6*, and *AtCESA4*, *7* and *8* are necessary for primary and secondary cell wall cellulose synthesis in *Arabidopsis*, respectively (20–23). A similar divergence of the *CESA* genes is also predicted in Barley (24).

The 10 *CESA* genes from *Arabidopsis* were analyzed using the ExpressionTree tool (Supplementary Figure S1A). Two tight clusters were evident; one consisting of *AtCESA4*, *7* and *8* and the other including *AtCESA1*, *3* and *6* corresponding to secondary and primary cell wall biosynthesis, respectively (20–23). Interestingly, *AtCESA2* and *AtCESA5* are tightly associated with the primary cell wall *AtCESAs*, and have recently been implicated to be functionally redundant to *AtCESA6* (22,25). Similar results using co-expression analysis were also obtained by ref. (26). Analogous to *Arabidopsis*, the expression of the eight Barley *HvCESA*s create two tight clusters consisting of *HvCESA1*, *2* and *6*, and *HvCESA5*/*7*, *4* and *8* (Supplementary Figure S1B), suggesting that these groups of *HvCESA*s form functional complexes in Barley. These data are consistent with results obtained by q-RT-PCR (24). The high sequence similarity of *HvCESA5* and *HvCESA7* makes it impossible to distinguish between these homologs (24). This type of analysis may thus provide researchers with a platform to infer functionally related gene products in *Arabidopsis* and Barley.

Co-expression using multiple bait genes—suberin biosynthesis

Genes that are involved in related processes are often co-expressed (17). Co-expression analyses therefore generally use a bait gene with a known function to target transcriptionally coordinated genes. This approach typically returns a list of genes that appear co-expressed with the bait gene. However, it is difficult to prioritize genes that are most relevant to the process that the bait gene is involved in. It therefore appears that an enrichment of such genes would be highly appreciated by biologists. GeneCAT utilizes two approaches to enrich genes for a given function. First, two or more genes that are involved in functionally related processes may be used as bait genes to more accurately identify target genes. Second, target genes that are true positives should in general also exhibit significant transcriptional coordination to each other, thus forming clusters of co-expressed genes (27). Several other tools provide the opportunity to apply such approaches, but GeneCAT is first to relate network information to the list of co-expressed genes. This process is done in three steps. In the first step an average co-expressed gene list is calculated for the bait genes. In the second step, a co-expressed gene network is created by measuring mutual co-expression ranks between the top 50 genes from the list in a pair-wise manner. Any two nodes (genes) that are connected with bold, normal or dashed lines display mutual ranks smaller than 10, 20 or 50, respectively. Blue nodes indicate bait genes and genes connected to these baits are colored green, orange and red if they are linked to any of the bait genes with bold, normal or dashed lines, respectively. The third step implements the color codes from the network to the co-expressed gene list, thus highlighting genes that exhibit high transcriptional connectivity to the bait genes and other genes in the list.

Since genes that are co-expressed tend to be functionally related, a typical co-expression list includes genes with overlapping annotations. This implies that the gene products may be functionally redundant. Consequently, any phenotypic traits may be masked by functional compensation if one gene is deleted. To identify genes that may be functionally redundant cross-wise BLAST analyses are performed for the top 150 genes in the co-expressed gene list. This analysis may thus give biologists information about functionally redundant genes and therefore candidates for additional mutant analyses.

To illustrate how the co-expression tool works we use a multigene co-expression approach for the suberin biosynthesis pathway from L-phenylalanine at AraCyc (http://www.arabidopsis.org/biocyc/index.jsp) as an example. Suberin is a waxy, polymeric plant cell wall constituent that regulates water transport and protects against pathogen attacks (28). We used one of the genes directly associated with the pathway, *OMT1* (At5g54160) as an initial bait gene. Supplementary Table S1 shows that several other genes in the pathway are co-expressed with *OMT1* such as *PAL1* (At2g37040), *PAL2* (At3g53260), *C4H* (At2g30490) and a caffeoyl-CoA-methyltransferase (At4g34050). To enrich for other genes associated with suberin biosynthesis, we then used these genes together

Table 1. [a]Co-expression analysis using multiple bait genes involved in suberin synthesis

