Review Article

# Comparative co-expression analysis in plant biology

Sara Movahedi, Michiel Van Bel, Ken S. Heyndrickx and Klaas Vandepoele*

Department of Plant Systems Biology, VIB, 9052 Gent, Belgium
Department of Plant Biotechnology and Bioinformatics, Ghent University, 9052 Gent, Belgium

* Corresponding author.

Mailing address:

Klaas Vandepoele
Department of Plant Systems Biology, VIB2-Universiteit Gent
Technologiepark 927, B-9052 Gent (Belgium)
Tel. 32-9-3313822; fax 32-9-3313809
E-mail: klaas.vandepoele@psb.vib-ugent.be

Running title : Comparative transcriptomics in plants

## Abstract

The analysis of gene expression data generated by high-throughput microarray transcript profiling experiments has shown that transcriptionally coordinated genes are often functionally related. Based on large-scale expression compendia grouping multiple experiments, this guilt-by-association principle has been applied to study modular gene programs, identify cis-regulatory elements, or predict functions for unknown genes in different model plants. Recently, several studies have demonstrated how, through the integration of gene homology and expression information, correlated gene expression patterns can be compared between species. The incorporation of detailed functional annotations as well as experimental data describing protein-protein interactions, phenotypes or tissue specific expression, provides an invaluable source of information to identify conserved gene modules and translate biological knowledge from model organisms to crops. In this review, we describe the different steps required to systematically compare expression data across species. Apart from the technical challenges to compute and display expression networks from multiple species, some future applications of plant comparative transcriptomics are highlighted.

Keywords: comparative genomics, expression analysis, bioinformatics, orthology

## Introduction

Comparative sequence analysis is a successful tool to study homologous gene families (genes sharing common ancestry), define conserved gene functions between orthologs (homologs separated by a speciation event), and identify lineage- and species-specific genes. Most annotations of newly sequenced genomes are based on similarity with sequences for which functional information is available. Apart from conserved sequences, inter-species differences provide important clues about evolutionary history and species-specific adaptations (Hardison, 2003). Accelerated by technological innovations, genome-wide data describing functional properties including gene expression, protein-protein interactions and protein-DNA interactions is becoming available for an increasing number of model organisms. Consequently, the integration of functional genomics information provides, apart from gene sequence data, an additional layer of information to study gene function and regulation across species (Tirosh, Bilu & Barkai, 2007).

Depending on the availability of expression profiling technologies and the evolutionary distances between the species under investigation, a number of different approaches can be applied to study expression profiles between organisms (Lu, Huggins & Bar-Joseph, 2009). The hybridization of samples from closely related species to the same microarray requires compatible experimental conditions and has been first used in studies comparing different Brassicaceae species (Gong, Li, Ma, Indu Rupassara & Bohnert, 2005, Hammond, Broadley, Craigon, Higgins, Emmerson, Townsend, White & May, 2005, Taji, Seki, Satou, Sakurai, Kobayashi, Ishiyama, Narusaka, Narusaka, Zhu & Shinozaki, 2004, Weber, Harada, Vess, Roepenack-Lahaye & Clemens, 2004). To monitor specific responses between more distantly related species, multiple microarray experiments are combined to first identify differentially expressed (DE) genes in each species independently, and then compare these genes among different species. Downstream comparative sequence analysis of DE genes between different species or kingdoms makes it possible to identify evolutionary conserved responsive gene families as well as species-specific components. In addition, unknown genes showing a conserved response shared between multiple species are interesting targets for detailed molecular characterization (Vandenbroucke, Robbens, Vandepoele, Inze, Van de Peer & Van Breusegem, 2008). Similarly, Mustroph and co-workers successfully applied a comparative meta-analysis of low-oxygen stress responses to identify several unknown plant-specific hypoxia responsive genes (Mustroph, Lee, Oosumi, Zanetti, Yang, Ma, Yaghoubi-Masihi, Fukao & Bailey-Serres, 2010). More recently, microarray data sets were integrated to study orthologs and specific biological processes between more distantly related plant species, including *Arabidopsis thaliana* (*Arabidopsis*), *Oryza sativa* (rice) and *Populus* (poplar). Two pioneering studies, comparing microarray expression profiles between *Arabidopsis* and rice, focused on conservation and divergence of light

regulation during seedling development and the analysis of global transcriptomes from representative organ types between both plant model systems (Jiao, Ma, Strickland & Deng, 2005, Ma, Chen, Liu, Jiao, Su, Li, Wang, Cao, Sun, Zhang, Bao, Li, Pedersen, Bolund, Zhao, Yuan, Wong, Wang & Deng, 2005). Similarly, Street and co-workers identified several transcription factors involved in leaf development based on cross-species expression analysis of orthologous genes between *Arabidopsis* and poplar (Street, Sjodin, Bylesjo, Gustafsson, Trygg & Jansson, 2008).

Although comparative expression analysis is most straightforward when compatible expression data sets are used that cover equivalent conditions for all species, only a small fraction of all available data in different species can be utilized in this approach (Tirosh *et al.*, 2007). To overcome these limitations, pioneering comparative transcriptomics studies have shown that comparing co-expression, instead of the raw expression values, provides a valid alternative to identify gene modules (set of co-expressed genes potentially sharing similar function and regulation) and study their evolution (Bergmann, Ihmels & Barkai, 2004, Stuart, Segal, Koller & Kim, 2003). Stuart and colleagues developed a computational approach to identify conserved biological functions in different species by looking for correlated patterns of gene expression in microarrays from humans, fruit flies, worms, and yeast (Stuart *et al.*, 2003). Similarly, the integration of genome-wide expression data was used to study the modular architecture of regulatory programs in six evolutionary distant organisms (Bergmann *et al.*, 2004).

In this manuscript we give an overview of the different steps to systematically compare microarray expression data across species based on recent comparative transcriptomics studies in plants. Apart from the retrieval, normalization and annotation of microarray expression information, challenges related to the detection of co-expressed genes, the accurate delineation of gene orthology and the integration of expression networks and homology data are highlighted. Two case studies are presented demonstrating how conserved co-expression can be used to functionally annotate genes and to discriminate between co-orthologs with varying levels of expression conservation. Finally, we discuss some properties of conserved expression modules in plants and highlight some future applications.

**Processing and integration of plant expression data**

Gene expression profiling of different samples reveals whether genes are transcriptionally induced or repressed as a reaction to a certain treatment, disease, or at different developmental stages. Consequently, it is a powerful tool for target discovery, disease classification, pathway analysis and monitoring of biotic or abiotic responses. Among different available microarray

technologies, such as Affymetrix, Agilent and Roche/NimbleGen, the Affymetrix GeneChip is one of the most popular platforms to quantify steady-state transcript abundances (shortly, gene expression). On Affymetrix oligonucleotide microarrays, tens of thousands of probes, typically covering 25nt, are attached to a solid surface. Other microarray platforms, like Agilent, use only a few but longer probes to measure expression of a specific gene (Hardiman, 2004). After sample preparation, the outcome of the probe-target hybridization is quantified and intensity values of each cell (feature) are saved in a CEL file for a specific experiment. Apart from the expression values, standardized descriptions of experimental conditions and protocols are stored using the MIAME/Plant standard to facilitate data sharing (Zimmermann, Schildknecht, Craigon, Garcia-Hernandez, Gruissem, May, Mukherjee, Parkinson, Rhee, Wagner & Hennig, 2006). A detailed description of various experimental parameters is essential if, in a later stage, the identification of compatible experimental conditions across species is required. Repositories like Gene Expression Omnibus (GEO) (Barrett & Edgar, 2006) or ArrayExpress (Parkinson, Sarkans, Kolesnikov, Abeygunawardena, Burdett, Dylag, Emam, Farne, Hastings, Holloway, Kurbatova, Lukk, Malone, Mani, Pilicheva, Rustici, Sharma, Williams, Adamusiak, Brandizi, Sklyar & Brazma, 2011) are public microarray archives and provide thousands of expression profiling studies (Figure 1). All available microarray data for a specific organism, mostly focusing on an individual platform, are frequently combined to build large-scale expression compendia (see for example PLEXdb (Wise, Caldo, Hong, Shen, Cannon & Dickerson, 2007)) which summarize expression profiles in tens or hundreds of different conditions (Fierro, Vandenbussche, Engelen, Van de Peer & Marchal, 2008). For each experiment, the CEL files are retrieved and subsequently processed using a Chip Description File (CDF) in order to obtain a raw intensity value per gene. A CDF file describes probe locations and probeset groupings on the chip. During microarray analysis, mostly performed using algorithms such as MAS5 (Affymetrix proprietary method) or RMA/GCRMA (Irizarry, Hobbs, Collin, Beazer-Barclay, Antonellis, Scherf & Speed, 2003), intensity values of individual probes are summarized for a probeset, typically representing a specific locus, gene or transcript. The final expression data set is a matrix of genes (rows) and conditions (columns), which is background-corrected, normalized and finally summarized (Quackenbush, 2002).

