Ghent University
Faculty of Sciences

Department of Plant Biotechnology and Bioinformatics
VIB Department of Plant Systems Biology

Comparative transcriptomics in plants

Sara Movahedi

June 2012

Promoter: Prof. Dr. Yves Van de Peer
Co-Promoter: Prof. Dr. Klaas Vandepoele

Dissertation submitted in fulfillment of the requirements for the degree of
Doctor (PhD) in Sciences, Bioinformatics

Voorzitter: Prof. Dr. Geert De Jaeger (WE09)
Secretaris: Prof. Dr. Yves Van De Peer (WE09) (promoter)

Leescommissie (ZAP UGent):
Prof. Dr. ir. Marnik Vuylsteke (WE09)
Prof. Dr. ir. Tim De Meyer (BW14)
Prof. Dr. Kathleen Marchal (WE09)

Leescommissie (niet ZAP UGent):
Dr. Stefanie De Bodt (WE09)
Dr. ir. Tim Van den Bulcke (UZ Antwerpen)

Overige leden examencommissie (ZAP UGent):
Prof. Dr. Klaas Vandepoele (WE09) (co-promoter)
Prof. Dr. Joke Vandesompele (GE02)

Day of the close defense: 12/01/2012
Day of the public defense: 08/06/2012

Abstract

Comparative genomics is the study of the structural and functional relationships between the genomes of different species or strains. Recently microarray experiments have yielded massive amounts of expression information for many genes under various conditions or in different tissues for different model species. Expression compendia grouping multiple microarray experiments performed in similar (or different) experimental condition make it possible to define correlated expression patterns between genes. Genes within such a coexpression cluster are expected to have more similar functionality compared to genes lacking expression similarity.

In this thesis the different steps required to systematically compare expression data across species are described and some future applications of plant comparative transcriptomics are highlighted. Then we analyzed if functionally related genes show coexpression in Arabidopsis and rice and developed a general framework to measure expression context conservation (ECC) for orthologous genes. Additionally, we studied the evolutionary parameters influencing ECC conservation and compared expression with sequence evolution. At the end, a new method is presented to define high quality tissue specific genes in seven different plant species; *A.thaliana* (Arabidopsis), *Z.mays* (Maize), *M.truncatula* (Medicago), *P.trichocarpa* (Poplar), *O.sativa* (Rice), *G.max* (Soybean) and *V.vinifera* (Grape) using Affymetrix microarray expression profiles. We also performed an in-depth study on the relationship between leaf tissue specific genes coexpression clusters, within a species and in comparison with other species for a set of strictly selected genes.

To my wonderful parents, *Parvin* and *MohammadMehdi*.

Acknowledgements

It gives me great pleasure in acknowledging the support and help of my promoter Yves Van de Peer. I consider it an honor to have the chance to work with him and his research group at UGent. I also acknowledge my co-promoter Klaas Vandepoele with immense gratitude. His energy, insights, and enthusiasm are inspiring. This thesis would not have been possible without his abundant help and invaluable assistance, support and guidance.

Becides, I would like to thank my thesis committee: Geert De Jaeger, Marnik Vuylsteke, Tim De Meyer, Kathleen Marchal, Tim Van den Bulcke, Joke Vandesompele, for their interesting discussions and insightful comments. Thanks specially to Stefanie De Bodt who has been extremely warm, welcoming and willing to help. We had lots of fruitful discussions during my PhD. I would also like to thank the people who first introduced me to the beauty of Bioinformatics: Ali Masoudi-Nejad, Ruy Juregui, Hayedeh Ahrabian and Abbas Nowzari-Dalini.

In my daily work I have been blessed with a friendly and cheerful group of fellow students. My table colleagues Sebastian Proost, Elisabeth Wischnitzki, Bram Slabbinck, Tine Blomme, Cindy Martens, Aminael Sanchez helped me truly enjoy my time at work. I had very nice collaborations with Michiel Van Bel. Lieven Sterck often made sure none of us starved and Yao-Cheng Lin made sure that we have fresh coffee all day long. Pierre Rouz was always encouraging, caring and uplifting others. Vanessa Vermeirssen, Stephane Rombauts, Kevin Vanneste, Phuong Thi Ngoc Cao, Pedro Pattyn, Bing Li, Krzysiek Wabnik, Ward Blonde, Erick Antezana, Thomas Abeel, Riet De Smet, Nathalie Pochet, Anagha Joshi, Grigoris Amoutzias, Eric Bonnet, Jeffrey Fawcett, Jonathan Gordon, Steven Maere, Jayson Gutirrez, Rahul Bhosale, Alex Gryzlov, Steven Robbens, Cedric

Simillion, Xin-Ying Ren and all my other colleagues at the Bioinformatics group in VIB, I spent lots of great moments full of joy and happiness with you all. A special thank of mine goes to Naira Naouar, Tine Casneuf, and my teachers Kristoffel Boudens and Masih Ghaffari for generously sharing their knowledge with me.