R-value	Affymetrix probe	Locus	Gene annotation
0.849	263845_at	At2g37040	Phenylalanine ammonia-lyase 1 (PAL1)
0.839	253276_at	At4g34050	Caffeoyl-CoA 3-*O*-methyltransferase, putative
0.797	251984_at	At3g53260	Phenylalanine ammonia-lyase 2 (PAL2)
0.768	248200_at	At5g54160	Quercetin 3-*O*-methyltransferase 1 (OMT1)
0.744	260913_at	At1g02500	S-adenosylmethionine synthetase 1 (SAM1)
0.744	267470_at	At2g30490	Trans-cinnamate 4-monooxygenase/cinnamic acid 4-hydroxylase (C4H)
0.739	256186_at	At1g51680	4-coumarate-CoA ligase 1/4-coumaroyl-CoA synthase 1 (4CL1)
0.725	248639_at	At5g48930	Transferase family protein, similar to anthranilate N-Hydroxycinnamoyl/benzoyltransferase
0.687	261933_at	At1g22410	2-dehydro-3-deoxyphosphoheptonate aldolase, putative
0.677	258047_at	At3g21240	4-coumarate-CoA ligase 2/4-coumaroyl-CoA synthase 2 (4CL2)
0.673	249910_at	At5g22630	Prephenate dehydratase family protein
0.649	262744_at	At1g28680	Transferase family protein
0.648	254192_at	At4g23850	Long-chain-fatty-acid-CoA ligase
0.642	260153_at	At1g52760	Esterase/lipase/thioesterase family protein
0.638	261749_at	At1g76180	Dehydrin (ERD14)
0.633	253277_at	At4g34230	Cinnamyl-alcohol dehydrogenase, putative
0.612	257771_at	At3g23000	CBL-interacting protein kinase 7 (CIPK7)
0.606	263426_at	At2g31570	Glutathione peroxidase, putative
0.602	256964_at	At3g13520	Arabinogalactan-protein (AGP12)
0.601	263838_at	At2g36880	S-adenosylmethionine synthetase, putative
0.598	250339_at	At5g11670	Malate oxidoreductase, putative
0.595	267212_at	At2g44060	Late embryogenesis abundant family protein/LEA family protein
0.594	246627_s_at	At2g45300	3-Phosphoshikimate 1-carboxyvinyltransferase
		At1g48860	3-Phosphoshikimate 1-carboxyvinyltransferase, putative
0.594	245780_at	At1g45688	Expressed protein
0.592	262237_at	At1g48320	Thioesterase family protein
0.591	248393_at	At5g52060	BAG domain-containing protein
0.590	247627_at	At5g60360	Cysteine proteinase, putative/AALP protein (AALP)
0.586	258852_at	At3g06300	Encodes a prolyl-4 hydroxylase
0.585	255552_at	At4g01850	S-adenosylmethionine synthetase 2 (SAM2)
0.585	264725_at	At1g22885	Expressed protein
0.584	263711_at	At2g20630	Protein phosphatase 2C, putative/PP2C, putative
0.560	256524_at	At1g66200	Glutamine synthetase, putative, similar to glutamine synthetase
0.579	254224_at	At4g23650	Calcium-dependent protein kinase, putative/CDPK
0.572	259516_at	At1g20450	Dehydrin (ERD10)
0.571	248573_at	At5g49720	endo-1,4-beta-glucanase KORRIGAN (KOR)
0.571	262619_at	At1g06550	enoyl-CoA hydratase/isomerase family protein

Five genes associated with suberin biosynthesis (blue color) were used as bait genes for the co-expression tool at GeneCAT.
[a]The table is truncated to comply with the journal format.

with the *OMT1* gene as bait genes for the multiple-bait gene co-expression analysis (Table 1; Figure 1). Several genes that are connected to shikimate, phenylpropanoid and chorismate biosynthesis are among the most highly ranked genes in the table. For example, two genes annotated as 4-coumarate-CoA-ligases (At1g51680 and At3g21240) are among the top-ranked genes (Table 1). These genes convert 4-coumarate into coumaryl-CoA linking the suberin biosynthesis and phenylpropanoid biosynthesis pathways. In addition, the crosswise BLAST analysis of the top 150 genes identified several putative homologs associated with suberin biosynthesis (Supplementary Table S2). These genes may consequently perform similar functions and may be considered as prime candidates for multiple mutant analyses. By using several connected bait genes for a given process it is therefore apparent that functionally associated genes may be enriched.