In contrast to gene-based arrays, tiling arrays contain a large number of probes that cover a complete chromosome or genome and can be used, apart from standard expression profiling, for various applications including the detection of novel transcripts, chromatin immunoprecipitation of transcription factor protein-DNA interactions, profiling of epigenetic modifications, or the detection of DNA polymorphisms (Gregory, Yazaki & Ecker, 2008). Although repeat sequences can interfere with the reliable measurement of genome-wide expression, high-density tiling arrays are independent of known gene annotations and therefore provide an unbiased approach for different

profiling studies. This is in contrast with the GeneChip platform, which measures the expression of a given sequence (i.e. gene or transcript) using multiple probes grouped in a probeset (see Supporting Information, Note I).

According to a survey executed on November 2011, there were thirteen Affymetrix GeneChip microarray platforms publicly available in the NCBI GEO database for different plants (eight dicots and five monocots, see Figure 1). The number of CEL files available for these species varies a lot, from only twenty for sugar cane (*Sacharum officinarum*) to more than 7000 for *Arabidopsis*. Apart from a developmental plant expression atlas generated for *Arabidopsis* (Schmid, Davison, Henz, Pape, Demar, Vingron, Scholkopf, Weigel & Lohmann, 2005), large-scale expression compendia have been constructed, using a variety of platforms, for other species as well. Examples include barley (*Hordeum vulgare*) (Druka, Muehlbauer, Druka, Caldo, Baumann, Rostoks, Schreiber, Wise, Close, Kleinhofs, Graner, Schulman, Langridge, Sato, Hayes, McNicol, Marshall & Waugh, 2006), Medicago (*Medicago truncatula)* (Benedito, Torres-Jerez, Murray, Andriankaja, Allen, Kakar, Wandrey, Verdier, Zuber, Ott, Moreau, Niebel, Frickey, Weiller, He, Dai, Zhao, Tang & Udvardi, 2008), rice (Jiao, Tausta, Gandotra, Sun, Liu, Clay, Ceserani, Chen, Ma, Holford, Zhang, Zhao, Deng & Nelson, 2009, Wang, Xie, Chen, Tang, Yang, Ye, Liu, Lin, Xu, Xiao & Zhang, 2010), tobacco (*Nicotiana tabacum*) (Edwards, Bombarely, Story, Allen, Mueller, Coates & Jones, 2010) and soybean (*Glycine max*) (Libault, Farmer, Joshi, Takahashi, Langley, Franklin, He, Xu, May & Stacey, 2010). Although many plant expression studies integrated all available expression data, in some cases condition-dependent or pre-defined expression compendia focusing on specific developmental stages, tissues or stress conditions have been generated to study specific gene functions (De Bodt, Carvajal, Hollunder, Van den Cruyce, Movahedi & Inze, 2010, Usadel, Obayashi, Mutwil, Giorgi, Bassel, Tanimoto, Chow, Steinhauser, Persson & Provart, 2009a). Additional procedures can be applied to remove low-quality samples or to remove samples that could generate biases within the final compendium (Table 1). The latter is typically achieved by applying a statistical selection procedure to only select independent conditions or, reversely, by first grouping similar conditions and only retaining a single experiment as a representative for a set of related microarray conditions (Movahedi, Van de Peer & Vandepoele, 2011, Mutwil, Klie, Tohge, Giorgi, Wilkins, Campbell, Fernie, Usadel, Nikoloski & Persson, 2011). Although these selection procedures allow for the detection of specific conditions providing new expression information compared to the samples already included in the compendium, the number of genes that can be reliable measured through a specific microarray platform also provides an important parameter when compiling expression compendia. As for some species the number of genes that can be measured using a microarray differs substantially from the number of annotated genes in the genome (Mutwil *et al.*, 2011), missing genes provide an important drawback for many microarray-based co-expression tools (see for example Figure 3B).

**Detection of gene clusters and construction of co-expression networks**

In order to compare genome-wide expression profiles between different species, most studies apply a clustering algorithm to search, based on a large-scale expression compendium, for groups of highly co-expressed genes per species (Figure 2). The idea of clustering is to study groups of genes, sharing similar expression patterns, instead of individual ones. There are many different gene expression clustering tools available and each has its own advantages and disadvantages. Most clustering methods apply a similarity or a distance measure together with other parameters such as the number of clusters, the minimum/maximum cluster size or a quality measure to construct gene co-expression clusters (Xu & Wunsch, 2005). Overall, it is not easy to do a fair evaluation of how well an algorithm will perform on typical expression data sets, and under which circumstances one algorithm should be preferred over another (D'Haeseleer, 2005, Usadel *et al.*, 2009a).

Two of the most commonly used similarity measures for gene expression data are Euclidean distance and Pearson correlation coefficient (PCC). Other examples of measures that have been applied in comparative plants co-expression studies are cosine and Spearman's correlation coefficient (Table 1). To identify clusters of genes showing expression similarity, very simple as well as complex graph-based clustering algorithms have been developed. The most simple methods rank, for a selected gene, all other genes based on a similarity measure (e.g. descending PCC values) and then select a predefined number of top best ranked genes. Alternatively, gene selection can also be applied by retaining all genes with a PCC value above a pre-defined threshold. Mutual ranks, defined as the geometrical average of the correlation ranks, are frequently applied to keep weak but significant gene co-expression relationships which would not be retained when applying a fixed absolute similarity threshold. A derivative, the highest reciprocal rank (HRR), considers the maximum rank for a pair of genes (Table 1). Application of these rank-based gene selection criteria are frequently used as a simple and fast substitute for more complex clustering algorithms since they generate a set of co-expressed genes for each query gene (i.e. gene-centric clustering, see Figure 2). In this case, the number of co-expression clusters is close or equal to the number of genes available in the expression data set and clusters are potentially overlapping on a genome-wide scale.

Apart from simple rank-based gene-centric clustering approaches, more advanced algorithms apply graph-theory to find groups of genes showing similar expression profiles. In general, a weighted graph of genes (nodes) is constructed where each pair of genes is connected by an edge and the edge weight is defined by the expression similarity between the genes. Graph-based clustering tools try to identify highly connected nodes (sub-graphs) in this expression network representing gene expression clusters. Whereas clique finders isolate fully connected sub-graphs,

other tools apply a variety of heuristic or statistical methods to find gene clusters. This can be done by considering only the first neighbors of a query (or seed) gene or all nodes within n steps away from the query gene (Node Vicinity Network, NVN). CAST (Cluster Affinity Search Technique) (Ben-Dor, Shamir & Yakhini, 1999, Vandepoele, Quimbaya, Casneuf, De Veylder & Van de Peer, 2009), the Confeito algorithm (Ogata, Sakurai, Suzuki, Aoki, Saito & Shibata, 2009), Weighted Gene Co-expression Network Analysis (WGCNA) (Langfelder & Horvath, 2008), Random Matrix Theory (RMT) (Luo, Yang, Zhong, Gao, Khan, Thompson & Zhou, 2007) and Heuristic Cluster Chiseling Algorithm (HCCA) (Mutwil, Usadel, Schutte, Loraine, Ebenhoh & Persson, 2010) are examples of graph-based algorithms which have been applied for defining gene co-expression clusters in plants (Table 1).

**Comparing co-expression networks across species**

A major objective in comparative expression studies is the systematic comparison of gene clusters across species using homologous or orthologous genes. Defining sequence-based orthologs is a powerful approach to link expression datasets across species (Table 1) and to identify genes with conserved gene functions or conserved modules that participate in similar biological processes (Bergmann *et al.*, 2004, Lu *et al.*, 2009, Stuart *et al.*, 2003). Although different approaches are available to identify homologous and orthologous genes (Koonin, 2005), most of them start from the output of a global all-against-all sequence similarity search. Whereas NCBI HomoloGene defines homologous genes in completely sequenced eukaryotic genomes (Sayers, Barrett, Benson, Bolton, Bryant, Canese, Chetvernin, Church, DiCuccio, Federhen, Feolo, Fingerman, Geer, Helmberg, Kapustin, Landsman, Lipman, Lu, Madden, Madej, Maglott, Marchler-Bauer, Miller, Mizrachi, Ostell, Panchenko, Phan, Pruitt, Schuler, Sequeira, Sherry, Shumway, Sirotkin, Slotta, Souvorov, Starchenko, Tatusova, Wagner, Wang, Wilbur, Yaschenko & Ye, 2011), the PFAM database provides information about conserved protein domains and families (Finn, Mistry, Tate, Coggill, Heger, Pollington, Gavin, Gunasekaran, Ceric, Forslund, Holm, Sonnhammer, Eddy & Bateman, 2010). Although reciprocal best hits (RBH) provide a practical solution to identify orthologs between closely related species, OrthoMCL and Inparanoid (Li, Stoeckert & Roos, 2003, Ostlund, Schmitt, Forslund, Kostler, Messina, Roopra, Frings & Sonnhammer) are more advanced methods to construct orthologous groups across genomes because they model, apart from orthology through RBH, also inparalogy (gene duplication events post-dating speciation). Consequently, species-specific gene family expansions are correctly represented in OrthoMCL orthologous groups while RBH approaches only retain a single gene as ortholog (excluding other inparalogs). In the latter case it is possible that erroneous conclusions about gene family expression evolution are drawn, especially if the expression profiles of the

