Published online 1 May 2018 *Nucleic Acids Research, 2018, Vol. 46, Web Server issue* ***W133–W140***
doi: 10.1093/nar/gky336

CoNekT: an open-source framework for comparative genomic and transcriptomic network analyses

Sebastian Proost[1,*] **and Marek Mutwil**[1,2,*]

[1]Max-Planck Institute for Molecular Plant Physiology, Am Muehlenberg 1, 14476 Potsdam, Germany and [2]School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore 637551, Singapore

Received January 27, 2018; Revised April 10, 2018; Editorial Decision April 17, 2018; Accepted April 18, 2018

ABSTRACT

The recent accumulation of gene expression data in the form of RNA sequencing creates unprecedented opportunities to study gene regulation and function. Furthermore, comparative analysis of the expression data from multiple species can elucidate which functional gene modules are conserved across species, allowing the study of the evolution of these modules. However, performing such comparative analyses on raw data is not feasible for many biologists. Here, we present CoNekT (Co-expression Network Toolkit), an open source web server, that contains user-friendly tools and interactive visualizations for comparative analyses of gene expression data and co-expression networks. These tools allow analysis and cross-species comparison of (i) gene expression profiles; (ii) co-expression networks; (iii) co-expressed clusters involved in specific biological processes; (iv) tissue-specific gene expression; and (v) expression profiles of gene families. To demonstrate these features, we constructed CoNekT-Plants for green alga, seed plants and flowering plants (*Picea abies, Chlamydomonas reinhardtii*, *Vitis vinifera*, *Arabidopsis thaliana*, *Oryza sativa*, *Zea mays* and *Solanum lycopersicum*) and thus provide a web-tool with the broadest available collection of plant phyla. CoNekT-Plants is freely available from http://conekt.plant.tools, while the CoNekT source code and documentation can be found at https://github.molgen.mpg.de/proost/CoNekT/.

INTRODUCTION

With the continuous improvement of sequencing technologies, the cost to generate a genome sequence has decreased nearly 8000-fold during the last decade (https://www.genome.gov/sequencingcostsdata/). Due to these improvements, RNA sequencing (RNA-Seq) became the method of choice to study transcript abundance. RNA-Seq allows detection of differentially expressed genes (1), assembly of coding sequences *de novo* in the absence of a reference genome (2), construction and analysis of expression atlases (3–5) and co-expression networks which can guide gene-function predictions (6,7). Combined with comparative genomics, these approaches can also be used to study transcriptional differences to understand phenotypic variation within and between species (8–10). Despite the advantages of RNA sequencing, it is important to note that the platform suffers from several biases, such as underrepresentation of reads stemming from transcripts that are either GC-rich, GC-poor, show low mappability (uniqueness) of a particular sequence compared to the rest of the genome (e.g. for recently duplicated paralogs), or are lowly expressed or very short (11,12). Furthermore, the disagreement between results obtained from different tools or parameter settings indicates that the platform and methods are continuously evolving, and the results should still be interpreted with care, especially when comparing gene expression across species (12).

While various tools exist to browse expression profiles and co-expression networks (8,13–16), they are often limited to few species and closed-source, which prevents users to create custom versions including their own data. To this end, we developed CoNekT (Co-expression Network Toolkit, https://github.molgen.mpg.de/proost/CoNekT). As CoNekT is open-source and available under the MIT license, researchers can create new online or in-house instances for their own data and expand CoNekT with features relevant to their research. To demonstrate the usefulness of the platform, we present CoNekT-Plants (http://conekt.plant.tools), which allows comparative analyses of six land plants and alga.

MATERIALS AND METHODS

Implementation and interface

CoNekT consists of two components: (i) a python-flask backend which processes requests, fetches data from the database, provides search functionality and serves web

[*]To whom correspondence should be addressed. Tel: +65 6904 7503; Email: mutwil@ntu.edu.sg
Correspondence may also be addressed to Sebastian Proost. Tel: +49 0 331 5678 155; Email: proost@mpimp-golm.mpg.de

pages and (ii) a front-end which includes various interactive visualizations based on charts.js, cytoscape.js (17), and phyD3.js (18). The Bootstrap CSS library is used to style all pages and is paired with jQuery.js and qtip2.js to add various dynamic elements such as tooltips and popups. Font-awesome is included for different glyphs and icons. A full overview of all included libraries in both the back- and front-end is included in the online documentation (https://github.molgen.mpg.de/proost/CoNekT/).

CoNekT contains pages for species, genes, gene families, co-expression clusters and neighborhoods, and others. These pages, in turn, contain graphs, tables and links relevant to the page. For example, gene pages indicate the gene's (i) description, (ii) gene family, (iii) phylogenetic tree, (iv) cDNA and protein sequences, (v) expression profile, (vi) co-expression neighborhood and cluster, (vii) similar neighborhoods in other species and (viii) Gene Ontology information (inferred by experimental evidence, InterProScan and co-expression network neighborhood) (Figure 1). A detailed description of all available features and instructions of how to deploy your own CoNekT web-server can be found at: https://github.molgen.mpg.de/proost/CoNekT/. The utilized packages and dependencies are listed at: https://github.molgen.mpg.de/proost/CoNekT/blob/master/requirements.txt/.