Forward genetics predictions using Map-O-Matic—photosynthesis

Identification of genes that correspond to phenotypic traits through forward genetic screens is typically time and resource consuming. The Map-O-Matic tool may be used to find genes that are likely to harbor mutations based on phenotypic similarities. The tool uses similar assumptions as regular co-expression approaches, namely, those genes involved in a specific biological process tend to be co-expressed.

To show how the Map-O-Matic tool works (Supplementary Figure S2) we have included an example based on photosynthesis. A mutant that is defective in photosynthesis was identified in *Arabidopsis* and the mutation was mapped to a genomic region of ~190 kb (29). This region is predicted to hold 57 genes, of which 49 were included on the ATH1 chip. To assess which of these genes that may be linked to the phenotypic trait we identified 47 genes that are associated with the keyword term '*photosystem'. We subsequently compiled a sub-database using the 47 genes associated with the photosystem search term, and ran cross-wise co-expression analyses between the 49 candidate genes and the 47 photosystem-associated genes using Map-O-Matic. The output from the Map-O-Matic analysis is displayed as a graph with the average co-expression for each of the

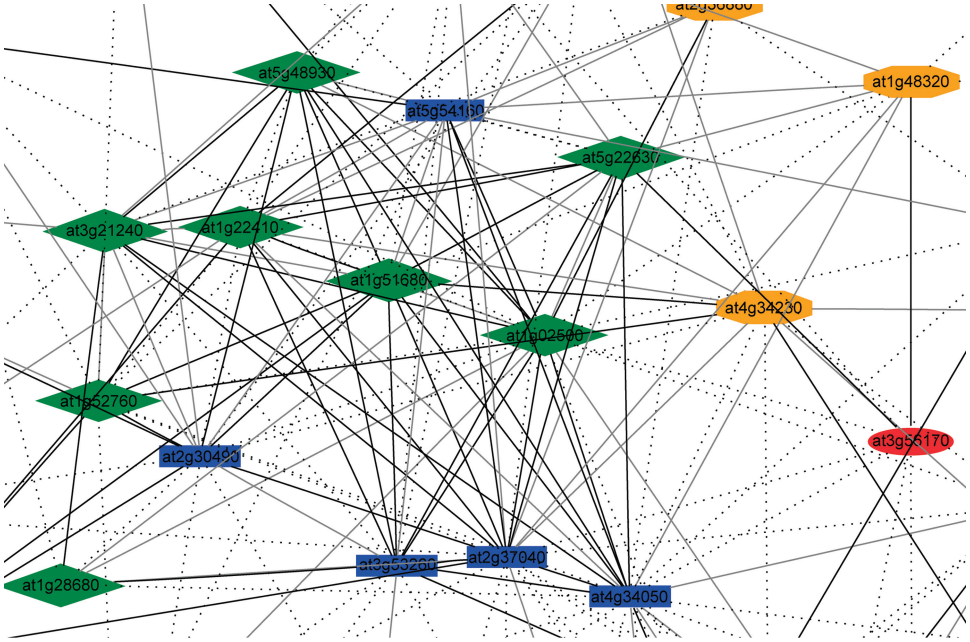

Figure 1. Co-expression network for multiple bait genes involved in suberin biosynthesis. Cropped co-expression network generated by the co-expression tool at GeneCAT using At2g37040, At4g34050, At3g53260, At5g54160 and At2g30490 as bait genes. Blue rectangular nodes indicate the bait genes for the analysis. Green diamonds, orange octagons and red ellipses indicate decreasing strength between node and the bait genes, respectively. Similarly, black, grey and dashed lines indicate decreasing strength between any two nodes.