inparalogs (or co-orthologs) have diverged. Whereas Inparanoid identifies orthologs and inparalogs in a pairwise manner, OrthoMCL can delineate orthologous clusters between multiple genomes in a single run. A detailed comparison of plants orthologs from multiple species revealed that 70-90% of OrthoMCL families could be confirmed by phylogenetic tree construction (Proost, Van Bel, Sterck, Billiau, Van Parys, Van de Peer & Vandepoele, 2009). Although phylogeny-based orthology predictions are available in a number of plant comparative genomics resources (Martinez, 2011), sequence similarity clustering methods are less computer intensive and more easily applicable. However, simple sequence similarity approaches have a higher risk of missing genes involved in complex many-to-many orthology relationships between more distantly related species (Kuzniar, van Ham, Pongor & Leunissen, 2008, Proost *et al.*, 2009, Van Bel, Proost, Wischnitzki, Movahedi, Scheerlinck, Van de Peer & Vandepoele, 2012). Reversely, protein domain-based methods might assign false orthology relationships between multi-domain protein coding genes that are only distantly related based on the presence of single frequently occurring domain (e.g. ankyrin repeat, WD40, F-box). Tools like CoGe or PLAZA provide synteny information to delineate putative orthologs (Lyons, Pedersen, Kane, Alam, Ming, Tang, Wang, Bowers, Paterson, Lisch & Freeling, 2008, Van Bel *et al.*, 2012), with the latter applying an ensemble approach to integrate results from different methods when searching for orthologous genes (PLAZA Integrative Orthology approach).

So far, most comparative expression analyses have combined gene expression clusters per species with homology information to identify conserved gene expression (Table 1). Examples in plants include Co-expressed biological Processes (CoP) (Ogata, Suzuki, Sakurai & Shibata, 2010), Expression Context Conservation (ECC) (Movahedi *et al.*, 2011), Plant Network (PLaNet) (Mutwil *et al.*, 2011) and STARNET2 (Jupiter, Chen & VanBuren, 2009) (Table 1). Although the CoP database simply provides a list of co-expressed genes in the other species starting from an individual query gene, the other tools include gene homology information to filter the co-expression information from the different species (see blue dashed lines in Figure 2). Gene expression is typically compared between species in a pairwise manner and, optionally, information about conserved genes in multiple species is combined (Mutwil *et al.*, 2011). Although this approach provides a first glimpse on the co-expressed genes that are conserved between different species (Humphry, Bednarek, Kemmerling, Koh, Stein, Gobel, Stuber, Pislewska-Bednarek, Loraine, Schulze-Lefert, Somerville & Panstruga, 2010), recently developed methods also apply statistical tests to verify if the number of shared orthologs between two expression clusters is significant (Chikina & Troyanskaya, 2011, Movahedi *et al.*, 2011, Mutwil *et al.*, 2011, Zarrineh, Fierro, Sanchez-Rodriguez, De Moor, Engelen & Marchal, 2011). Since most approaches use gene homology or orthology information to connect co-expression networks between different species, larger co-expression clusters will logically also yield a higher number of shared orthologs. Similarly, for genes involved in many-to-many orthology

relationships, the probability to have shared orthologs between co-expression clusters is also higher compared to small families with one-to-one orthology relationships. As shown in Figure S2, the application of a statistical significance test can be used to objectively define if, based on the gene co-expression cluster sizes and homologous genes or families, the number of shared orthologs is significantly higher than expected by chance. In comparative studies where the homologous genes from the different species can be classified using one-to-one orthology, the hypergeometric distribution and Pearson's chi-square test have been used to estimate if the number of shared orthologs is significant (Chikina & Troyanskaya, 2011, Zarrineh *et al.*, 2011). However, for species with many multi-gene families like plants (Vandepoele & Van de Peer, 2005), the application of empirical significance testing using a permutation test provides a more reliable alternative as the probability of finding shared orthologs between two expression clusters differs for genes belonging to families with different sizes. To the best of our knowledge, only PLANET and ECC applied a statistical evaluation taking into consideration different gene family sizes (Table 1), the latter including different null models to reliably estimate the significance levels of conserved co-expression controlling for network properties such as connectivity (i.e. the degree distribution of co-expressed genes within the network) or tissue specificity (Movahedi *et al.*, 2011). As a consequence, these models correct for specific expression breadth biases that might exist in co-expression clusters for certain genes when performing statistical evaluation.

To determine the most optimal conserved co-expression module, the recently developed COMODO method uses a cross-species co-clustering approach that simultaneously evaluates the homology relations and the extension of co-expression seed modules. Starting from seeds in each species, these seed modules are gradually expanded (by addition of co-expressed genes ranked using PCC similarity information) in each of the species until a pair of modules is found for which the number of shared orthologs is statistically optimal (Zarrineh *et al.*, 2011). Although this approach explores the two-dimensional parameter landscape (Figure S2) to find the best co-expression module definition, it is still required to pre-specify a co-expression stringency value for seed identification.

Complementary to two-step approaches which first define expression clusters and then filters co-expressed edges in the networks using gene homology information, Ficklin and Feltus (Ficklin & Feltus, 2011) used a global network alignment approach to combine the co-expression topology and homology information and to delineate conserved modules. Although this approach successfully identified several conserved modules between rice and maize, the applied method did not include a statistical evaluation of the conserved sub-graphs.

**Functional annotation and network visualization**

To study the biological processes behind conserved co-expression modules, different functional annotation systems as well as experimental data have been used. Although several studies relied on Gene Ontology (GO) annotations to identify enriched gene functions within conserved modules, information from KEGG pathways (Kanehisa, Goto, Furumichi, Tanabe & Hirakawa, 2010), Reactome (Tsesmetzis, Couchman, Higgins, Smith, Doonan, Seifert, Schmidt, Vastrik, Birney, Wu, D'Eustachio, Stein, Morris, Bevan & Walsh, 2008) or MapMan (Usadel, Poree, Nagel, Lohse, Czedik-Eysenberg & Stitt, 2009b) has also been exploited (Table 1). Gene annotation enrichment analysis is a high-throughput strategy that increases the likelihood for investigators to identify biological processes most pertinent to their study, based on an underlying enrichment algorithm (Huang da, Sherman & Lempicki, 2009). The integration of known protein-protein interactions, tissue specific expression or phenotypic information from mutant lines provides an additional level of experimental information that has been used to characterize conserved modules (Ficklin & Feltus, 2011, Movahedi *et al.*, 2011, Mutwil *et al.*, 2011).

Graphviz and Cytoscape (Smoot, Ono, Ruscheinski, Wang & Ideker, 2011) are frequently applied software tools to graphically integrate expression networks, homology information and functional annotations (Table 1). Typically, genes are depicted by nodes while different edge attributes are used to represent expression similarity and homology information within and between species (Figure 3A). Although functional information about individual genes can be displayed using node attributes based on color, shape or outline thickness, the wealth of GO, KEGG or MapMan functional categories as well as various experimental properties makes it difficult to summarize all information in one single view. Although filtering on specific gene functions or a GO biological process provides a practical solution to reduce network complexity, the construction of meta-networks (also referred to as module or ontology networks) makes it possible to explore regulatory interactions between groups of functionally related genes rather than between individual genes (Table 1). Furthermore, meta-networks are an important instrument to identify regulatory interactions and cross-talk between different processes (Mutwil *et al.*, 2011).

Although both STARNET2 and PlaNet host a website where users can browse co-expression networks, only the latter can be used to successfully generate cross species networks due to missing rice HomoloGene information in STARNET2. Although Mohavedi et al. and Ficklin et al. published several examples of conserved co-expression modules between *Arabidopsis*-rice and rice-maize (Ficklin & Feltus, 2011, Movahedi *et al.*, 2011), respectively, an online resource to browse these conserved modules is currently unavailable. The COP database displays small co-expression networks for individual genes but reports conserved orthologs between two co-expression clusters from different species in a textual manner. Clearly, it remains an important challenge to provide an

interactive web-browser application where, apart from the co-expression networks from multiple species, different functional annotations, phenotypes, protein-protein interactions, and complex orthology gene relationships can also be displayed.