Data acquisition for CoNekT-Plants

To demonstrate the web-server, we introduce CoNekT-Plants, which contains data from seven species (Table 1), including green alga *Chlamydomonas reinhardtii*, gymnosperm *Picea abies*, two monocots (*Oryza sativa*, *Zea mays*) and three dicotyledonous plants (*Vitis vinifera*, *Arabidopsis thaliana*, and *Solanum lycopersicum*). For each species, publically available RNA-Seq data was obtained through the Sequence Read Archive's 'Run Selector' (https://www.ncbi.nlm.nih.gov/sra/) (19). These samples were downloaded, converted to fastq files (using SRA Tools, https://www.ncbi.nlm.nih.gov/books/NBK158900/) and processed using LSTrAP (6), which maps reads to the genome using TopHat (20) and determines transcript abundance for each gene using HTSeq-count (21). LSTrAP used the output from HTSeq-count to calculate Transcripts Per Kilobase Million (TPM) values, which normalize for read count and gene length (12). The expression values are represented as an expression matrix, where the genes are present in rows and the samples in columns. The mapping statistics included in LSTrAP were used to detect and discard samples that showed either (i) low mapping to the genome (<65%), (ii) low mapping to coding sequences (<40%) or (iii) too few useful reads (<8M reads mapping to the genome). Additionally, using LSTrAP's heatmap tool, the output was screened for outliers, which were removed from the final dataset. The remaining samples were used to construct expression matrices and co-expression networks. For *Arabidopsis thaliana*, experimentally determined functional annotation (Gene Ontology terms) was obtained from www.arabidopsis.org. Additionally, for all species, InterProScan v5.18 (22) was used to detect protein domains and obtain predicted functional annotation. To obtain orthologs, OrthoFinder v1.1.8 (23) was used to group genes into orthogroups and construct phylogenetic trees, using Diamond to determine sequence similarities with settings at default values (24). Sequence similarities reported by Diamond were clustered using MCL to group homologous genes into gene families (25). Note that all above mentioned steps can be performed in LSTrAP, and the output can be directly used in CoNekT.

Co-expression networks in CoNekT are based on Highest Reciprocal Rank (HRR) metric score of 100 or better (8), which is related to a robust rank-based metric used to identify co-expressed genes (26). Groups of densely connected genes, called co-expression clusters, were detected using the Heuristic Cluster Chiseling Algorithm (27). Using CoNekT's graphical admin interface, the expression and genomic data were added to the platform (see instructions on https://github.molgen.mpg.de/proost/CoNekT/). Through the same interface, multiple analyses were started, such as (i) the Heuristic Cluster Chiseling Algorithm (HCCA), to find clusters of co-expressed genes in the networks (27); (ii) Gene Ontology term over-representation to elucidate the functional annotation of co-expression clusters and co-expression network neighborhoods (reported as enrichment fold-changes and *P*-values); (iii) identification of similar co-expression network clusters and neighborhoods within and across species, by employing Expression Context Conservation (ECC) value. The value is a Jaccard Index of gene families found in the two compared neighborhoods or clusters (9).

RESULTS AND DISCUSSION

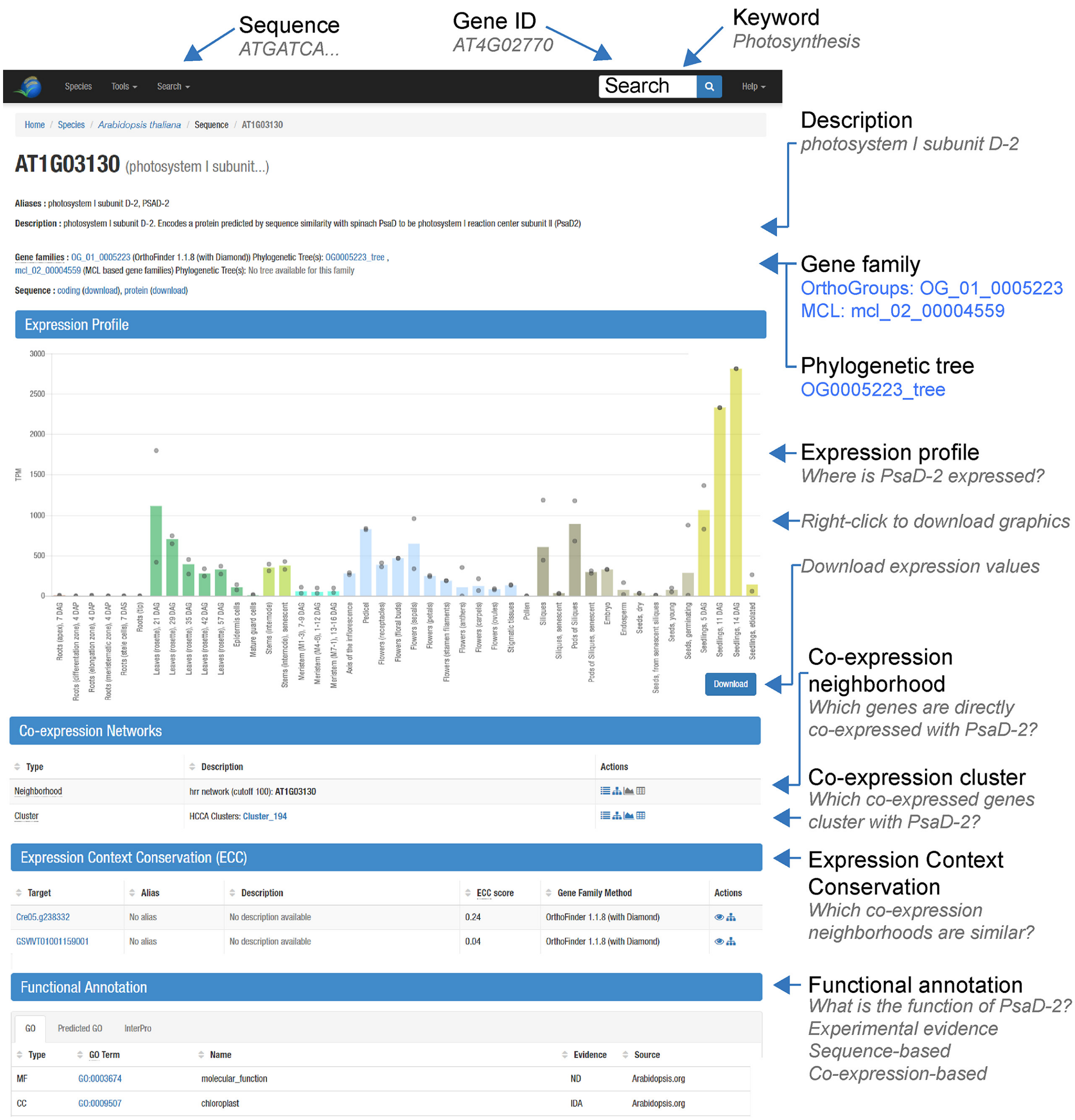
Querying CoNekT

CoNekT features three modes to search for relevant content. First, the keyword search, available from the landing page and upper right corner (Figure 1), accepts gene IDs (e.g. *At4g32410*), Gene Ontology term IDs (e.g. GO:0008810), keywords (e.g. 'cellulose') and InterPro domains (e.g. cellulose_synt), and returns genes, GO terms and InterPro domains that match the query. Second, the advanced search function available in Search/Search(advanced) menu on the top of the page can be used to retrieve genes with a specific combination of functional annotation, GO IDs and/or InterPro domains. Third, relevant genes can be retrieved by sequence similarity with BLAST (Search/BLAST).