Niloufer Irani, Amandine Radziejwoski, Nubia Eloy, Megan Ladd, Andry Andriankaja, Kristiina Himanen, Barbara Berckmans, Indrani Attiganal and Joost Altink thanks for the bright lights of your friendship. Dora Szakonyi, Yunlong Du, Pierre Hilson, Nino Villarroel, I never forget all those friendly chats and long discussions at the afternoon coffee breaks. Wilson Ardiles you always treated me and other Iranians as your special friends. I will always keep your kindness in my heart. My dearest Mansour Karimi, your generous support and welcoming style made the big move from Iran to Belgium a truly enjoyable experience for me.

I faced a serious health issue during the second and third year of my study. It was not possible to survive those difficult days without the love and support of my wonderful friends. Sarieh Ghorbani, Saman soltani, Mahsa vahdat, Soren Seifi, Nima Navidi, Maryam Hakim Hashemi, Farzaneh Ghazavi, Behrouz Hassannia, Sepehr Baratiniko, Elnaz Karimian, Setareh Gorji, khosro Mehdikhanlou, Nafiseh Lashgari, Adel Pezeshki, Farzaneh Razavi, Mohammad Miranzadeh, Nasim Navidi, Reza Karimi, HamidReza Karimian, Parisa Kord Jamshidi, Shahrouz Mirzaalizadeh, Soheil Radiom, Mohammad Moradzadeh, Fereshteh Bozorgi, Reza Khoshsiar and all my other Iranian friends in Belgium, I cannot find words to express my appreciation to your sincere support and attention. Besides, I thank you for all those wonderful moments we spent together: trips, campings, parties, etc. I will miss you all.

Dear friend Zeynab Biglaripour, thank you so much for the excellent design of my thesis book cover. I like it very much and I will recommend you to everyone.

My lovely friends Pooneh Kalhorzadeh, Morvarid Farhang Ghahremani, Leila Kheibarshekan, Maryam Atef Yekta, Moones Harandy, Shirin Nasr-Isfahani, Solmaz Shahalizadeh, Samira Jamnejad, Sandra Botelho, Evan-

gelia Dougali, and Ying He thanks for being such a beautiful part of my life, your friendship means more to me than you will ever know. Thank you for all the moments we have shared, moments filled with shared dreams and wishes, secrets, laughter, and tears, and above all, friendship. Each precious second will be treasured in my heart forever.

Thanks to my colleagues at Rijk Zwaan, Remco Ursem, Rob Dirks, Marcel Kerkveld, Rao Xiangyu, Martijn van Elk, Raoul Frijters, Aat Vogelaar, Evert Gutteling, Taco Jesse, Tineke Sonneveld, Kees van Dun, Dilly van Noort, Han de Rooij, Jasper de Joode, Roy Koopmans, Ruud van Hof, Silvester de Nooijer and many others for their kind support as I was learning my new job while at the same time finalizing my thesis.

A warm thank you goes out to my beloved family, Parvin Adabi, MohammadMehdi Movahedi, MohammadAli Movahedi, Mitra Movahedi, Amirhossein Mirafzal, Nassim Nassirinia and Asal Mirafzal for always encouraging and believing in me. Their love and support has been constant and invaluable.

Finally, I would like to thank everybody who has directly or indirectly contributed to the success of this thesis; I express my sincere apology that I could not mention them personally.

Sara Movahedi
28 May 2012

This project was funded by the Research Foundation-Flanders and the Belgian Federal Science Policy Office: IUAP P6/25 (BioMaGNet).

Contents

List of Figures

LIST OF FIGURES

List of Tables

LIST OF TABLES

List of Abbreviations

CAST Cluster Affinity Search Technique

CDF Chip Description File

CoP Coexpressed biological Processes

DE Differentially Expressed

EC Expression Coherence

ECC Expression Context Conservation

GEO Gene Expression Omnibus

GO Gene Ontology

GOMO Gene Ontology for Motifs

HCCA Heuristic Cluster Chiseling Algorithm

HRR Highest Reciprocal Rank

JI Jaccard Index

MAS Affymetrix Micorarray Suite

NGS Next Generation Sequencing

NVN Node Vicinity Network

PCC Pearson Correlation Coefficient

QC Quality Control

LIST OF TABLES

RMA Robust Multichip Average

RMT Random Matrix Theory

SNP Single-Nucleotide Polymorphisms

SVG Scalable Vector Graphics

TF Transcription Factor

WGCNA Weighted Correlation Network Analysis

WTSS Whole Transcriptome Shotgun Sequencing

1

Introduction

1.1 Chapter overview

The goal of this chapter it to provide an introduction to basic knowledge necessary for better understanding of the thesis. It begins with a brief overview of the field of bioinformatics. Afterward, the main subject of this thesis, comparative analysis of gene (co-)expression data, is briefly introduced and essential related concepts are explained. The chapter is finalized with an introductory to microarray technologies and a description on key principles of designing a custom made CDF package.