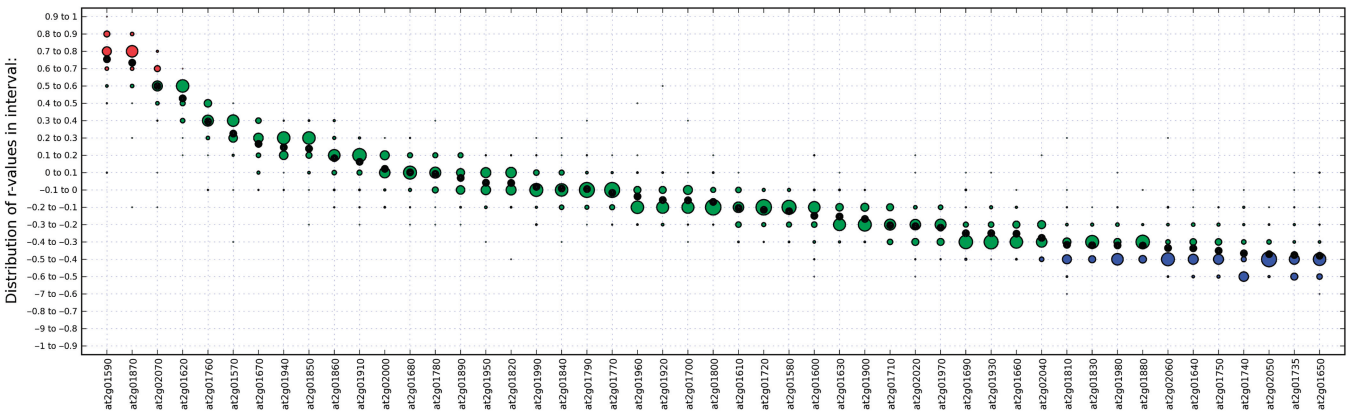

Figure 2. Map-O-Matic analysis of a photosynthesis mutant. Forty-nine genes corresponding to a genomic region of ~190 kb that was mapped for a photosynthetic defect was crosswise compared for co-expression with 47 genes associated with the keyword photosystem. Each of the 49 bait genes on the graph is ranked by average coefficient of correlation across the comparison with the 47 photosystem genes. The bait genes are displayed in descending order from left to right, according to average correlation coefficient (depicted as a black dot). Circle sizes are proportional to fraction of database genes correlating with *r*-values in given interval to a bait gene. Circles depicting *r*-values >0.6 are colored red, $r < -0.4$ are colored blue, while $0.6 > r > -0.4$ are colored green.

49 candidate genes against the 47 photosystem-associated genes (Figure 2). The top 5 genes of the 49 candidate genes are all highly co-expressed with most of the photosystem-associated genes (Figure 2). The gene that corresponded to the phenotypic trait was mapped to At2g01590 (29), which also was the gene that ranked as the most highly co-expressed gene with the photosystem genes of the 49 genes in the region. The gene ranked second in the analysis, At2g01870, is annotated as 'expressed protein'. Based on its high co-expression with the photosystem genes we suggest that this gene product may also play a direct role in photosynthetic processes. We believe that the Map-O-Matic tool is a powerful way to predict genes that are likely to be involved in specified biological processes.

Combining BLAST and co-expression using Rosetta—cellulose synthases

Orthologs in different species can be inferred through BLAST analyses and sequence comparison. These orthologs are then predicted to perform similar molecular functions in the different organisms. If they do perform similar functions we would also expect that other genes involved in the same process would have corresponding orthologs in the different species. Combining BLAST and

Table 2. [a]Rosetta analysis comparing primary and secondary cellulose biosynthesis in *A. thaliana*