Gene expression profiles

The pattern of gene expression can reveal where and when a specific gene is active and thus can suggest the gene's function. For example, uncharacterized genes with specific expression in roots might be essential for root development. To visualize gene expression levels, Transcripts Per Kilobase Million (TPM) values were grouped by tissue, condition and/or developmental stage for each gene (Figure 1). These profiles can be exported as png/jpg graphics or as a table.

CoNekT allows comparisons of gene expression across species, where average expression in predefined organs is shown. CoNekT-Plants was configured to show gene



Sequence
ATGATCA...

Gene ID
AT4G02770

Keyword
Photosynthesis

AT1G03130 (photosystem I subunit...)

Aliases : photosystem I subunit D-2, PSAD-2

Description : photosystem I subunit D-2. Encodes a protein predicted by sequence similarity with spinach PsaD to be photosystem I reaction center subunit II (PsaD2)

Gene families : OG_01_0005223 (OrthoFinder 1.1.8 (with Diamond)) Phylogenetic Tree(s): OG0005223_tree , mcl_02_00004559 (MCL based gene families) Phylogenetic Tree(s): No tree available for this family

Sequence : coding (download), protein (download)

Expression Profile

Co-expression Networks

Type	Description	Actions
Neighborhood	hrr network (cutoff 100): **AT1G03130**	
Cluster	HCCA Clusters: Cluster_194	

Expression Context Conservation (ECC)

Target	Alias	Description	ECC score	Gene Family Method	Actions
Cre05.g238332	No alias	No description available	0.24	OrthoFinder 1.1.8 (with Diamond)	
GSVIVT01001159001	No alias	No description available	0.04	OrthoFinder 1.1.8 (with Diamond)	

Functional Annotation

Type	GO Term	Name	Evidence	Source
MF	GO:0003674	molecular_function	ND	Arabidopsis.org
CC	GO:0009507	chloroplast	IDA	Arabidopsis.org

Description
photosystem I subunit D-2

Gene family
OrthoGroups: OG_01_0005223
MCL: mcl_02_00004559

Phylogenetic tree
OG0005223_tree

Expression profile
Where is PsaD-2 expressed?

Right-click to download graphics

Download expression values

Co-expression neighborhood
Which genes are directly co-expressed with PsaD-2?

Co-expression cluster
Which co-expressed genes cluster with PsaD-2?

Expression Context Conservation
Which co-expression neighborhoods are similar?

Functional annotation
What is the function of PsaD-2?
Experimental evidence
Sequence-based
Co-expression-based

Figure 1. Gene page contents exemplified with Arabidopsis *PsaD-2*. The gene page provides information (as tables) and links (in blue) specific to the gene. The links allow quick access to the co-expression neighborhood, cluster, gene family and phylogenetic tree of *PsaD-2*.

Table 1. Species included in CoNekT-Plants

Organism	Genome source	Class	Number of samples (retained)	Number of nodes	Number of HCCA clusters
Arabidopsis thaliana	TAIR10	Eudicot	913	27 172	479
Chlamydomonas reinhardtii	Phytozome v5.5	Chlorophyceae	605	17 741	273
Oryza sativa	Phytozome v7.0	Monocot	750	39 717	662
Picea abies	ConGenIE v1.0	Pinopsida	148	66 632	1814
Solanum lycopersicum	ITAG 3.10	Eudicot	706	34 879	612
Vitis vinifera	Genescope 12x	Eudicot	612	26 346	499
Zea Mays	Ensembl Plants AGPv4	Monocot	574	39 000	728

The table indicates the genome source, phylogenetic class, number of RNA-seq samples that passed the LSTrAP quality control, number of nodes (genes) and the number of co-expression clusters identified by HCCA algorithm.

expression in roots/rhizoids, leaves, stems, female reproduction (containing ovaries, pistrils), male reproduction (containing pollen, anthers) and flower/seeds/spores. This cross-species comparative expression analysis is available from Tools/Create heatmap/Comparative window or by clicking on the 'View comparative expression as a heatmap:' link on a gene family page. We illustrate such a heatmap with photosystem I subunit family D (PsaD, http://conekt.plant.tools/family/view/5224), which is involved in photosynthesis (28). The heatmap can be accessed by clicking on the 'row-normalized' link in 'View comparative expression as heatmap:' line. As expected of photosynthesis-related genes, they show the highest expression in leaves and virtually no expression in roots or male reproduction, which contains non-photosynthesizing pollen and anthers (Figure 2A). While this example illustrates that the PsaD family genes have conserved expression, the heatmap could be used to rapidly identify genes with changed expression.

Expression specificity

Since gene expression patterns can reveal gene function, extracting genes expressed specifically in a given organ, tissue or condition can be used to predict gene function. To detect expression specificity, CoNekT uses specificity measure (SPM, ranges between zero and one, where one indicates the gene is exclusively expressed in the tissue) (29), Tau (high values indicate that a profile is specific in a tissue), and entropy (indicates how much a profile fluctuates across all tissues, where genes with very specific or very stable expression have low entropy) (30). The application of one or more of these metrics allows users to search for expression profiles specific for one tissue or condition (30). To illustrate the tool, we selected 'Tools/Find specific profiles,' chose 'Tissue specificity' as method and selected Arabidopsis 'Meristems' as condition. The output is returned as a table, where rows are genes, columns contain descriptions, and the SPM, entropy and Tau values. For meristems in Arabidopsis, the tool returned a table with 146 genes with SPM>0.85 (default value) and showed known factors influencing meristem and flower development, such as *LEAFY*, *CLAVATA3*, *AINTEGUMENTA*, *DORNROSCHEN* and others (31–34) (See Supplementary Table S1 for full list). Thus, the presence of these known factors indicates that the tool can retrieve relevant genes.