## Studying conserved gene functions using comparative co-expression analysis

To demonstrate the power of comparative co-expression methods to study gene functions across species, Figure 3A displays the result of a comparative transcriptomics analysis for the *Arabidopsis* gene ETG1 (AT2G40550). Whereas this gene was previously described as a conserved E2F target gene with unknown function (Vandepoele, Vlieghe, Florquin, Hennig, Beemster, Gruissem, Van de Peer, Inze & De Veylder, 2005), recent experimental work revealed it has an essential role in sister chromatin cohesion during DNA replication (Takahashi, Quimbaya, Schubert, Lammens, Vandepoele, Schubert, Matsui, Inze, Berx & De Veylder, 2010). To identify the biological role of ETG1 and verify whether it is part of a conserved co-expression module in plants, we first characterized the gene's co-expression context based on a general *Arabidopsis* expression compendium from CORNET (De Bodt *et al.*, 2010). Retrieval of the 50 most co-expressed genes based on the PCC yielded a set of genes showing a strong GO enrichment towards 'cellular DNA replication' (90-fold enrichment, p-value 1.33e-36). Enrichment analysis for known plant cis-regulatory elements using ATCOECIS (Vandepoele *et al.*, 2009) yielded enrichment for the E2F binding site TTTCCCGC (18-fold enrichment, p-value 1.41e-18), confirming that ETG1 is a putative E2F target gene. To explore whether this functional enrichment is evolutionary conserved, we first searched for ETG1 orthologs using the PLAZA 2.0 Integrative Orthology Viewer in species for which microarray data is publicly available. Whereas poplar, maize and rice have one ETG1 ortholog (PT19G07260, ZM03G04050 and OS01G07260, respectively), two copies were found in soybean (GM04G39990 and GM06G14860). Next, for each species a general expression compendium was compiled using Affymetrix experiments from GEO and the top-50 co-expressed genes were isolated in these organisms as well. Finally, the number of shared orthologs between the different co-expression clusters was determined and the resulting conserved modules were delineated (Figure 3A). Based on the ETG1 *Arabidopsis* co-expression cluster, 9 and 13 orthologous genes were conserved with the co-expression clusters for poplar and rice, respectively. Whereas for both species the fraction of conserved orthologs is much higher than expected by chance (p-value <1e-5, see inset Figure 3A), the functions of these orthologs (MCM2-5, MCM7, RPA70B, RPA70D and POLA3) as well as the expression context conservation in both monocots and dicots lend support for the conserved role of ETG1 in DNA replication. Querying the CoP database for ETG1 reports a smaller number of co-expressed genes but confirms the functional

enrichment towards DNA replication as well as the shared orthologs MCM3, MCM6 and POL3A between *Arabidopsis* and rice. Whereas the PlaNet platform did not directly confirm the biological role of ETG1 in DNA replication based on the *Arabidopsis* co-expression cluster, the comparative analysis confirmed that up to ten known DNA replication genes showed conserved co-expression in other plants. Examples included multiple replication factors, two ribonucleotide reductases, PCNA, ORC2 and different DNA polymerase subunits.

Based on the frequent nature of many-to-many gene orthology relationships in plants, mediated by large-scale duplication events (Van de Peer, Fawcett, Proost, Sterck & Vandepoele, 2009), comparative transcriptomics also offers a practical solution to identify functional homologs in multi-gene families (Chikina & Troyanskaya, 2011). Apart from detecting conserved gene modules, the ECC method can also be applied to identify orthologs and inparalogs with conserved co-expression between different species for which large-scale expression data is available. For a set of 21 ubiquitin-activating enzyme homologs from seven species (Figure 3B), the systematic examination of conserved co-expression between all family members makes it possible to explore whether duplicates show different conservation patterns. Application of the ECC method using the 50 most co-expressed genes revealed that, for those orthologs which have expression data, in poplar, Medicago, soybean, Arabidopsis and maize ECC patterns with orthologs from other species were different between inparalogs. This result reveals that for at least five species both co-orthologs with conserved and non-conserved co-expression contexts exist, making the transfer of biological information between different species challenging.

**Biological applications and future directions**

Hypothesis-driven gene discovery remains one of the most promising applications for co-expression networks. Whereas this principle is not new in plant genomics (Usadel *et al.*, 2009a), the analysis of expression networks between more distantly related species exploits the assumption that predicted gene-function associations that occur by chance within one organism will not be conserved in a multi-species data set. Indeed, several plant studies identified conserved expression modules related to photosynthesis, translation, cell cycle and DNA metabolism, both in dicots and monocots (Ficklin & Feltus, 2011, Movahedi *et al.*, 2011, Mutwil *et al.*, 2011). As a consequence, the analysis of conserved modules with enriched gene functions and the comparison of gene sets with enriched phenotypes provide an invaluable approach for biological gene discovery in model species and to translate new gene functions to species with agricultural or economical value. Reversely, the analysis of orthologous genes lacking expression conservation might reveal biological adaptations linking

genotype to phenotype (Tirosh *et al.*, 2007). Based on the statistical evaluation of genes lacking shared orthologs between *Arabidopsis* and rice genes, Movahedi and co-workers reported that non-conserved ECC genes involved in stress response and signal transduction could provide a connection between regulatory evolution and environmental adaptations (Movahedi *et al.*, 2011).

The integration of new experiments describing specific transcriptional responses or tissue specific expression will provide, apart from GO annotations, an important complementary source of functional information to annotate homologs and to transfer biological knowledge between species based on conserved gene modules,. Nevertheless, this would require that, for example using ontology-based experimental annotations (De Bodt *et al.*, 2010, Jaiswal, Avraham, Ilic, Kellogg, McCouch, Pujar, Reiser, Rhee, Sachs, Schaeffer, Stein, Stevens, Vincent, Ware & Zapata, 2005), similar conditions in different species could easily be identified within public databases covering thousands of profiling experiments. The recently developed Expressolog Tree Viewer, part of the Bio-Array Resource for Plant Biology website (http://bar.utoronto.ca/), demonstrates how in several cases equivalent conditions between different plants can be identified and how direct comparisons of expression profiles between homologous genes can be used to identify (co-)orthologs showing conserved spatial-temporal expression. Nevertheless, as divergence time and morphological differences between species increase (e.g. between monocotyledonous and eudicotyledonous plants), finding equivalent tissues becomes challenging. Consequently, and in contrast to co-expression comparisons (Figure 3B), this setup only allows for a limited number of conditions that can directly be compared across homologs of different species.

The application of next-generation sequencing to quantify plant transcriptomes (RNA-Seq) will generate new opportunities to study and compare expression profiles between species (Figure 1). For example, detailed comparisons of different alternative transcripts within a co-expression network context will provide important information about the biological processes different splicing variants are involved in. Furthermore, studying alternative transcript expression levels within a comparative framework will generate new insights into the evolution and functional significance of alternative splicing in plants. However, the development and application of robust data processing and normalization methods will be essential in order to combine RNA-Seq experiments with varying sequencing depths into uniform and comparable expression compendia (Tarazona, Garcia-Alcalde, Dopazo, Ferrer & Conesa, 2011).

In conclusion, the rapid accumulation of genome-wide data describing both plant genome sequences and a variety of functional properties will require the continuous development of systems biology approaches as well as user-friendly databases to extract biological knowledge and exchange information between experimental and computational plant biologists.

**Acknowledgements**

We thank Annick Bleys for help in preparing the manuscript and Yves Van de Peer for general support. K.S.H. is indebted to the Agency for Innovation by Science and Technology (IWT) in Flanders for a predoctoral fellowship. K.V. acknowledges the support of Ghent University (Multidisciplinary Research Partnership “Bioinformatics: from nucleotides to networks”). This project is funded by the Research Foundation–Flanders and the Belgian Federal Science Policy Office: IUAP P6/25 (BioMaGNet).

**References**


Barrett T. & Edgar R. (2006) Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. *Methods Enzymol*, **411**, 352-369.

Ben-Dor A., Shamir R. & Yakhini Z. (1999) Clustering gene expression patterns. *J Comput Biol*, **6**, 281-297.

Benedito V.A., Torres-Jerez I., Murray J.D., Andriankaja A., Allen S., Kakar K., Wandrey M., Verdier J., Zuber H., Ott T., Moreau S., Niebel A., Frickey T., Weiller G., He J., Dai X., Zhao P.X., Tang Y. & Udvardi M.K. (2008) A gene expression atlas of the model legume Medicago truncatula. *Plant J*, **55**, 504-513.

Bergmann S., Ihmels J. & Barkai N. (2004) Similarities and differences in genome-wide expression data of six organisms. *PLoS Biol*, **2**, E9.

Chikina M.D. & Troyanskaya O.G. (2011) Accurate Quantification of Functional Analogy among Close Homologs. *PLoS Comput Biol*, **7**, e1001074.

D'Haeseleer P. (2005) How does gene expression clustering work? *Nat Biotechnol*, **23**, 1499-1501.

De Bodt S., Carvajal D., Hollunder J., Van den Cruyce J., Movahedi S. & Inze D. (2010) CORNET: a user-friendly tool for data mining and integration. *Plant Physiol*, **152**, 1167-1179.