1.2 Bioinformatics: An introduction

Bioinformatics is the application of computer science techniques to solve biological problems. Our genetic identity is encoded in long molecules called DNA. DNA codes for genes which code for proteins which determine the biological makeup of any living organism. It is important to understand the variation of DNA molecule and its errors which can cause disease and biological disorders. Bioinformatics provides necessary tools and methods to uncover the wealth of biological information hidden in sequence, structure, literature and other biological data. Bioinformatics deals with algorithms, databases, information and computation theory, software engineering, data mining, image processing, modeling and simulation, statistics and many other concepts in order to improve knowledge of biology and medicine by management and analysis of biological data.

Biological databases

Biological databases consist of effectively stored biological data in various areas. Biological databases can be grouped into two main categories; primary or archived databases such as information and annotation of DNA and protein sequences, DNA and protein structures and DNA and protein expression profiles and secondary databases which are based on results of analysis on the primary resources, for instance information on sequence patterns or motifs, variants and mutations and evolutionary relationships. EMBL database for nucleotide sequence data, UniProtKB/Swiss-Prot protein database and PDBe a 3D protein structure database are examples of a few popular biological databases.

Biological applications

Biological databases can not provide meaningful information for biologists unless their data is analyzed by bioinformatic tools and software programs. Nowadays a wide range of biological data analysis tools are publicly available. For instance sequence analysis tools, protein function analysis tools, gene expression analysis tools and RNA-seq data analysis tools. Fields of Molecular medicine, Microbial genomics, Agriculture, Animals and Comparative studies are only a few scientific areas which take advantage of bioinformatics applications (1, 2, 3).

1.3 Comparative genomics: from sequence to expression

1.3.1 Introduction

Comparative genomics is the study of the similarities and differences of genome structure and function across various biological species. The goal of comparative genomics is to understand the evolutionary processes of genomes, find new genes, determine function of genes, proteins or regulatory networks, and to discover non-coding functional elements of the genomes. The objective of this study is ranging from whole genome comparisons to gene expression analysis. The raw data used for comparative genomics analysis can be produced in-house or retrieved from public databases such as NCBI which provides very wide range of biological data like Genomes, Genes, Proteins, SNPs, ESTs, Taxonomy, etc.

Comparative sequence analysis provides valuable information about the functional parts

encoded by a genome, mainly by exploring the conserved DNA sequences that code for proteins or RNAs or regulatory elements (4). Inter-species comparisons have two major applications: conservation between species helps to detect and characterize functional elements whereas differences can reveal biological adaptations linking genotype with phenotype (5). Recently, the increase in functional genomics data has transformed comparative approaches from basic sequence analysis to detailed studies of functional attributes such as gene expression or protein-protein interactions (6, 7). For instance, microarray experiments have yielded large amounts of genome-wide expression information under various conditions or in different tissues for several model species.

1.3.2 Tools and Concepts

This section introduces essential tools and concepts needed to understand the different aspects of comparative coexpression analysis which is the main subject of this thesis.

Gene expression, coexpression, coexpression network

Gene expression is the process by which information from a protein-coding gene is converted into mRNA via transcription and then to protein via translation. Non-protein coding genes are involved only in transcription and not translation, and the product is a functional RNA such as ribosomal RNA (rRNA) or transfer RNA (tRNA) (8).

Genes with similar expression patterns in a relatively large percentage of samples are defined as coexpressed genes. Two of the most popular methods to measure coexpression of a pair of gene are Pearson Correlation Coefficient (PCC) and Euclidean distance. There are many different methods which group coexpressed genes and make coexpression clusters, for instance k-means algorithm (9), the Cluster Affinity Search Technique CAST (10), the Confeito algorithm (11), Weighted Gene Coexpression Network Analysis (WGCNA) (12) and Heuristic Cluster Chiseling Algorithm (HCCA) (13).

The pairwise relation of highly correlated gene expression profiles across microarray samples (i.e. coexpressed genes) can be described in a coexpression network which correlated genes (nodes) are connected to each other through an undirected edge with the weight of distance value (e.g. PCC value).

Gene coexpression relationship can only reflects mRNA-level regulation and not the protein-level regulation therefore to predict the gene functional interactions many studies integrate gene coexpression information with protein-protein interaction and gene

functional annotation. The incorporation of detailed information on aforesaid concepts and gene homology between species can simplify the process of identifying conserved gene modules and translating biological knowledge from model organisms to crops (14, 15).