Comparative expression specificity

When orthologous genes are specifically expressed in similar tissues/organs across different species, it further strengthens the evidence of their importance for the function of that tissue, as conserved expression profiles are unlikely to appear by chance (8,9,35). Furthermore, orthologs that show conserved expression are more likely to be functionally equivalent, which can be used to resolve often unclear phylogenetic relationships caused by gene duplications (8,9). To this end, the 'Tools/Compare specificities' tool compares two lists of specifically expressed genes within or across species. As an example, we compared orthologs specifically expressed in Arabidopsis and rice pollen (Selected method 'Tissue specificity', default SPM, select 'pollen' for both species), which revealed 103 orthogroups that were specifically expressed in pollen of the two plants. The results include a set of well-known genes related to pollen development and fertilization, such as *CSLD1*, *CSLD4*, *COBL10*, *LIP1*, *LIP2*, *TIP5;1*, *FH3*, *AKT6* (36–41), but also a host of other genes potentially important for these processes in both species (See Supplementary Table S2 for full list). Thus, similarly to the tool above, this feature can reveal genes relevant for the development of a tissue of interest, with the additional advantage of highlighting conserved and biologically relevant genes.

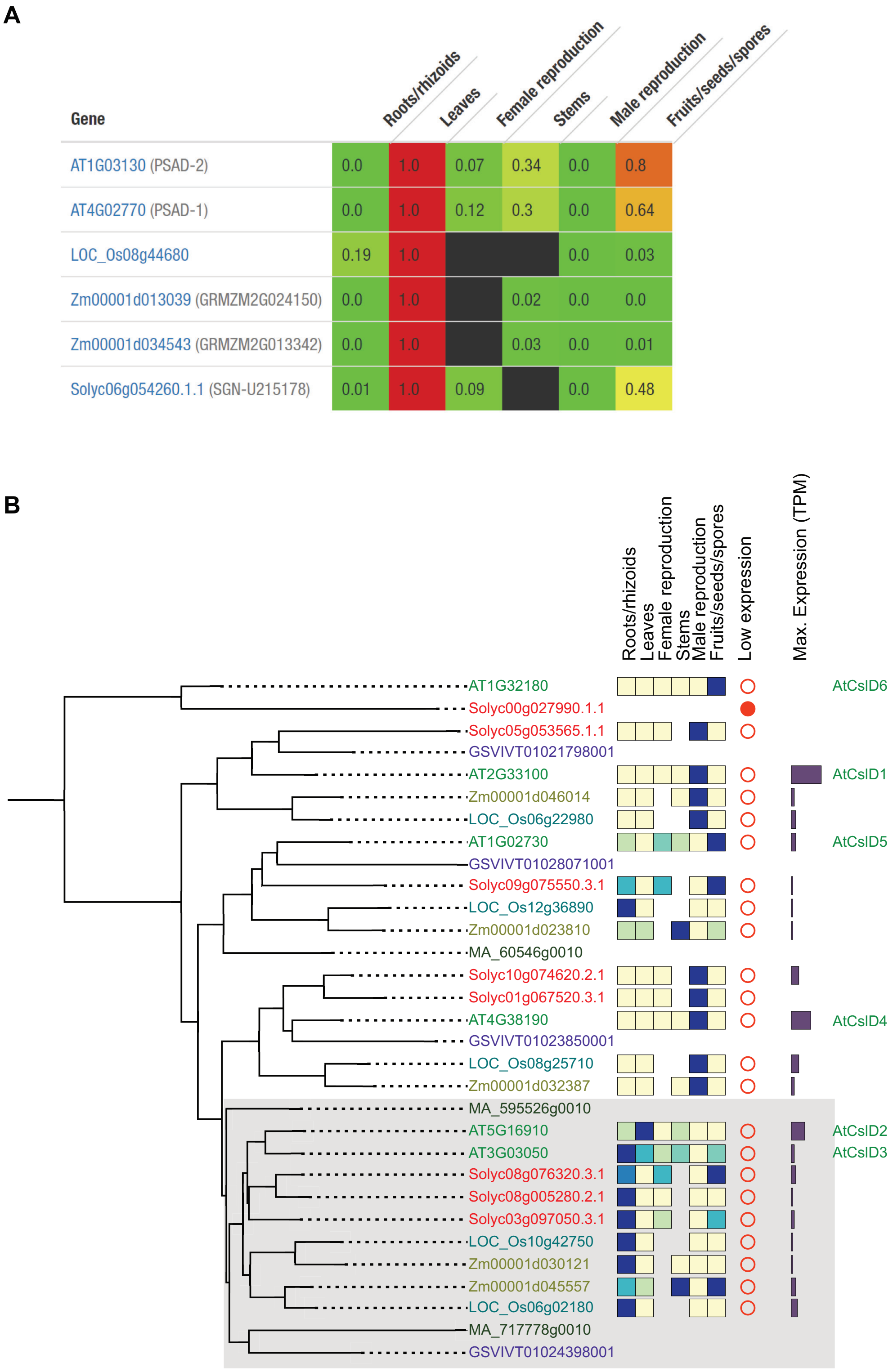
Phylogenetic and expression analysis

Phylogenetic trees provide the most detailed view of speciations and duplications, and their timing, between homologous genes. However, phylogenetic trees do not reveal sub- or neo-functionalization of genes that might be apparent when investigating expression data (10,42,43). To remedy this, CoNekT combines interactive phylogenetic trees with the comparative expression profile heatmaps, which allows elucidating potential sub- or neo-functionalization events.