Druka A., Muehlbauer G., Druka I., Caldo R., Baumann U., Rostoks N., Schreiber A., Wise R., Close T., Kleinhofs A., Graner A., Schulman A., Langridge P., Sato K., Hayes P., McNicol J., Marshall D. & Waugh R. (2006) An atlas of gene expression from seed to seed through barley development. *Funct Integr Genomics*, **6**, 202-211.

Edwards K.D., Bombarely A., Story G.W., Allen F., Mueller L.A., Coates S.A. & Jones L. (2010) TobEA: an atlas of tobacco gene expression from seed to senescence. *BMC Genomics*, **11**, 142.

Ficklin S.P. & Feltus F.A. (2011) Gene Coexpression Network Alignment and Conservation of Gene Modules between Two Grass Species: Maize and Rice. *Plant Physiol*, **156**, 1244-1256.

Fierro A.C., Vandenbussche F., Engelen K., Van de Peer Y. & Marchal K. (2008) Meta Analysis of Gene Expression Data within and Across Species. *Curr Genomics*, **9**, 525-534.

Finn R.D., Mistry J., Tate J., Coggill P., Heger A., Pollington J.E., Gavin O.L., Gunasekaran P., Ceric G., Forslund K., Holm L., Sonnhammer E.L., Eddy S.R. & Bateman A. (2010) The Pfam protein families database. *Nucleic Acids Res*, **38**, D211-222.

Gong Q., Li P., Ma S., Indu Rupassara S. & Bohnert H.J. (2005) Salinity stress adaptation competence in the extremophile Thellungiella halophila in comparison with its relative Arabidopsis thaliana. *Plant J*, **44**, 826-839.

Gregory B.D., Yazaki J. & Ecker J.R. (2008) Utilizing tiling microarrays for whole-genome analysis in plants. *Plant J*, **53**, 636-644.

Hammond J.P., Broadley M.R., Craigon D.J., Higgins J., Emmerson Z.F., Townsend H.J., White P.J. & May S.T. (2005) Using genomic DNA-based probe-selection to improve the sensitivity of high-density oligonucleotide arrays when applied to heterologous species. *Plant Methods*, **1**, 10.

Hardiman G. (2004) Microarray platforms--comparisons and contrasts. *Pharmacogenomics*, **5**, 487-502.

Hardison R.C. (2003) Comparative genomics. *PLoS Biol*, **1**, E58.

Huang da W., Sherman B.T. & Lempicki R.A. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*, **37**, 1-13.

Humphry M., Bednarek P., Kemmerling B., Koh S., Stein M., Gobel U., Stuber K., Pislewska-Bednarek M., Loraine A., Schulze-Lefert P., Somerville S. & Panstruga R. (2010) A regulon conserved in monocot and dicot plants defines a functional module in antifungal plant immunity. *Proc Natl Acad Sci U S A*, **107**, 21896-21901.

Irizarry R.A., Hobbs B., Collin F., Beazer-Barclay Y.D., Antonellis K.J., Scherf U. & Speed T.P. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249-264.

Jaiswal P., Avraham S., Ilic K., Kellogg E.A., McCouch S., Pujar A., Reiser L., Rhee S.Y., Sachs M.M., Schaeffer M., Stein L., Stevens P., Vincent L., Ware D. & Zapata F. (2005) Plant Ontology (PO): a Controlled Vocabulary of Plant Structures and Growth Stages. *Comp Funct Genomics*, **6**, 388-397.

Jiao Y., Ma L., Strickland E. & Deng X.W. (2005) Conservation and divergence of light-regulated genome expression patterns during seedling development in rice and Arabidopsis. *Plant Cell*, **17**, 3239-3256.

Jiao Y., Tausta S.L., Gandotra N., Sun N., Liu T., Clay N.K., Ceserani T., Chen M., Ma L., Holford M., Zhang H.Y., Zhao H., Deng X.W. & Nelson T. (2009) A transcriptome atlas of rice cell types uncovers cellular, functional and developmental hierarchies. *Nat Genet*, **41**, 258-263.

Jupiter D., Chen H. & VanBuren V. (2009) STARNET 2: a web-based tool for accelerating discovery of gene regulatory networks using microarray co-expression data. *BMC Bioinformatics*, **10**, 332.

Kanehisa M., Goto S., Furumichi M., Tanabe M. & Hirakawa M. (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res*, **38**, D355-360.

Koonin E.V. (2005) Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet*, **39**, 309-338.

Kuzniar A., van Ham R.C., Pongor S. & Leunissen J.A. (2008) The quest for orthologs: finding the corresponding gene across genomes. *Trends Genet*, **24**, 539-551.

Langfelder P. & Horvath S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, **9**, 559.

Li L., Stoeckert C.J., Jr. & Roos D.S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*, **13**, 2178-2189.

Libault M., Farmer A., Joshi T., Takahashi K., Langley R.J., Franklin L.D., He J., Xu D., May G. & Stacey G. (2010) An integrated transcriptome atlas of the crop model Glycine max, and its use in comparative analyses in plants. *Plant J*, **63**, 86-99.

Lu Y., Huggins P. & Bar-Joseph Z. (2009) Cross species analysis of microarray expression data. *Bioinformatics*, **25**, 1476-1483.

Luo F., Yang Y., Zhong J., Gao H., Khan L., Thompson D.K. & Zhou J. (2007) Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory. *BMC Bioinformatics*, **8**, 299.

Lyons E., Pedersen B., Kane J., Alam M., Ming R., Tang H., Wang X., Bowers J., Paterson A., Lisch D. & Freeling M. (2008) Finding and comparing syntenic regions among Arabidopsis and the outgroups papaya, poplar, and grape: CoGe with rosids. *Plant Physiol*, **148**, 1772-1781.

Ma L., Chen C., Liu X., Jiao Y., Su N., Li L., Wang X., Cao M., Sun N., Zhang X., Bao J., Li J., Pedersen S., Bolund L., Zhao H., Yuan L., Wong G.K., Wang J. & Deng X.W. (2005) A microarray analysis of the rice transcriptome and its comparison to Arabidopsis. *Genome Res*, **15**, 1274-1283.

Martinez M. (2011) Plant protein-coding gene families: emerging bioinformatics approaches. *Trends Plant Sci*, **in press**.

Movahedi S., Van de Peer Y. & 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.

Mustroph A., Lee S.C., Oosumi T., Zanetti M.E., Yang H., Ma K., Yaghoubi-Masihi A., Fukao T. & Bailey-Serres J. (2010) Cross-kingdom comparison of transcriptomic adjustments to low-oxygen stress highlights conserved and plant-specific responses. *Plant Physiol*, **152**, 1484-1500.

Mutwil M., Klie S., Tohge T., Giorgi F.M., Wilkins O., Campbell M.M., Fernie A.R., Usadel B., Nikoloski Z. & Persson S. (2011) PlaNet: Combined Sequence and Expression Comparisons across Plant Networks Derived from Seven Species. *Plant Cell*, **23**, 895-910.

Mutwil M., Usadel B., Schutte M., Loraine A., Ebenhoh O. & Persson S. (2010) Assembly of an interactive correlation network for the Arabidopsis genome using a novel heuristic clustering algorithm. *Plant Physiol*, **152**, 29-43.

Ogata Y., Sakurai N., Suzuki H., Aoki K., Saito K. & Shibata D. (2009) The prediction of local modular structures in a co-expression network based on gene expression datasets. *Genome Inform*, **23**, 117-127.

Ogata Y., Suzuki H., Sakurai N. & Shibata D. (2010) CoP: a database for characterizing co-expressed gene modules with biological information in plants. *Bioinformatics*, **26**, 1267-1268.

Ostlund G., Schmitt T., Forslund K., Kostler T., Messina D.N., Roopra S., Frings O. & Sonnhammer E.L. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res*, **38**, D196-203.

Parkinson H., Sarkans U., Kolesnikov N., Abeygunawardena N., Burdett T., Dylag M., Emam I., Farne A., Hastings E., Holloway E., Kurbatova N., Lukk M., Malone J., Mani R., Pilicheva E., Rustici G., Sharma A., Williams E., Adamusiak T., Brandizi M., Sklyar N. & Brazma A. (2011) ArrayExpress update--an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Res*, **39**, D1002-1004.

Proost S., Van Bel M., Sterck L., Billiau K., Van Parys T., Van de Peer Y. & Vandepoele K. (2009) PLAZA: A Comparative Genomics Resource to Study Gene and Genome Evolution in Plants. *Plant Cell*.

Quackenbush J. (2002) Microarray data normalization and transformation. *Nat Genet*, **32 Suppl**, 496-501.