Homology, orthology and paralogy

Two genes with a common ancestral DNA sequence are called homolog. In 1970, Walter Fitch (16) defined the concepts of orthology and paralogy to distinguish between two types of homology relationships. Orthologous genes are separated by a speciation event from their common ancestor, whereas paralogous genes are separated by the event of genetic duplication (17). There are different approaches available to identify homologous and orthologous genes (18) which most of them start from the output of a global all-against-all sequence similarity search. OrthoMCL and Inparanoid (19, 20) are two advanced methods to construct orthologous groups across genomes.

Functional annotation, Gene Ontology (GO), Plant Ontology (PO)

Gene functional annotations provide biological information for a gene. The Gene Ontology (GO) (21) project provides a controlled vocabulary of defined terms representing gene product properties. Similarly, Plant Ontology (PO) (22) provides an ontology of plant structures and growth stages.

There are many tools and softwares publicly available which use GO for different purposes. For instance Gene Ontology for Motifs (GOMO) is a comparative genomics tool for assigning functional roles to DNA regulatory motifs from DNA sequence (23, 24) or Blast2GO (B2G) joins similarity search based GO annotation and functional analysis in one universal application (25). PO terms are being used by databases such as TAIR (26), NASC (27), Gramene (28, 29), SGN (30) and MaizeGDB (31) to describe expression patterns of genes and the phenotypes of mutants and natural variants.

Protein-protein interactions

Protein-protein interactions provide information on binding two or more proteins together which are often involved in a similar biological function (32). PlaPID a database of protein-protein interactions in plants (33), the MIPS mammalian protein-protein interaction database (34), NIA mouse protein-protein interaction database (35), DIP

the database of interacting proteins (36) and INTERACT an object oriented protein-protein interaction database (37) are a few available protein-protein interaction databases. There are also many softwares available which interrogate protein-protein interaction databases, for instance CORNET (38) PPI tool investigates protein-protein interaction databases for any Arabidopsis gene query.

Tissue specificity, expression breadth, Tau (τ)

Expression breadth is the number of tissues (or experiments) where a gene is expressed. Tissue specific genes are highly expressed only in a limited number of tissues (39). Some studies define tissue specific genes by applying a threshold on expression breadth and some use a more accurate method such as Tau (τ) (40) (see section 4.5.7).

Guilt by association

Guilt by association principle for analyzing gene networks states that genes which are associated or interacting are more likely to share function. The combination of guilt by association and gene co-expression analysis together with other techniques enables us to assign roles to previously uncharacterized genes (15, 41).

1.3.3 Comparative analysis of (co-)expression networks

Expression compendia grouping multiple microarray experiments make it possible to define correlated expression patterns between genes (42, 43). Genes within a coexpression cluster are expected to have more similar functionality than those without expression similarity. In other words, functionally related genes are often coexpressed, both in closely and more distantly related species (6, 44). In one of the first gene coexpression network studies, in 2003, Stuart et al. studied 3182 DNA microarrays from humans, flies, worms, and yeast and defined 22,163 pairs of genes with conserved coexpression pattern. They believe many of these cases provide strong evidence for evolution of new genes in core biological functions and experimentally confirmed some of the predictions implied by these links (6).

To compare expression data between different species, microarrays can be processed in three different manners (45). One approach combines multiple microarray experiments to identify differentially expressed genes in each species independently and then compare these genes among different species (46). This method requires for each species

orthology information to link gene expression states across species and can be applied to both closely or more distantly related species. Another approach hybridizes samples from different but closely related species to the same microarray and requires similar experimental conditions as well as orthology information. Finally, separate arrays can be used to sample similar experimental conditions for different species and all of them are analyzed together to investigate expression evolution of orthologs, paralogs, or specific functional categories (45). The latter approach was used, for example, to investigate species-specific gene duplications in human and mouse (47). In addition to comparison of expression profiles between orthologous genes, some studies also try to combine expression information with sequence evolution and gene function. Bergmann et al. integrated expression data of six species with their genomic sequence information to identify coexpression conservation and to improve functional gene annotation. Based on graph theory, transcriptional networks were inferred revealing that highly connected genes are often conserved and have essential functional rules (44). Sets of genes that are conserved at both sequence and expression levels among multiple species are expected to play a key role in biological responses (6, 45).

Although comparative expression analysis is most straightforward when compatible expression data sets are used that cover equivalent conditions for all species, in this approach only a small fraction of all available data in different species can be utilized (5). Furthermore, how similar gene expression is modulated in distantly related species is still not understood, even when considering compatible conditions. To overcome these limitations, more advanced transcriptomics studies have shown that comparing coexpression, instead of the raw expression values, provides a valid alternative to identify regulatory modules and study their evolution (6). For instance, gene expression between the four distantly related species *Caenorhabditis elegans*, *Drosophila melanogaster*, *H. sapiens* and *Saccharomyces cerevisiae* was compared by means of a coexpression meta-analysis (14).