To demonstrate the usefulness of the tool, we show a tree of Cellulose Synthase-like D (CslD) gene family, involved in tip growth in plants (Figure 3) (36). To obtain the tree, we first navigated to the page of *AtCslD1* (*At2g33100,* http://conekt.plant.tools/sequence/view/45080), which is involved in pollen tube growth (36), and clicked on the link to the family's phylogenetic tree (OG0000579_tree, http://conekt.plant.tools/tree/view/12121). The tree revealed that most of the CslD family genes are either expressed in roots or male reproductive tissues (Figure 2B). For example, *AtCslD3* (*At3g03050*) shows high expression in roots, which is in line with the gene's involvement in root hair growth (36). Furthermore, genes from other plants that are found in the *AtCslD3* clade (indicated by the gray box) show root-specific expression, suggesting that these genes are also involved in root hair growth. Conversely, most genes in the *AtCslD1* (*At2g33100*) and *AtCslD4* (*At4g38190*) clade show predominantly male reproduction-specific expression, which is in line with their function in pollen tube growth (36). *AtCslD5* (*At1g02730*) was shown to be involved in cell plate formation (44), suggesting that the clade the gene is found in has gained new function (Figure 2B). Finally, since the oldest lineage in the root hair and pollen tube clades is spruce (genes starting with *MA_*), we can postulate that the gene duplication that created the sub-specialized pollen tube / root hair genes took place in the common ancestor of seed plants. To conclude, the combination of phylogenetic trees and expression can be used to identify functional innovations found in gene families.

Comparative co-expression network analyses

Genes with similar expression profiles (co-expressed genes) are often functionally related, and consequently, co-expression analysis is a robust method for gene function prediction (8,14,45,46). Co-expressed genes can be represented as a network, where nodes represent genes and edges are drawn between co-expressed nodes. Co-expression patterns can be conserved across species (even over large phylogenetic distance) (47–49), and this property can be used



Gene	Roots/rhizoids	Leaves	Female reproduction	Stems	Male reproduction	Fruits/seeds/spores
AT1G03130 (PSAD-2)	0.0	1.0	0.07	0.34	0.0	0.8
AT4G02770 (PSAD-1)	0.0	1.0	0.12	0.3	0.0	0.64
LOC_Os08g44680	0.19	1.0			0.0	0.03
Zm00001d013039 (GRMZM2G024150)	0.0	1.0		0.02	0.0	0.0
Zm00001d034543 (GRMZM2G013342)	0.0	1.0		0.03	0.0	0.01
Solyc06g054260.1.1 (SGN-U215178)	0.01	1.0	0.09		0.0	0.48

Figure 2. Comparative expression analysis. (**A**) Expression profile of Arabidopsis (genes starting with AT), rice (LOC), maize (Zm) and tomato (Solyc) PsaD gene family in roots, leaves, female reproduction (e.g. ovaries, stigma), stems, male reproduction (e.g. pollen, anthers) and fruits. The expression values of each gene were normalized by diving by their maximum, and range from 1 (red, maximum expression) to 0 (green, no expression). Missing expression data is shown with a black box (e.g. female reproduction and stems for rice). (**B**) Phylogenetic tree of the Cellulose Synthase-like D (CSLD) gene family. The heatmap shows the expression level in different tissues, full red dots show genes with low-expression and the bar on the right indicates the maximum expression level (TPM). The color of a gene identifier indicates the species. The added gray box contains genes that have shifted towards being expressed in roots. Note that OrthoFinder tree nodes do not contain bootstrap values, and should be interpreted with care. Missing data is indicated by absent box; for example, spruce has insufficient expression data to provide an informative expression.