Bait: At4g32410	Target 1: At5g44030
At4g32410 cellulose synthase *AtCesA1* At5g64740 cellulose synthase *AtCesA6* At5g05170 cellulose synthase *AtCesA3* At5g09870 cellulose synthase *AtCesA5* At4g39350 cellulose synthase *AtCesA2*	At5g17420 cellulose synthase *AtCesA7* At5g44030 cellulose synthase *AtCesA4* At4g18780 cellulose synthase *AtCesA8*
At5g60920 phytochelatin synthetase (COBRA)	At5g15630 COBRA-like 4
At1g05850 chitinase-like protein 1 (CTL1)	At3g16920 CTL2
At5g49720 endo-1, 4-beta-glucanase (KOR)	At1g19940 glycosyl hydrolase family 9 protein
At3g23820 NAD-dependent epimerase/dehydratase	At2g28760 NAD-dependent epimerase/dehydratase
	At5g59290 UDP-glucuronic acid decarboxylase (UXS3)
At4g12880 plastocyanin-like domain-containing protein	At5g26330 plastocyanin-like domain-containing protein At3g27200 plastocyanin-like domain-containing protein At1g72230 plastocyanin-like domain-containing protein At1g22480 plastocyanin-like domain-containing protein
At5g03040 calmodulin-binding family protein	At2g33990 similar to calmodulin-binding protein At3g59690 calmodulin-binding family protein At3g15050 calmodulin-binding family protein
At3g16850″glycoside hydrolase family 28 protein	At3g42950 glycoside hydrolase family 28 protein
	At1g80170 polygalacturonase, putative
At1g75500 nodulin MtN21 family protein	At3g45870 integral membrane family protein/nodulin
At3g15480 expressed protein	At4g27435 expressed protein
At1g41830 multicopper oxidase type I family protein	At5g03260 laccase, putative At2g38080 laccase, putative At5g01190 similar to laccase At5g05390 laccase, putative At2g29130 laccase, putative At5g60020 laccase, putative
At3g02350 glycosyl transferase family 8 protein	At5g54690 glycosyl transferase family 8 protein At1g19300 glycosyl transferase family 8 protein
At5g12250 tubulin beta-6 chain (TUB6) At1g20010 tubulin beta-5 chain (TUB5)	At5g12250 tubulin beta-6 chain (TUB6) At5g23860 tubulin beta-8 chain (TUB8)

Probe sets displaying BLAST score $e \leq 10^{-7}$ are placed in the same row by Rosetta.
[a]The table is truncated to comply with the journal format.

co-expression analyses may consequently reveal 'true' orthologous processes that are conserved in different organisms.

Plant species typically contain comparatively large gene families (30). This implies that several gene products may perform similar functions in different organs, tissues and/or developmental stages. It may therefore also be relevant to compare co-expression lists between these homologs to investigate functional conservation within a single species.

To demonstrate the application of the Rosetta tool (Supplementary Figure S3), we compared the cellulose synthesis machineries both in *Arabidopsis* and between *Arabidopsis* and Barley. Primary cell wall *CESA1* and secondary cell wall *CESA4* in *Arabidopsis* were used as bait and target (i.e. *Arabidopsis* versus *Arabidopsis*), respectively, for the Rosetta analysis. Using this bait and target, Rosetta identified *AtCESA1*, *AtCESA3*, *AtCESA6* and *AtCESA4*, *AtCESA7*, *AtCESA8* as being associated with primary and secondary cell wall synthesis,

respectively, based on the individual genes co-expression profiles (Table 2). Present in the co-expression lists were also genes that are common between the two processes. These include *COBRA* (At5g60920) and *CTL1* (At1g05850) and *COBRA*-like 4 (At5g15630) and *CTL2* (At3g16920) that are associated with primary and secondary cellulose synthesis, respectively (Table 2). The COBRA and CTL gene products affect primary and secondary cell wall biosynthesis, although their specific functions are unclear (19). Several other genes, such as glucanases, family 8 glycosyltransferases and arabinogalactan proteins, also appear to have homologs associated with primary and secondary cellulose production, respectively.

To identify genes associated with secondary cell wall biosynthesis in Barley we used *AtCESA4* from *Arabidopsis* as bait gene and used BLAST to identify targets in Barley (i.e. *Arabidopsis* versus Barley). Rosetta identified 14 probe sets in Barley that have similar sequences compared to *AtCESA4* in *Arabidopsis* (Supplementary Table S3). Similar to above, Rosetta recognized genes that are common between the co-expressed gene list for *AtCESA4* and the co-expressed gene lists for the 14 Barley probe sets. The comparison of the co-expression profiles revealed that three of the Barley probe sets, corresponding to Contig9658_at, Contig20165_at and Contig_15116_at, had the most similar co-expression profiles to *AtCESA4* in *Arabidopsis* (Supplementary Table S3). BLAST analysis revealed that these probe sets correspond to secondary cell wall *HvCESA4*, *HvCESA7* and *HvCESA5/7*, respectively. Similar to the analysis in *Arabidopsis*, Rosetta also identified putative *COBRA*-like 4 and *CTL2* orthologs associated with the secondary *HvCESA*s in Barley (Supplementary Table S4). Thus, Rosetta may rapidly identify homologs that are involved in similar biological processes within and across different organisms and may therefore be used to infer 'true' orthologs.