In chapter 2 we extensively study comparative co-expression analysis in plant biology. In this chapter different common steps which are taken to compare expression data between plant species is discussed. We also compare in details a few examples of comparative expression analysis in plants which mostly identify conserved gene expression by comparing gene clusters across species using homologous or orthologous genes. A

number of these studies are Coexpressed biological Processes (CoP) (48), Expression Context Conservation (ECC) (49), Plant Network (PLaNet) (50) and STARNET2 (51).

1.4 Principles of microarray technology

1.4.1 Introduction

Microarray technologies, invented in late 1980s, provide new tools that changes the way scientific experiments are carried out. Microarrays interrogate thousands of genes or entire genomes simultaneously.
Microarrays are 2D array on a solid substrate (usually a glass slide or silicon thin-film cell) that contain an ordered series of samples (DNA, RNA, protein, tissue). The type of microarray depends upon the material placed onto the slide:

- DNA microarrays, such as cDNA microarrays, oligonucleotide microarrays and SNP microarrays
- MMChips, for surveillance of microRNA populations
- Protein microarrays
- Tissue microarrays
- Cellular microarrays (also called transfection microarrays)
- Chemical compound microarrays
- Antibody microarrays
- Carbohydrate arrays (glycoarrays)

In general microarrays are used for differing expression of genes over time between tissues and disease states, identification of complex genetic diseases, drug discovery and toxicology studies, mutation/polymorphism detection (SNPs), pathogen analysis or other purposes.

1.4.2 DNA microarrays (GeneChip)

The most commonly used microarray is the DNA microarray (also commonly known as gene chip, DNA chip, or biochip) which determines the expression levels of genes in a sample, commonly termed expression profiling. Microarray experiment have variable applications. One of most widespread uses of microarrays is to quantify and compare the expression level of genes in different conditions in order to define the function of particular genes in a specific condition such as stress, tissue or chemical environment. A gene can up- or down-regulate, or stay unchanged during a particular condition. One of the most popular applications of microarrays is to assign a specific function to a gene which reacts to an applied condition the same as a number of functionally known genes. By assumption, genes that share common regulatory patterns also share the same function.(52)

In 1998 Eisen et al. (42) introduced the idea of gene coexpression across multiple microarray experiments. Central to the concept of coexpression is the idea that in order for two gene products to work together in the same pathway or process they must be present at the same time. Therefore, it is logical to suppose that two genes that are expressed together under a range of experimental conditions are likely to share a common functionality (53).

1.4.3 Tiling arrays

Tiling array is another technology which was made to detect expression of transcripts or alternatively splice forms which may not have been previously known or predicted. Tiling Arrays are very similar to microarray chips, in the sense that in both cases, labeled target molecules are hybridized to unlabeled probes fixed on to a solid surface. The difference is that tiling arrays consist of short overlapping (or in very close proximity) probes, approximately 25bp, covering the entire genome or contiguous regions of the genome. Therefore, these arrays identify not only novel transcripts, but also their intron-exon boundaries. Three manufacturers of tiling arrays are Affymetrix, NimbleGen and Agilent. Their products vary in probe length and spacing.

As shown on the tree in figure 1.1 (generated using iTOL (54)) there are at the moment Affymetrix microarrays available for 32 different species (last updated on November 2011 (55)), mostly covering plants and animals. Whereas 3 *Gene Expression Analysis*

arrays cover all species, *Tiling arrays* are available for only a few plants, animals and yeasts, and *Exon and Gene Level arrays* are only available for human, mouse and rat.