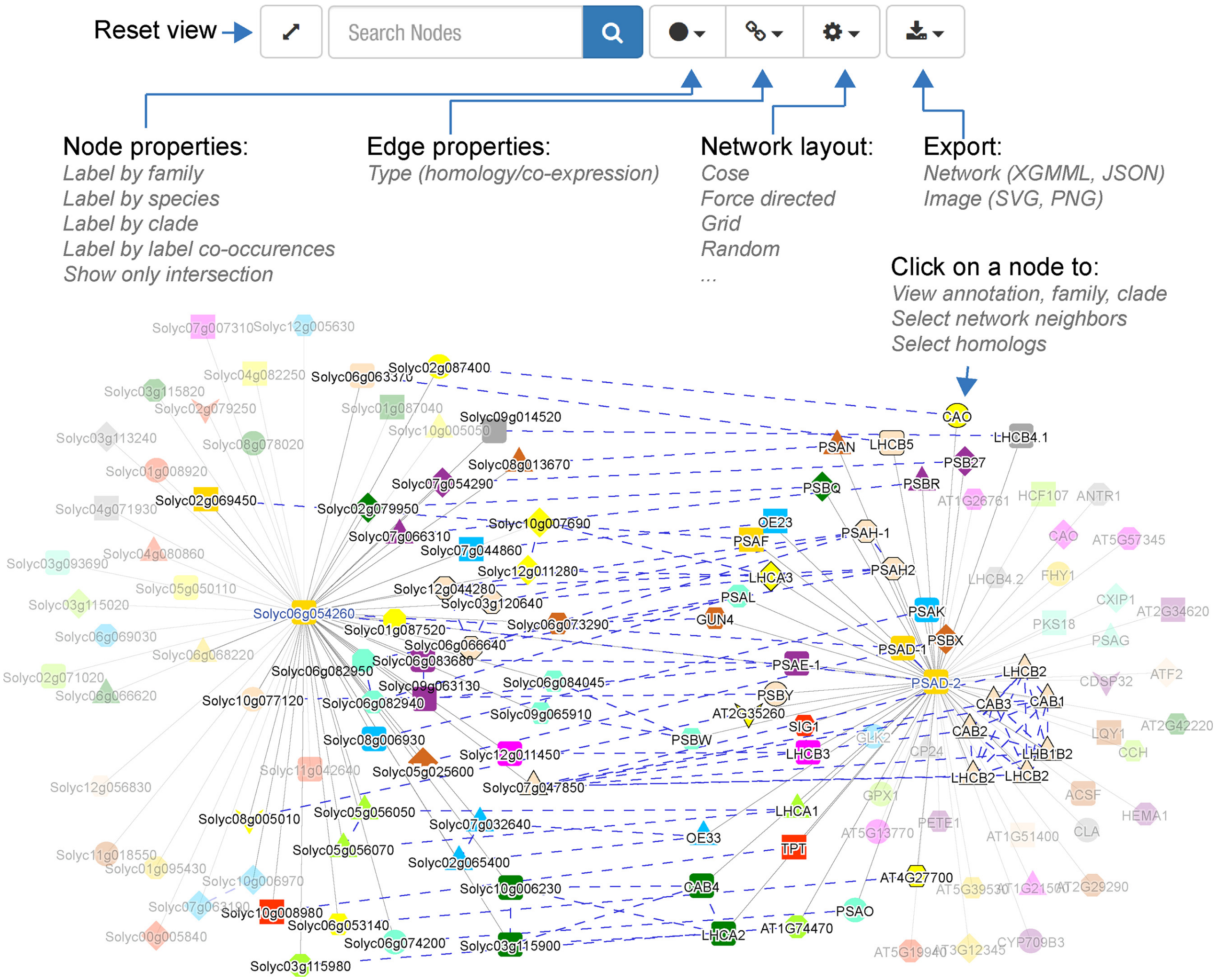


Figure 3. Comparative network analysis of *AtPsaD-2* and Solyc06g054260. *AtPsaD-2* and tomato ortholog *Solyc06g054260* are shown together with their co-expression neighborhood (co-expressed genes are connected using solid gray lines). Nodes with the same shape and color are members of the same OrthoGroup. Orthologs found in both neighborhoods are connected with dashed blue lines. The indicated menus are used to change the node and edge labels, network layout and export the networks as images and Cytoscape-compatible data. For clarity, non-conserved nodes were made semi-transparent, and the two query genes are connected by a solid edge.

to transfer functional knowledge from one species to another (8,50–52). Recently, it was demonstrated that subsections of the co-expression network comprised of functionally related genes, also known as gene modules, were duplicated within one species (43). To detect these conserved and duplicated modules, CoNekT uses the Expression Context Conservation (ECC) value (9) to detect which network regions contain more gene families in common than expected by chance (53).

Users can visualize a gene and its direct co-expression partners (neighborhood) or all genes within a co-expression cluster (Figure 3). The interactive networks provide an intuitive interface that allows nodes to be colored based on various parameters (e.g., gene family, phylogenetic clade and others) and can be searched by gene name, alias or annotation. Edges can be colored and shown/hidden by edge weight. Furthermore, different graph layouts are supported, and the networks can be exported as vector (SVG) or pixel-based (PNG) graphics.

To exemplify a comparative co-expression analysis, we used Arabidopsis *PsaD-2*, which is part of photosystem I complex. On the *PsaD-2* gene page (*At1g03130*, http://conekt.plant.tools/sequence/view/35615), a tomato ortholog (*Solyc06g054260.1.1*) with an ECC score of 0.33 was found. By clicking on ‘View ECC pair as graph’ glyph, CoNekT detected conserved photosynthetic components of photosystem I (gene IDs starting with PSA), II (gene IDs starting with PSB), Light Harvesting Complex (gene IDs starting with LHC) in both genes’ neighborhoods (orthologs are connected by dashed lines, Figure 3).

Searching for functionally enriched clusters across species

To detect co-expression clusters containing genes associated with specific functions, CoNekT precalculates GO enrichment for all clusters, which can be searched using the

'Tools\Find enriched clusters' feature. For instance, looking for clusters enriched for GO term 'reproduction' in all seven species found in CoNekT-Plants yielded 18 clusters significantly enriched for this term. By clicking on 'Compare profiles in this cluster' icon in 'Action' menu, users can quickly screen which cluster is acting in the tissues of interest. Such search revealed that *Arabidopsis thaliana* cluster 17 and maize cluster 6 were significantly enriched for genes involved in reproduction (adjusted P-value < 0.01) and expressed in pollen and anthers (Supplementary Table S3), indicating that these clusters are involved in a male reproductive process. Such analysis is therefore a good starting point to identify genes relevant for a biological process of interest.

CONCLUSIONS

CoNekT is a modern web-platform that provides an intuitive interface for combining large-scale expression data with functional and genomic information. This allows users to extract tissue-specific genes, to compare tissue-specific transcriptomes between species and to leverage co-expression networks to predict gene function. These networks can be compared in a broad phylogenetic context. As CoNekT is open-source, researchers can create a version which includes their own RNA-Seq data and disseminate this online. Expert users can dive into the code and implement advanced features designed to answer their specific research questions, without having to re-implement core components such as gene families, expression profiles and co-expression network browsers.

DATA AVAILABILITY

The co-expression networks are available for download at http://conekt.plant.tools/species/. Source code and documentation can be found on https://github.molgen.mpg.de/proost/CoNekT/.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

CoNekT-Plants is hosted at the Max Planck Institute of Molecular Plant Physiology MPIMP in Potsdam-Golm, Germany and we would like to thank Andreas Donath for tech support. Furthermore, we would like to thank EvoRepro members for testing and providing valuable feedback, especially Jörg Becker, Ann-Catherin Lindner, David Twell, and Mark Johnson, and Camilla Ferrari for proofreading the manuscript. Finally, we would like to express our gratitude to Maximilian Funk for help with licenses.

Author Contributions: CoNekT was designed and implemented by S.P. who also prepared the data and build CoNekT-Plants with input from M.M. Both S.P. and M.M. wrote the manuscript.

FUNDING

ERA-CAPS grant EVOREPRO for funding.

Conflict of interest statement. None declared.