Sayers E.W., Barrett T., Benson D.A., Bolton E., Bryant S.H., Canese K., Chetvernin V., Church D.M., DiCuccio M., Federhen S., Feolo M., Fingerman I.M., Geer L.Y., Helmberg W., Kapustin Y., Landsman D., Lipman D.J., Lu Z., Madden T.L., Madej T., Maglott D.R., Marchler-Bauer A., Miller V., Mizrachi I., Ostell J., Panchenko A., Phan L., Pruitt K.D., Schuler G.D., Sequeira E., Sherry S.T., Shumway M., Sirotkin K., Slotta D., Souvorov A., Starchenko G., Tatusova T.A., Wagner L., Wang Y., Wilbur W.J., Yaschenko E. & Ye J. (2011) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, **39**, D38-51.

Schmid M., Davison T.S., Henz S.R., Pape U.J., Demar M., Vingron M., Scholkopf B., Weigel D. & Lohmann J.U. (2005) A gene expression map of Arabidopsis thaliana development. *Nat Genet*, **37**, 501-506.

Smoot M.E., Ono K., Ruscheinski J., Wang P.L. & Ideker T. (2011) Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, **27**, 431-432.

Street N.R., Sjodin A., Bylesjo M., Gustafsson P., Trygg J. & Jansson S. (2008) A cross-species transcriptomics approach to identify genes involved in leaf development. *BMC Genomics*, **9**, 589.

Stuart J.M., Segal E., Koller D. & Kim S.K. (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science*, **302**, 249-255.

Taji T., Seki M., Satou M., Sakurai T., Kobayashi M., Ishiyama K., Narusaka Y., Narusaka M., Zhu J.K. & Shinozaki K. (2004) Comparative genomics in salt tolerance between Arabidopsis and aRabidopsis-related halophyte salt cress using Arabidopsis microarray. *Plant Physiol*, **135**, 1697-1709.

Takahashi N., Quimbaya M., Schubert V., Lammens T., Vandepoele K., Schubert I., Matsui M., Inze D., Berx G. & De Veylder L. (2010) The MCM-binding protein ETG1 aids sister chromatid cohesion required for postreplicative homologous recombination repair. *PLoS Genet*, **6**, e1000817.

Tarazona S., Garcia-Alcalde F., Dopazo J., Ferrer A. & Conesa A. (2011) Differential expression in RNA-seq: A matter of depth. *Genome Res*, **21**, 2213-2223.

Tirosh I., Bilu Y. & Barkai N. (2007) Comparative biology: beyond sequence analysis. *Curr Opin Biotechnol*, **18**, 371-377.

Tsesmetzis N., Couchman M., Higgins J., Smith A., Doonan J.H., Seifert G.J., Schmidt E.E., Vastrik I., Birney E., Wu G., D'Eustachio P., Stein L.D., Morris R.J., Bevan M.W. & Walsh S.V. (2008) Arabidopsis reactome: a foundation knowledgebase for plant systems biology. *Plant Cell*, **20**, 1426-1436.

Usadel B., Obayashi T., Mutwil M., Giorgi F.M., Bassel G.W., Tanimoto M., Chow A., Steinhauser D., Persson S. & Provart N.J. (2009a) Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. *Plant Cell Environ*, **32**, 1633-1651.

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

Van Bel M., Proost S., Wischnitzki E., Movahedi S., Scheerlinck C., Van de Peer Y. & Vandepoele K. (2012) Dissecting plant genomes with the PLAZA comparative genomics platform. *Plant Phys*, **158**, 590-600.

Van de Peer Y., Fawcett J.A., Proost S., Sterck L. & Vandepoele K. (2009) The flowering world: a tale of duplications. *Trends Plant Sci*, **14**, 680-688.

Vandenbroucke K., Robbens S., Vandepoele K., Inze D., Van de Peer Y. & Van Breusegem F. (2008) Hydrogen peroxide-induced gene expression across kingdoms: a comparative analysis. *Mol Biol Evol*, **25**, 507-516.

Vandepoele K., Quimbaya M., Casneuf T., De Veylder L. & Van de Peer Y. (2009) Unraveling transcriptional control in Arabidopsis using cis-regulatory elements and coexpression networks. *Plant Physiol*, **150**, 535-546.

Vandepoele K. & Van de Peer Y. (2005) Exploring the plant transcriptome through phylogenetic profiling. *Plant Physiol*, **137**, 31-42.

Vandepoele K., Vlieghe K., Florquin K., Hennig L., Beemster G.T., Gruissem W., Van de Peer Y., Inze D. & De Veylder L. (2005) Genome-wide identification of potential plant E2F target genes. *Plant Physiol*, **139**, 316-328.

Wang L., Xie W., Chen Y., Tang W., Yang J., Ye R., Liu L., Lin Y., Xu C., Xiao J. & Zhang Q. (2010) A dynamic gene expression atlas covering the entire life cycle of rice. *Plant J*, **61**, 752-766.

Weber M., Harada E., Vess C., Roepenack-Lahaye E. & Clemens S. (2004) Comparative microarray analysis of Arabidopsis thaliana and Arabidopsis halleri roots identifies nicotianamine synthase, a ZIP transporter and other genes as potential metal hyperaccumulation factors. *Plant J*, **37**, 269-281.

Wise R.P., Caldo R.A., Hong L., Shen L., Cannon E. & Dickerson J.A. (2007) BarleyBase/PLEXdb. *Methods Mol Biol*, **406**, 347-363.

Xu R. & Wunsch D., 2nd (2005) Survey of clustering algorithms. *IEEE Trans Neural Netw*, **16**, 645-678.

Zarrineh P., Fierro A.C., Sanchez-Rodriguez A., De Moor B., Engelen K. & Marchal K. (2011) COMODO: an adaptive coclustering strategy to identify conserved coexpression modules between organisms. *Nucleic Acids Res*, **39**, e41.

Zimmermann P., Schildknecht B., Craigon D., Garcia-Hernandez M., Gruissem W., May S., Mukherjee G., Parkinson H., Rhee S., Wagner U. & Hennig L. (2006) MIAME/Plant - adding value to plant microarrray experiments. *Plant Methods*, **2**, 1.

**Table 1. Overview of cross-species co-expression studies in plants.**

| | STARNET2 | CoP | PlaNet | Maize- rice | ECC |
|---|---|---|---|---|---|
| **Species** | H. sapiens (human), R. norvegicus (rat), M. musculus (mouse), G. gallus (chicken), D. rerio (zebrafish), D. melanogaster (fly), C.elegans (worm) , S. cerevisiae (baker's yeast), A. thaliana (thale cress), O. sativa (rice) | A. thaliana, O. sativa, P. trichocarpa (poplar), G. max (soybean), T. aestivum (wheat), H. vulgare (barley), V. vinifera (grape), Z. mays (maize) | A. thaliana, O. sativa, M.truncatula - M. sativa (Medicago), P. trichocarpa, G. max, T. aestivum, H. vulgare | Z. mays, O. sativa | A. thaliana, O. sativa |
| **Source of microarray data (1)** | GEO | GEO, ArrayExpress | GEO, ArrayExpress | GEO | GEO |
| **Sample bias filtering** | no | no | yes | no | yes |
| **Filtering low-quality samples** | no | no | yes(deleted residuals) | yes (R/arrayQualityMetrics) | no |
| **Microarray normalization (2)** | custom-made CDF + RMA | MAS5 | RMA | RMA | custom-made CDF + RMA |
| **Primary co-expression measure (3)** | PCC | cosine correlation coefficient | Highest Reciprocal Rank (based on PCC) | PCC | PCC |
| **Clustering algorithm (4)** | gene-centric | Confeito algorithm extracting highly interconnected sub-graphs | graph-based (NVN, HCCA) | graph-based (WGCNA, RMT) | gene-centric |
| **Gene homology detection** | NCBI HomoloGene | Best hit orthologous gene (BLASTn) | PFAM | Reciprocal Best Hits | OrthoMCL |
| **Cross-species expression analysis** | filtering homology links between co-expression clusters | list of co-expressed genes in other species based on individual query gene | filtering and quantification homology links between co-expression clusters | network alignment (mixed co-expression topology and homology; IsoRankN) | filtering and quantification homology links between co-expression clusters |
| **Statistical model (5)** | no | no | permutation test | no | permutation test |
| **Bio-classification, functional annotation** | GO (terms linked to AMIGO), Entrez ID, interaction data (protein, DNA, RNA) | GO (Biological Process), KEGG PATHWAYS, KaPPA-View 4, and biological processes of Gene Ontology | MapMan, phenotype | GO, InterPro, KEGG, phenotype | GO, Reactome, MapMan |
| **Functional enrichment analysis** | hypergeometric distribution + Bonferroni correction | no | fisher exact test + Benjamini-Hochberg correction | fisher exact test | hypergeometric distribution + Benjamini-Hochberg correction |
| **Reference** | (Jupiter *et al.*, 2009) | (Ogata *et al.*, 2010) | (Mutwil *et al.*, 2011) | (Ficklin & Feltus, 2011) | (Movahedi *et al.*, 2011) |
| **Algorithm available (6)** | no | no | yes | no | no |

| | | | | | |
|---|---|---|---|---|---|
| **Website cross-species co-expression clusters** | http://vanburenlab.medicine.tamhsc.edu/starnet2.html | http://webs2.kazusa.or.jp/kagiana/cop0911/ | http://aranet.mpimp-golm.mpg.de/ | not available | not available |
| **Visualization (7)** | Graphviz | SVG | Graphviz | | Cytoscape |
| **Comment** | HeatSeeker cross-species analysis using color maps | | meta-network of co-expression clusters | comparison of functional enrichments between co-expression clusters using Kappa | integration data about tissue specificity, protein evolution (Ka) and promoter cis-regulatory elements |