CONCLUDING REMARKS

Several tools use transcriptional coordination of genes to prioritize genes associated with a specific biological function. However, combining gene expression analyses with other data sources may give researchers additional information. GeneCAT combines sequence homology and co-expression and therefore provides a multidimensional platform for exploring gene co-expression and functional redundancies between homologs within and across different species such as *Arabidopsis* and Barley. Rapid advances in other large-scale approaches, such as protein–protein interactions and metabolomics, may in the near future be combined with the tools presented here to generate a more in depth view of cellular processes in higher plants. To facilitate an easily accessible exploratory platform for plant biologists we have linked web interfaces for several other genome tools through the GeneCAT FAQs page.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank Ann Loraine, Bjorn Usadel and Peter Ulvskov for their useful comments on the article. Financial support was provided by the Max-Planck Society to M.M. and S.P. Funding to pay the Open Access publication charges for this article was provided by Max Planck Society; Max Planck Institute of Molecular Plant Physiology.

Conflict of interest statement. None declared.

REFERENCES

1. Parkinson,H., Kapushesky,M., Shojatalab,M., Abeygunawardena,N., Coulson,R., Farne,A., Holloway,E., Kolesnykov,N., Lilja,P., Lukk,M. *et al*. (2007) ArrayExpress—a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res*., **35(Database issue)**, D747–D750.
2. Zimmermann,P., Hirsch-Hoffmann,M., Hennig,L. and Gruissem,W. (2004) GENEVESTIGATOR. Arabidopsis microarray database and analysis toolbox. *Plant Physiol*., **136**, 2621–2632.
3. 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(Web Server issue)**, W504–W509.
4. 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.
5. Steinhauser,D., Usadel,B., Luedemann,A., Thimm,O. and Kopka,J. (2004) CSB.DB: a comprehensive systems-biology database. *Bioinformatics*, **20**, 3647–3651.
6. 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(Database issue)**, 869.
7. Geisler-Lee,J., O'Toole,N., Ammar,R., Provart,N.J., Millar,A.H. and Geisler,M. (2004) A predicted interactome for Arabidopsis. *Plant Physiol*., **45**, 317–329.
8. Stuart,J.M., Segal,E., Koller,D. and Kim,S.K. (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science*, **302**, 249–255.
9. Bergmann,S., Ihmels,J. and Barkai,N. (2004) Similarities and differences in genome-wide expression data of six organisms. *PLoS Biol*., **2**, E9.
10. Brown,D.M., Zeef,L.A., Ellis,J., Goodacre,R. and Turner,S.R. (2005) Identification of novel genes in Arabidopsis involved in secondary cell wall formation using expression profiling and reverse genetics. *Plant Cell*, **17**, 2281–2295.
11. Persson,S., Wei,H., Milne,J., Page,G.P. and Somerville,C.R. (2005) Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets. *Proc. Natl Acad. Sci. USA*, **102**, 8633–8638.
12. Wei,H., Persson,S., Mehta,T., Srinivasasainagendra,V., Chen,L., Page,G.P., Somerville,C. and Loraine,A. (2006) Transcriptional coordination of the metabolic network in Arabidopsis. *Plant Physiol*., **142**, 762–774.
13. Hirai,M.Y., Sugiyama,K., Sawada,Y., Tohge,T., Obayashi,T., Suzuki,A., Araki,R., Sakurai,N., Suzuki,H., Aoki,K. *et al*. (2007) Omics-based identification of Arabidopsis Myb transcription factors regulating aliphatic glucosinolate biosynthesis. *Proc. Natl Acad. Sci. USA*, **104**, 6478–6483.
14. Fredslund,J. (2006) PHY.FI: fast and easy online creation and manipulation of phylogeny color figures. *BMC Bioinform*., **7**, 315.
15. Rhee,S., Beavis,W., Berardini,T., Chen,G., Dixon,D., Doyle,A., Garcia-Hernandez,M., Huala,E., Lander,G. and Montoya,M. (2003) The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to