REFERENCES

1. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
2. Haas,B.J., Papanicolaou,A., Yassour,M., Grabherr,M., Blood,P.D., Bowden,J., Couger,M.B., Eccles,D., Li,B., Lieber,M. *et al.* (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.*, **8**, 1494–1512.
3. Fasoli,M., Dal Santo,S., Zenoni,S., Tornielli,G.B., Farina,L., Zamboni,a., Porceddu,a., Venturini,L., Bicego,M., Murino,V. *et al.* (2012) The grapevine expression atlas reveals a deep transcriptome shift driving the entire plant into a maturation program. *Plant Cell*, **24**, 3489–3505.
4. Sibout,R., Proost,S., Hansen,B.O., Vaid,N., Giorgi,F.M., Ho-Yue-Kuang,S., Legée,F., Cézart,L., Bouchabké-Coussa,O., Soulhat,C. *et al.* (2017) Expression atlas and comparative coexpression network analyses reveal important genes involved in the formation of lignified cell wall in Brachypodium distachyon. *New Phytol.*, **215**, 1009–1025.
5. Schmid,M., Davison,T.S., Henz,S.R., Pape,U.J., Demar,M., Vingron,M., Schölkopf,B., Weigel,D. and Lohmann,J.U. (2005) A gene expression map of Arabidopsis thaliana development. *Nat. Genet.*, **37**, 501–506.
6. Proost,S., Krawczyk,A. and Mutwil,M. (2017) LSTrAP: Efficiently combining RNA sequencing data into co-expression networks. *BMC Bioinformatics*, **18**, 444.
7. Langfelder,P. and Horvath,S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, **9**, 559.
8. Mutwil,M., Klie,S., Tohge,T., Giorgi,F.M., Wilkins,O., Campbell,M.M., Fernie,A.R., Usadel,B., Nikoloski,Z. and Persson,S. (2011) PlaNet: combined sequence and expression comparisons across plant networks derived from seven species. *Plant Cell*, **23**, 895–910.
9. Movahedi,S., Van de Peer,Y. and Vandepoele,K. (2011) Comparative network analysis reveals that tissue specificity and gene function are important factors influencing the mode of expression evolution in Arabidopsis and rice. *Plant Physiol.*, **156**, 1316–1330.
10. Ruprecht,C., Vaid,N., Proost,S., Persson,S. and Mutwil,M. (2017) Beyond genomics: studying evolution with gene coexpression networks. *Trends Plant Sci.*, **22**, 298–307.
11. Risso,D., Schwartz,K., Sherlock,G. and Dudoit,S. (2011) GC-Content normalization for RNA-Seq data. *BMC Bioinformatics*, **12**, 480.
12. Conesa,A., Madrigal,P., Tarazona,S., Gomez-Cabrero,D., Cervera,A., McPherson,A., Szcześniak,M.W., Gaffney,D.J., Elo,L.L., Zhang,X. *et al.* (2016) A survey of best practices for RNA-seq data analysis. *Genome Biol.*, **17**, 16.
13. Proost,S. and Mutwil,M. (2017) PlaNet: Comparative Co-Expression network analyses for plants. In: van Dijk,ADJ (ed). *Plant Genomics Databases: Methods and Protocols*. Springer, NY, pp. 213–227.
14. Aoki,Y., Okamura,Y., Tadaka,S., Kinoshita,K. and Obayashi,T. (2015) ATTED-II in 2016: A plant coexpression database towards Lineage-Specific coexpression. *Plant Cell Physiol.*, **2**, doi:10.1093/pcp/pcv165.
15. Sundell,D., Mannapperuma,C., Netotea,S., Delhomme,N., Lin,Y.C., Sjödin,A., Van de Peer,Y., Jansson,S., Hvidsten,T.R. and Street,N.R. (2015) The plant genome integrative explorer Resource: PlantGenIE.org. *New Phytol.*, **208**, 1149–1156.
16. Netotea,S., Sundell,D., Street,N.R. and Hvidsten,T.R. (2014) ComPlEx: conservation and divergence of co-expression networks in A. thaliana, Populus and O. sativa. *BMC Genomics*, **15**, 106.
17. Franz,M., Lopes,C.T., Huck,G., Dong,Y., Sumer,O. and Bader,G.D. (2015) Cytoscape.js: A graph theory library for visualisation and analysis. *Bioinformatics*, **32**, 309–311.
18. Kreft,L., Botzki,A., Coppens,F., Vandepoele,K. and Van Bel,M. (2017) PhyD3: A phylogenetic tree viewer with extended phyloXML support for functional genomics data visualization. *Bioinformatics*, **33**, 2946–2947.
19. Leinonen,R., Sugawara,H. and Shumway,M. (2011) The sequence read archive. *Nucleic Acids Res.*, **39**, D19–D21.