1) GEO: Gene Expression Omnibus
2) RMA: Robust Multichip Average; CDF: Chip Description File; MAS: Affymetrix Micorarray Suite
3) PCC: Pearson Correlation Coefficient
4) NVN: node vicinity network; HCCA: heuristic cluster chiseling algorithm; WGCNA: weighted correlation network analysis; RMT: random matrix theory
5) ECC includes the construction of a null model controlling for network connectivity or tissue specific expression
6) PLANET: http://aranet.mpimp-golm.mpg.de/download/
7) SVG: Scalable Vector Graphics

**Figure legends**

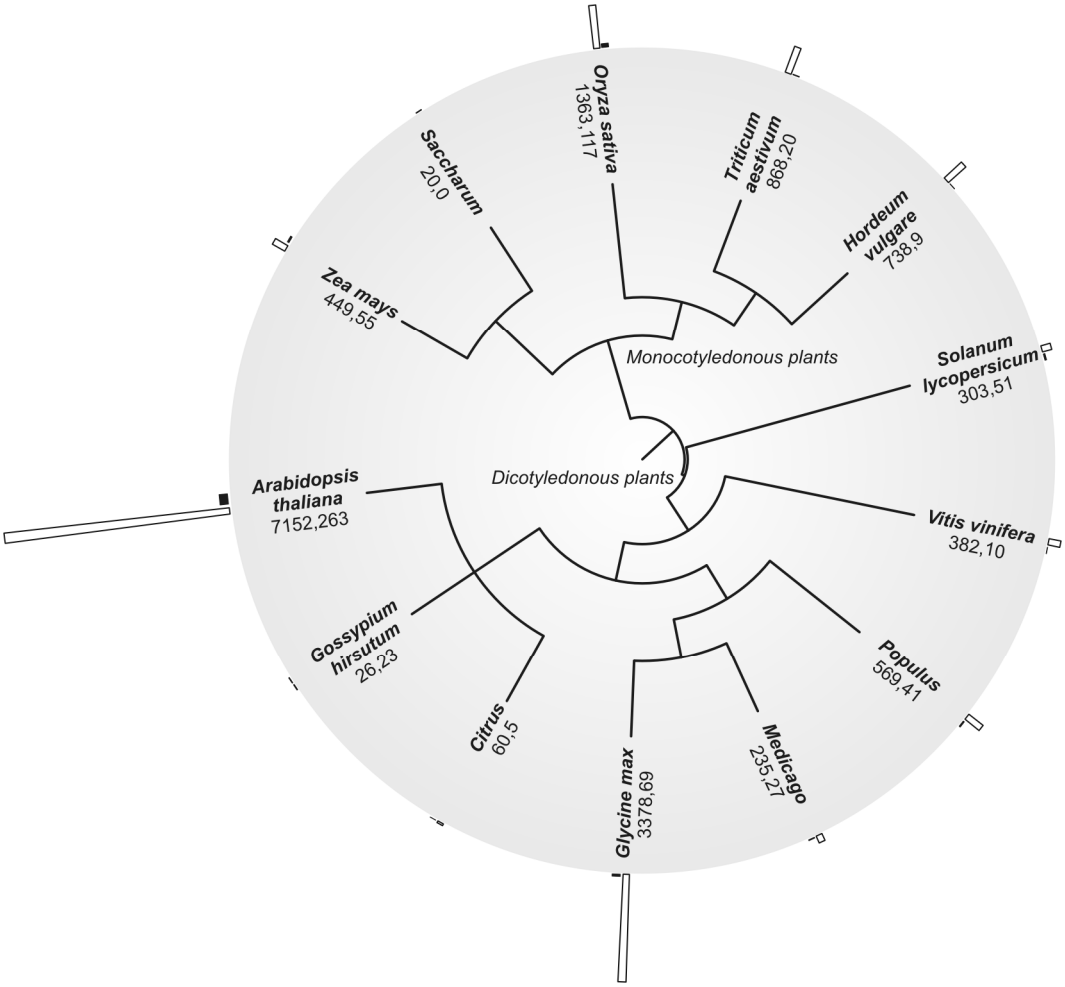



**Figure 1. Overview of publicly available expression data for different plant species.** White and black bars indicate for each species the number of Affymetrix GeneChip microarray experiments (CEL files) in the NCBI Gene Expression Omnibus database and the number of Transcriptome experiments from the NCBI Short Read Archive (SRA), respectively. Values below the species name indicate the number of available CEL files and Transcriptome SRA experiments (November 2011), respectively.