Arabidopsis biology, research materials and community. *Nucleic Acids Res.*, **31**, 224–228.
16. Schmid,M., Davison,T.S., Henz,S.R., Pape,U.J., Demar,.M., Vingron,M., Scholkopf,B., Weigel,D. and Lohmann,J.U. (2005) A gene expression map of Arabidopsis thaliana development. *Nat. Genet.*, **37**, 501–506.
17. Shen,L., Gong,J., Caldo,R.A., Nettleton,D., Cook,D., Wise,R.P. and Dickerson,J.A. (2005) BarleyBase—an expression profiling database for plant genomics. *Nucleic Acids Res.*, **33(Database issue)**, D614–D618.
18. Druka,A., Muehlbauer,G., Druka,I., Caldo,R., Baumann,U., Rostoks,N., Schreiber,A., Wise,R., Close,T., Kleinhofs,A. *et al.* (2006) An atlas of gene expression from seed to seed through Barley development. *Funct. Integr. Genomics*, **6**, 202–211.
19. Somerville,C. (2006) Cellulose synthesis in higher plants. *Annu. Rev. Cell Dev. Biol.*, **22**, 53–78.
20. Arioli,T., Peng,L., Betzner,A.S., Burn,J., Wittke,W., Herth,W., Camilleri,C., Hofte,H., Plazinski,J., Birch,R. *et al.* (1998) Molecular analysis of cellulose biosynthesis in Arabidopsis. *Science*, **279.**
21. Fagard,M., Desnos,T., Desprez,T., Goubet,F., Refregier,G., Mouille,G., McCann,M., Rayon,C., Vernhettes,S. and Hofte,H. (2000) PROCUSTE1 encodes a cellulose synthase required for normal cell elongation specifically in roots and dark-grown hypocotyls of Arabidopsis. *Plant Cell*, **12**, 2409–2424.
22. Persson,S., Paredez,A., Carroll,A., Palsdottir,H., Doblin,M., Poindexter,P., Khitrov,N., Auer,M. and Somerville,C.R. (2007) Genetic evidence for three unique components in primary cell-wall cellulose synthase complexes in Arabidopsis. *Proc. Natl Acad. Sci. USA*, **104**, 15566–15571.
23. Turner,S.R. and Somerville,C.R. (1997) Collapsed xylem phenotype of Arabidopsis identifies mutants deficient in cellulose deposition in the secondary cell wall. *Plant Cell*, **9**, 689–701.
24. Burton,R.A., Shirley,N.J., King,B.J., Harvey,A.J. and Fincher,G.B. (2004) The CesA gene family of Barley. Quantitative analysis of transcripts reveals two groups of co-expressed genes. *Plant Physiol.*, **134**, 224–236.
25. Desprez,T., Juraniec,M., Crowell,E.F., Jouy,H., Pochylova,Z., Parcy,F., Hofte,H., Gonneau,M. and Vernhettes,S. (2007) Organization of cellulose synthase complexes involved in primary cell wall synthesis in Arabidopsis thaliana. *Proc. Natl Acad. Sci. USA*, **104**, 15572–15577.
26. Jen,C.H., Manfield,I.W., Michalopoulos,I., Pinney,J.W., Willats,W.G, Gilmartin,P.M. and Westhead,D.R. (2006) The Arabidopsis co-expression tool (ACT): a WWW-based tool and database for microarray-based gene expression analysis. *Plant J.*, **46**, 336–348.
27. Aoki,K., Ogata,Y. and Shibata,D. (2007) Approaches for extracting practical information from gene coexpression networks in plant biology. *Plant Cell Physiol.*, **48**, 381–390.
28. Franke,R. and Schreiber,L. (2007) Suberin—a biopolyester forming apoplastic plant interfaces. *Curr. Opin. Plant Biol.*, **10**, 252–259.
29. Muraoka,R., Okuda,K., Kobayashi,Y. and Shikanai,T. (2006) A eukaryotic factor required for accumulation of the chloroplast NAD(P)H dehydrogenase complex in Arabidopsis. *Plant Physiol.*, **142**, 1683–1689.
30. Cooke,R., Raynal,M., Laudie,M. and Delseny,M. (1997) Identification of members of gene families in Arabidopsis thaliana by contig construction from partial cDNA sequences: 106 genes encoding 50 cytoplasmic ribosomal proteins. *Plant J.*, **11**, 1127–1140.