20. Trapnell,C., Pachter,L. and Salzberg,S.L. (2009) TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.
21. Anders,S., Pyl,P.T. and Huber,W. (2015) HTSeq-A Python framework to work with high-throughput sequencing data. *Bioinformatics*, **31**, 166–169.
22. Quevillon,E., Silventoinen,V., Pillai,S., Harte,N., Mulder,N., Apweiler,R. and Lopez,R. (2005) InterProScan: Protein domains identifier. *Nucleic Acids Res.*, **33**, W116–W120.
23. Emms,D.M. and Kelly,S. (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.*, **16**, 157.
24. Buchfink,B., Xie,C. and Huson,D.H. (2015) Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*, **12**, 59–60.
25. van Dongen,S. (2000) Graph clustering by flow simulation. *Graph Stimul. by flow Clust*. **PhD thesis**, University of Utrecht.
26. 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.
27. Mutwil,M., Usadel,B., Schütte,M., Loraine,A., Ebenhöh,O. and Persson,S. (2010) Assembly of an interactive correlation network for the Arabidopsis genome using a novel heuristic clustering algorithm. *Plant Physiol.*, **152**, 29–43.
28. Ihnatowicz,A., Pesaresi,P., Varotto,C., Richly,E., Schneider,A., Jahns,P., Salamini,F. and Leister,D. (2004) Mutants for photosystem I subunit D of Arabidopsis thaliana: effects on photosynthesis, photosystem I stability and expression of nuclear genes for chloroplast functions. *Plant J.*, **37**, 839–852.
29. Xiao,S.J., Zhang,C., Zou,Q. and Ji,Z.L. (2010) TiSGeD: a database for tissue-specific genes. *Bioinformatics*, **26**, 1273–1275.
30. Kryuchkova-Mostacci,N. and Robinson-Rechavi,M. (2017) A benchmark of gene expression tissue-specificity metrics. *Brief. Bioinform.*, **18**, 205–214.
31. Chandler,J.W. and Werr,W. (2017) DORNRÖSCHEN, DORNRÖSCHEN-LIKE, and PUCHI redundantly control floral meristem identity and organ initiation in Arabidopsis. *J. Exp. Bot.*, **68**, 3457–3472.
32. Manchado-Rojo,M., Weiss,J. and Egea-Cortines,M. (2014) Validation of Aintegumenta as a gene to modify floral size in ornamental plants. *Plant Biotechnol. J.*, **12**, 1053–1065.
33. Parcy,F., Bomblies,K. and Weigel,D. (2002) Interaction of LEAFY, AGAMOUS and TERMINAL FLOWER1 in maintaining floral meristem identity in Arabidopsis. *Development*, **129**, 2519–2527.
34. Monniaux,M., McKim,S.M., Cartolano,M., Thévenon,E., Parcy,F., Tsiantis,M. and Hay,A. (2017) Conservation vs divergence in LEAFY and APETALA1 functions between Arabidopsis thaliana and Cardamine hirsuta. *New Phytol.*, **216**, 549–561.
35. Movahedi,S., Van Bel,M., Heyndrickx,K.S. and Vandepoele,K. (2012) Comparative co-expression analysis in plant biology. *Plant, Cell Environ.*, **35**, 1787–1798.
36. Bernal,A.J., Yoo,C.-M., Mutwil,M., Jensen,J.K., Hou,G., Blaukopf,C., Sørensen,I., Blancaflor,E.B., Scheller,H.V. and Willats,W.G.T. (2008) Functional analysis of the cellulose synthase-like genes CSLD1, CSLD2, and CSLD4 in tip-growing arabidopsis cells. *Plant Physiol.*, **148**, 1238–1253.
37. Li,S., Ge,F.R., Xu,M., Zhao,X.Y., Huang,G.Q., Zhou,L.Z., Wang,J.G., Kombrink,A., McCormick,S., Zhang,X.S. *et al.* (2013) Arabidopsis COBRA-LIKE 10, a GPI-anchored protein, mediates directional growth of pollen tubes. *Plant J.*, **74**, 486–497.
38. Liu,J., Zhong,S., Guo,X., Hao,L., Wei,X., Huang,Q., Hou,Y., Shi,J., Wang,C., Gu,H. *et al.* (2013) Membrane-bound RLCKs LIP1 and LIP2 are essential male factors controlling male-female attraction in Arabidopsis. *Curr. Biol.*, **23**, 993–998.
39. Soto,G., Fox,R., Ayub,N., Alleva,K., Guaimas,F., Erijman,E.J., Mazzella,A., Amodeo,G. and Muschietti,J. (2010) TIP5;1 is an aquaporin specifically targeted to pollen mitochondria and is probably involved in nitrogen remobilization in Arabidopsis thaliana. *Plant J.*, **64**, 1038–1047.
40. Ye,J., Zheng,Y., Yan,A., Chen,N., Wang,Z., Huang,S. and Yang,Z. (2009) Arabidopsis Formin3 directs the formation of actin cables and polarized growth in pollen tubes. *Plant Cell*, **21**, 3868–3884.
41. Li,D.D., Guan,H., Li,F., Liu,C.Z., Dong,Y.X., Zhang,X.S. and Gao,X.Q. (2017) Arabidopsis shaker pollen inward K+channel SPIK functions in SnRK1 complex-regulated pollen hydration on the stigma. *J. Integr. Plant Biol.*, **59**, 604–611.
42. Patel,R.V., Nahal,H.K., Breit,R. and Provart,N.J. (2012) BAR expressolog identification: expression profile similarity ranking of homologous genes in plant species. *Plant J.*, **71**, 1038–1050.
43. Ruprecht,C., Proost,S., Hernandez-Coronado,M., Ortiz-Ramirez,C., Lang,D., Rensing,S.A., Becker,J.D., Vandepoele,K. and Mutwil,M. (2017) Phylogenomic analysis of gene co-expression networks reveals the evolution of functional modules. *Plant J.*, **90**, 447–465.
44. Gu,F., Bringmann,M., Combs,J., Yang,J., Bergmann,D. and Nielsen,E. (2016) The Arabidopsis CSLD5 functions in cell plate formation in a cell cycle dependent manner. *Plant Cell*, **28**, 1722–1737.
45. Lee,I., Ambaru,B., Thakkar,P., Marcotte,E.M. and Rhee,S.Y. (2010) Rational association of genes with traits using a genome-scale gene network for Arabidopsis thaliana. *Nat. Biotechnol.*, **28**, 149–156.
46. Rhee,S.Y. and Mutwil,M. (2014) Towards revealing the functions of all genes in plants. *Trends Plant Sci.*, **19**, 212–221.
47. 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.
48. Zarrineh,P., Sánchez-Rodríguez,A., Hosseinkhan,N., Narimani,Z., Marchal,K. and Masoudi-Nejad,A. (2014) Genome-scale co-expression network comparison across Escherichia coli and Salmonella enterica serovar Typhimurium reveals significant conservation at the regulon level of local regulators despite their dissimilar lifestyles. *PLoS One*, **9**, e102871.
49. Gerstein,M.B., Rozowsky,J., Yan,K.-K., Wang,D., Cheng,C., Brown,J.B., Davis,C.A., Hillier,L., Sisu,C., Li,J.J. *et al.* (2014) Comparative analysis of the transcriptome across distant species. *Nature*, **512**, 445–448.
50. Ruprecht,C., Mutwil,M., Saxe,F., Eder,M., Nikoloski,Z. and Persson,S. (2011) Large-Scale Co-Expression approach to dissect secondary cell wall formation across plant species. *Front. Plant Sci.*, **2**, 1–13.
51. Park,C.Y., Wong,A.K., Greene,C.S., Rowland,J., Guan,Y., Bongo,L.A., Burdine,R.D. and Troyanskaya,O.G. (2013) Functional knowledge transfer for high-accuracy prediction of under-studied biological processes. *PLoS Comput. Biol.*, **9**, e1002957.
52. Tzfadia,O., Diels,T., De Meyer,S., Vandepoele,K., Aharoni,A. and Van de Peer,Y. (2016) CoExpNetViz: comparative co-expression networks construction and visualization tool. *Front. Plant Sci.*, **6**, 1194.
53. Ruprecht,C., Mendrinna,A., Tohge,T., Sampathkumar,A., Klie,S., Fernie,A.R., Nikoloski,Z., Persson,S. and Mutwil,M. (2016) FamNet: a framework to identify multiplied modules driving pathway expansion in plants. *Plant Physiol.*, **170**, 1878–1894.