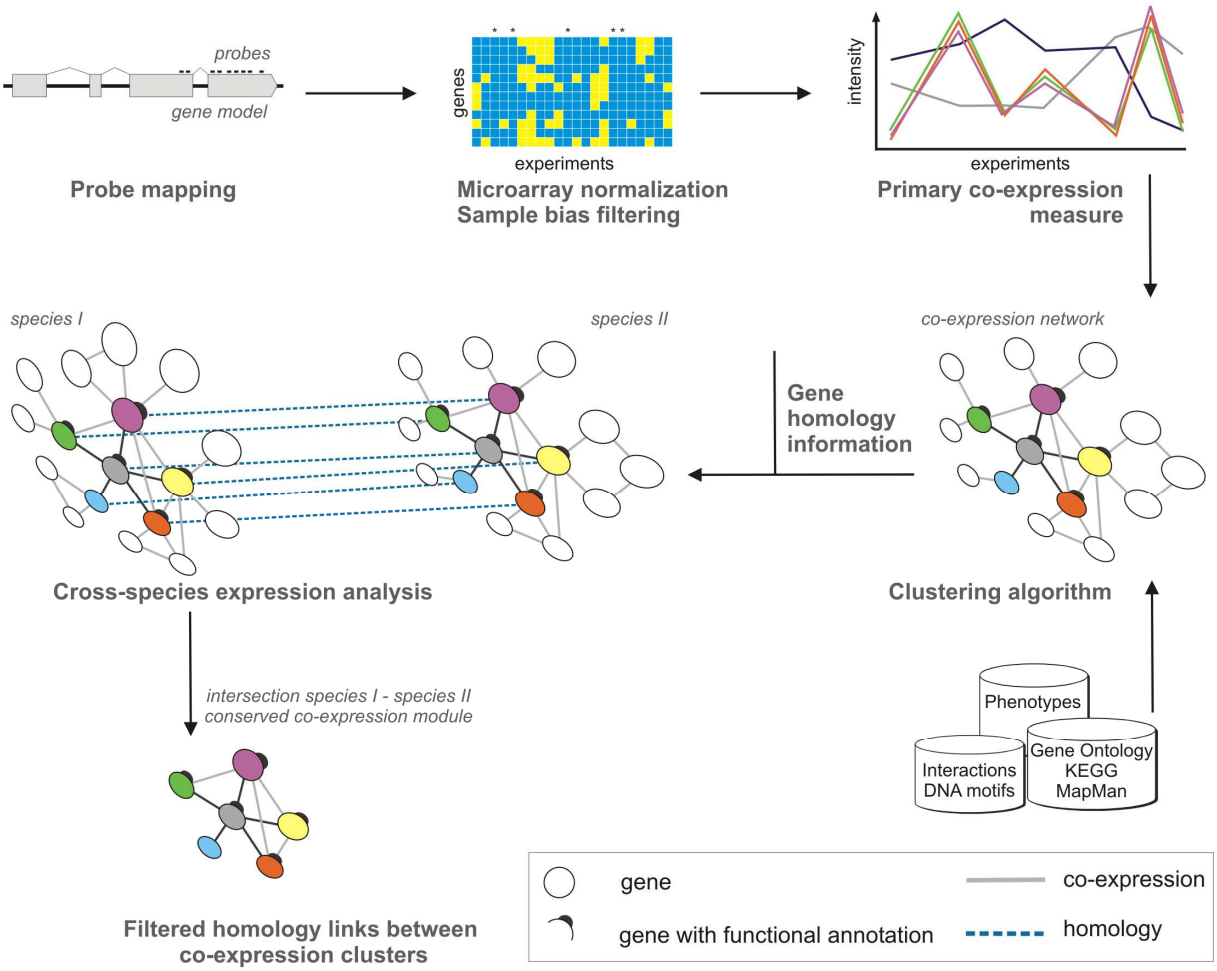
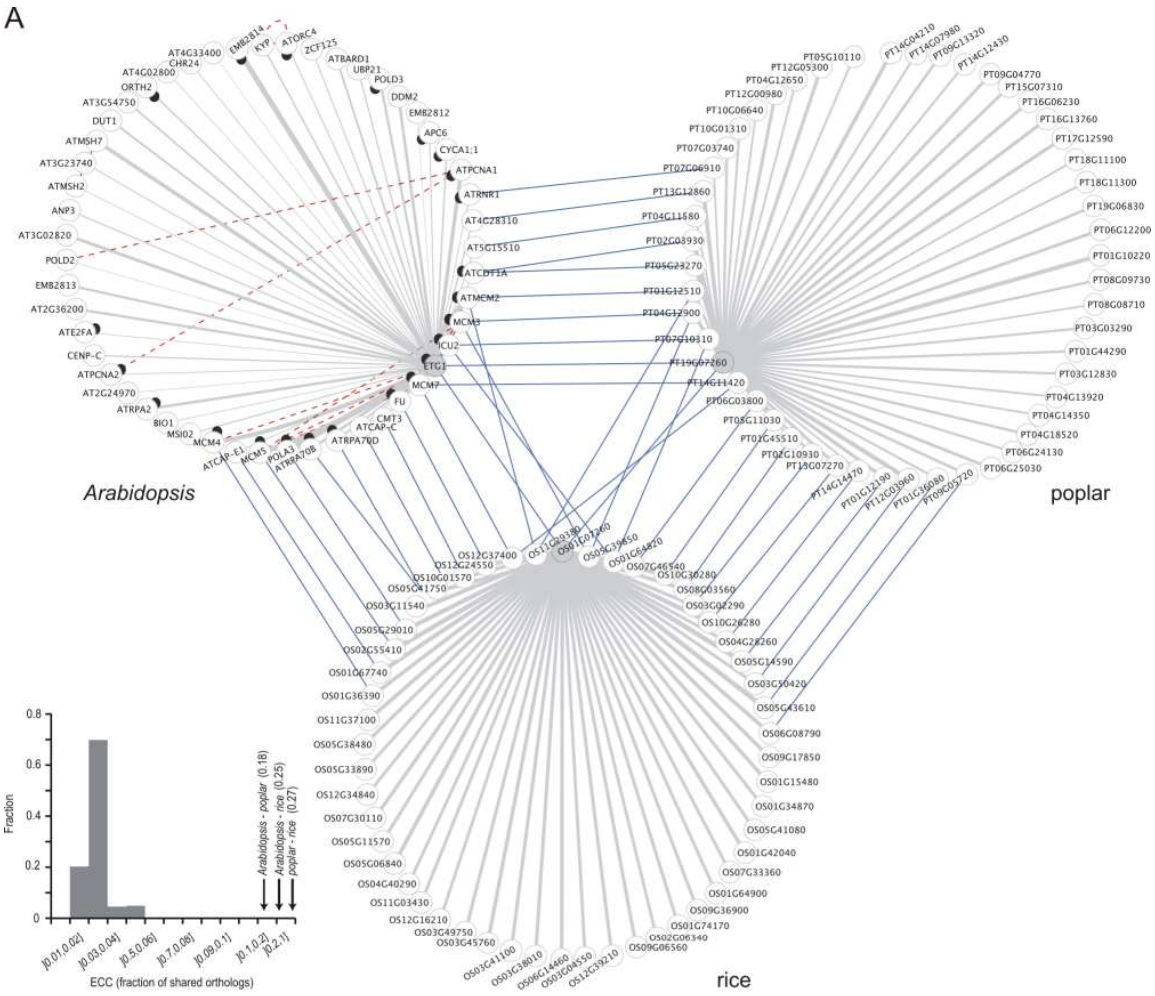



**Figure 2. Workflow for cross-species expression network analysis.** Asterisk above the gene-experiment matrix indicate potentially redundant experiments which can cause a sample bias when computing gene expression similarities. In the co-expression graph circles denote genes while lines indicate expression similarity. Black co-expression lines indicate the first neighbors of the gray query gene (gene-centric cluster) while gray co-expression lines indicate the indirect neighbors (extended node vicinity). Blue lines indicate homologous gene relationships which, when superimposed on the co-expression networks, indicate conserved gene modules.


A
Arabidopsis
poplar
rice
Fraction
ECC (fraction of shared orthologs)
Arabidopsis - poplar (0.18)
Arabidopsis - rice (0.25)
poplar - rice (0.27)

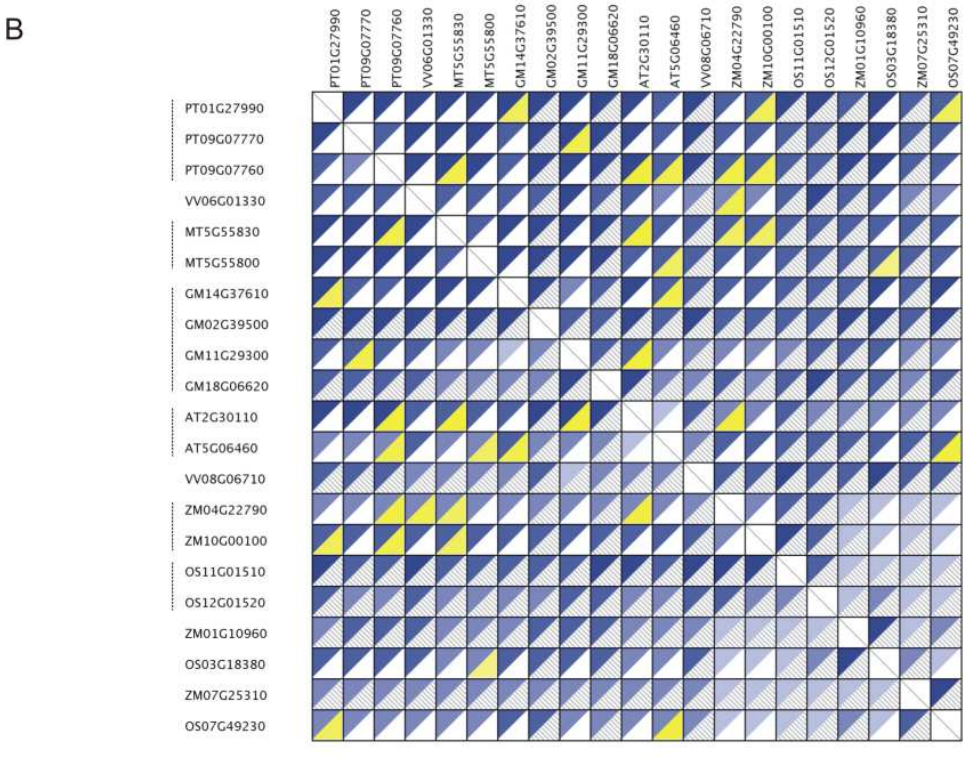

B
PT01G27990
PT09G07770
PT09G07760
VV06G01330
MT5G55830
MT5G55800
GM14G37610
GM02G39500
GM11G29300
GM18G06620
AT2G30110
AT5G06460
VV08G06710
ZM04G22790
ZM10G00100
OS11G01510
OS12G01520
ZM01G10960
OS03G18380
ZM07G25310
OS07G49230

**Figure 3. Plant orthologs with conserved co-expression.** (A) Co-expression context analysis for the Arabidopsis ETG1 gene and its orthologs in poplar and rice (based on PLAZA 2.0 annotations). Grey edges represent co-expression links between ETG1 (query gene) and its top 50 coexpressed genes, weighted by the PCC value. Red dashed edges denote protein-protein interactions, black add-ons are used to indicate genes with known GO annotations for cell cycle and/or DNA replication, and blue edges depict orthology. The inset displays a histogram of the ECC background model (expected number of shared orthologs for random clusters with equal sizes as real co-expression clusters) while the arrows indicate the ECC scores for the different ETG1 co-expression context comparisons. (B) Systematic evaluation of orthology and conserved co-expression using the ECC method for a set of 21 homologs (encoding ubiquitin-activating enzyme E1) from *Arabidopsis*, grape, Medicago, maize, poplar, rice and soybean (AT, VV, MT, ZM, PT, OS and GM prefixes, respectively). Groups of inparalogous genes are indicated using dashed vertical lines. Upper-left triangles denote the sequence-based orthologous relationship between the genes, with a darker shade of blue indicating a higher number of evidence types reported by the PLAZA 2.0 Integrative Orthology approach. The lower-right yellow triangles denote gene pairs with significant ECC scores (p-value < 0.05), white triangles represent gene pairs lacking a significant number of hared orthologs (p-value ≥0.05) and darker shades of yellow indicate a higher fraction of shared orthologs. Arced sections denote missing expression data for at least one of the genes. ECC scores are only computed between genes from different species.