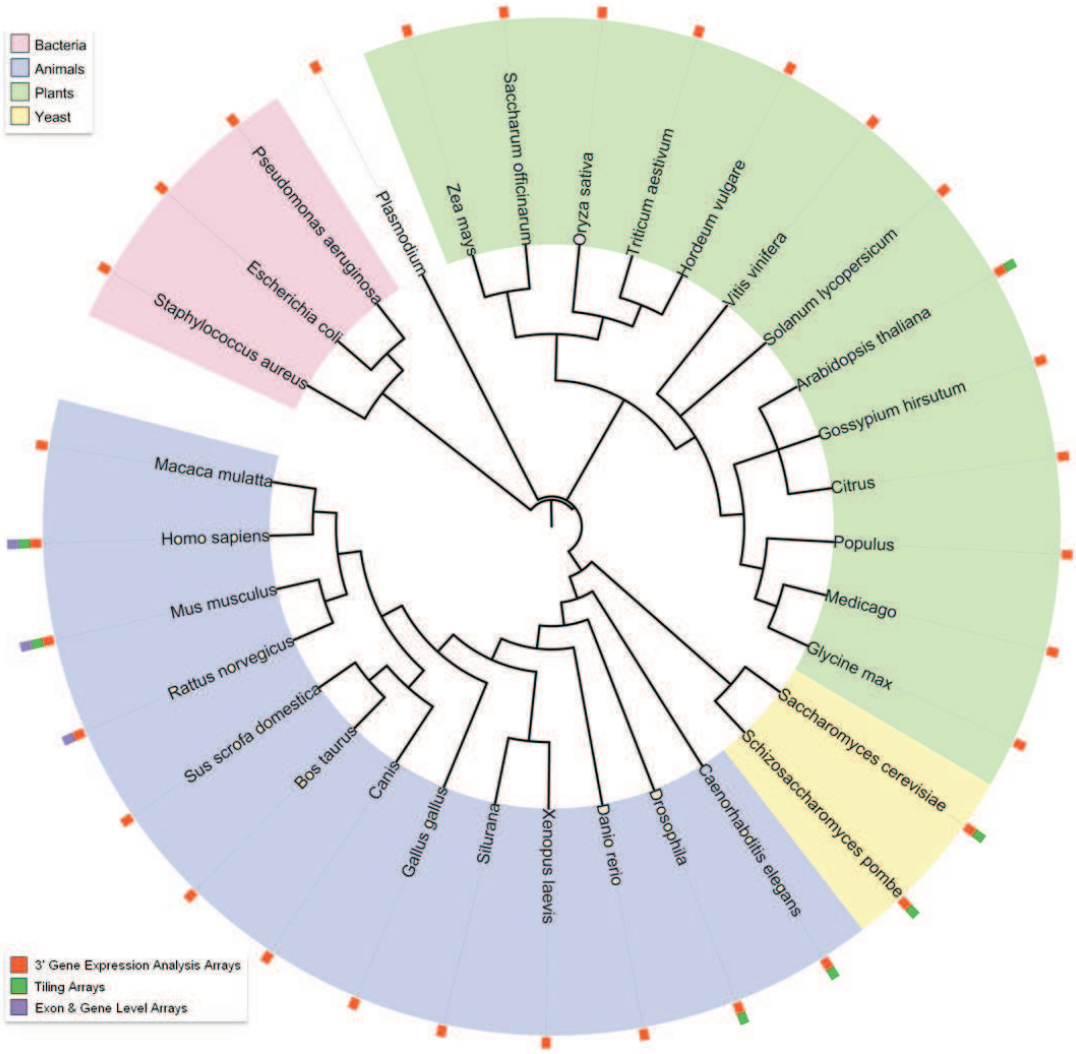


Figure 1.1: Affymetrix platform, species coverage - There are Affymetrix microarrays available for 32 different species. Including plants, animals, bacteria and yeast.

1.4.4 RNA-seq

RNA-Seq also called "Whole Transcriptome Shotgun Sequencing" (WTSS) is a recently developed approach to transcriptome profiling that uses deep-sequencing technologies

and determines the cDNA sequence directly (56). At the moment the Illumina IG, Applied Biosystems SOLiD and Roche 454 Life Science systems have been already applied for the purpose of high-throughput sequencing technology.
RNA-seq is highly accurate for quantifying expression levels and can also be used to sequence the transcriptome of organisms with undetermined genomic sequences. It has very low background signal comparing to DNA microarrays because DNA sequences can be mapped to unique regions of the genome with high certainty.

1.5 Custom made CDF file

A DNA microarray consists of thousands of microscopic spots of DNA oligonucleotides, called features, each containing picomoles (10-12 moles) of a specific DNA sequence, known as probes (or reporters). These can be a short section of a gene or other DNA element that are radioactively labeled or otherwise marked and used to hybridize to a cDNA or cRNA sample (called target) under high-stringency conditions. Hybridization with probes is detected optically (see figure 1.2). In Affymetrix microarrays, the probes (which can be tens of thousands in an array) are attached to a solid surface of silicon chip. Other microarray platforms, such as Illumina, use microscopic beads. Intensity value of each cell (feature) of the resulted image for a specific experiment is saved in a CEL file. GEO website (57) has a big dataset of these files. For each specific experiment, the CEL file needs to be processed via the CDF file in order to have a raw intensity value per gene. Then we need to transform intensity to expression values using algorithms such as MAS5 (Affymetrix proprietary method) or RMA/GCRMA (58). Resulted expression profile is a matrix of genes (rows) and conditions (columns), which is background-corrected, normalized and finally summarized.
Depending on the chip design and the laboratory protocol, microarrays can be used to measure mRNA expression levels, detect changes in genomic regions of interest (resequencing), tile probes over an entire genome for various applications (novel transcripts, ChIP, epigenetic modifications) (Tiling arrays), detect which known SNPs are in the tested DNA (SNP array).
Whereas the GeneChip platform measures the expression of a gene using multiple probes grouped in a probeset[1], the initial probeset definitions provided by Affymetrix

[1] A probe set is a collection of probes designed to interrogate a given sequence.

frequently rely on draft gene annotations and therefore might suffer from cross-hybridization events (i.e. probes matching to different genes). For instant, remapping all 631,066 rice (*Oryza sativa*) probes to the rice gene models (TIGR5) shows that 43% of the original probesets contain one or more cross-hybridizing probes. These cross-hybridization effects have an essential role in imprecise profiling of low-copy-number molecules (59). Since it is for all microarray studies important to ascertain that a probe specifically measures the intended gene/transcript, the inaccuracies present in the original Affymetrix probesets made us to develop a generic software tool to construct high-quality probesets (or Chip Description Files)[1]. This can be done by mapping unique probes at the gene level or at the transcript level (Appendix A). As an example, a new rice gene CDF file was created that now allows to reliably measure expression levels for approximately 75% of all non-transposable rice gene models (using 8 or more probes per probeset).

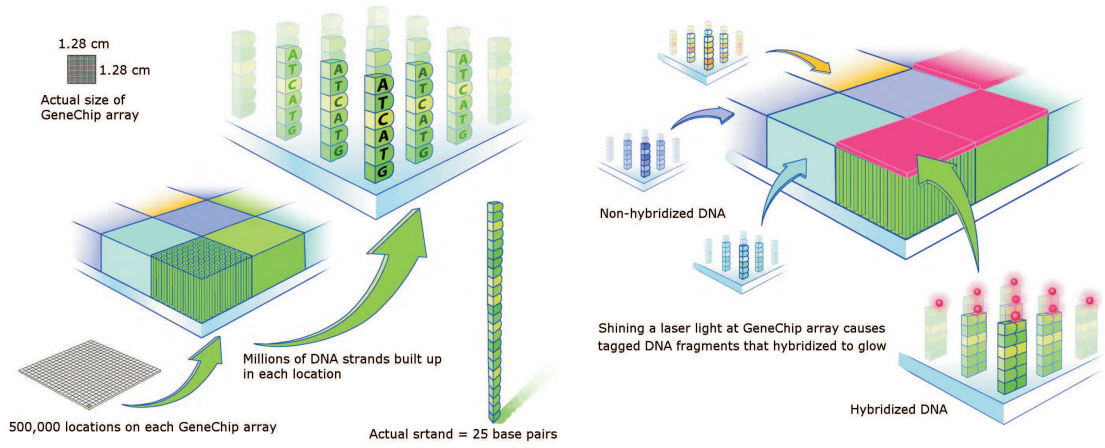


Figure 1.2: The basic construction of an Affymetrix chip - An illustration of Affymetrix chip (source: www.affymetrix.com).

[1]CDF file describes probe locations and probeset groupings on the chip

2

Comparative co-expression analysis in plant biology

Review Article

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

2.1 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.

2.2 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 (4). Accelerated by technological innovations, genome-wide data describing functional properties, such as 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 (5).

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 (45). The hybridization of samples from closely related species to the same microarray requires compatible experimental conditions and has been used in studies comparing different *Brassicaceae* species (60, 61, 62). 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 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 (63). 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 (46). 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 (64, 65). 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 (66).

Although comparative expression analysis is most straightforward when compatible expression data sets are used that cover equivalent conditions for all species, in this approach only a small fraction of all available data in different species can be utilized (5). To overcome these limitations, pioneering comparative transcriptomics studies have shown that comparing coexpression, instead of the raw expression values, provides a valid alternative to identify modules (sets of coexpressed genes potentially sharing similar regulation) and study their evolution (6, 44). 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 (6). Similarly, the integration of genome-wide expression data was used to study the modular architecture of regulatory programs in six evolutionary distant organisms (44).

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.

2.3 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 25 nt, are attached to a solid surface. Other microarray platforms, like Agilent, use only a few but longer probes to measure gene expression (67). 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 standard to facilitate data sharing (68). 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) (69) or ArrayExpress (70) are public microarray archives and provide data of thousands of expression profiling studies (Figure 2.1). 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 (71)) which summarize expression profiles in tens or hundreds of different conditions (72). 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 (58), 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 (73).

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 (74). Although repeat sequences can interfere with the reliable measurement of genome-wide expression, high-density tiling arrays are independent of known gene annotations and

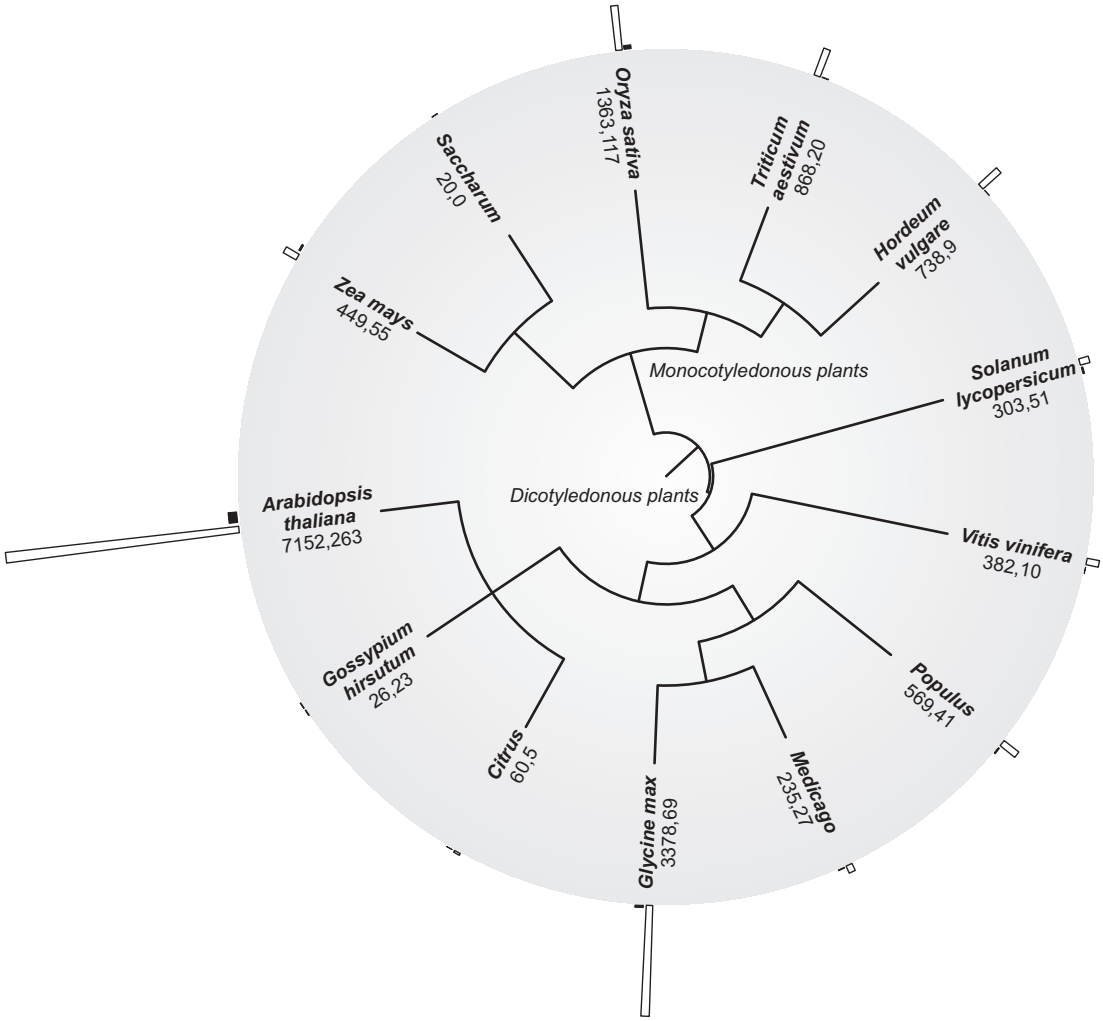


Figure 2.1: Overview of publicly available expression data for different plant species - White and black bars indicate for each species the number of Affymetrix GeneChip microarray experiments (CEL files) in the NCBI Gene Expression Omnibus database and the number of Transcriptome experiments from the 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.

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 2.11).

According to a survey executed on November 2011, there were thirteen Affymetrix GeneChip microarray platforms publicly available in the NCBI GEO for different plant models (eight dicots and five monocots, see Figure 2.1). 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 the plant expression atlas generated for Arabidopsis (75), large-scale expression compendia have been constructed, using a variety of platforms, for other species as well. Examples include barley (*Hordeum vulgare*) (76), *Medicago truncatula* (77), rice (78, 79), tobacco (*Nicotiana tabacum*) (80) and soybean (*Glycine max*) (81). 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 (38, 41). Additional procedures can be applied to remove low-quality samples or to remove samples that could generate biases within the final compendium (Table 2.1). The latter is typically achieved by applying a statistical selection procedure to retain only independent conditions or, reversely, by first grouping similar conditions and retaining only a single experiment as a representative for a set of related microarray conditions (49, 50). 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 (50), missing genes provide an important drawback for many microarray-based co-expression tools (see for example Figure 2.2).

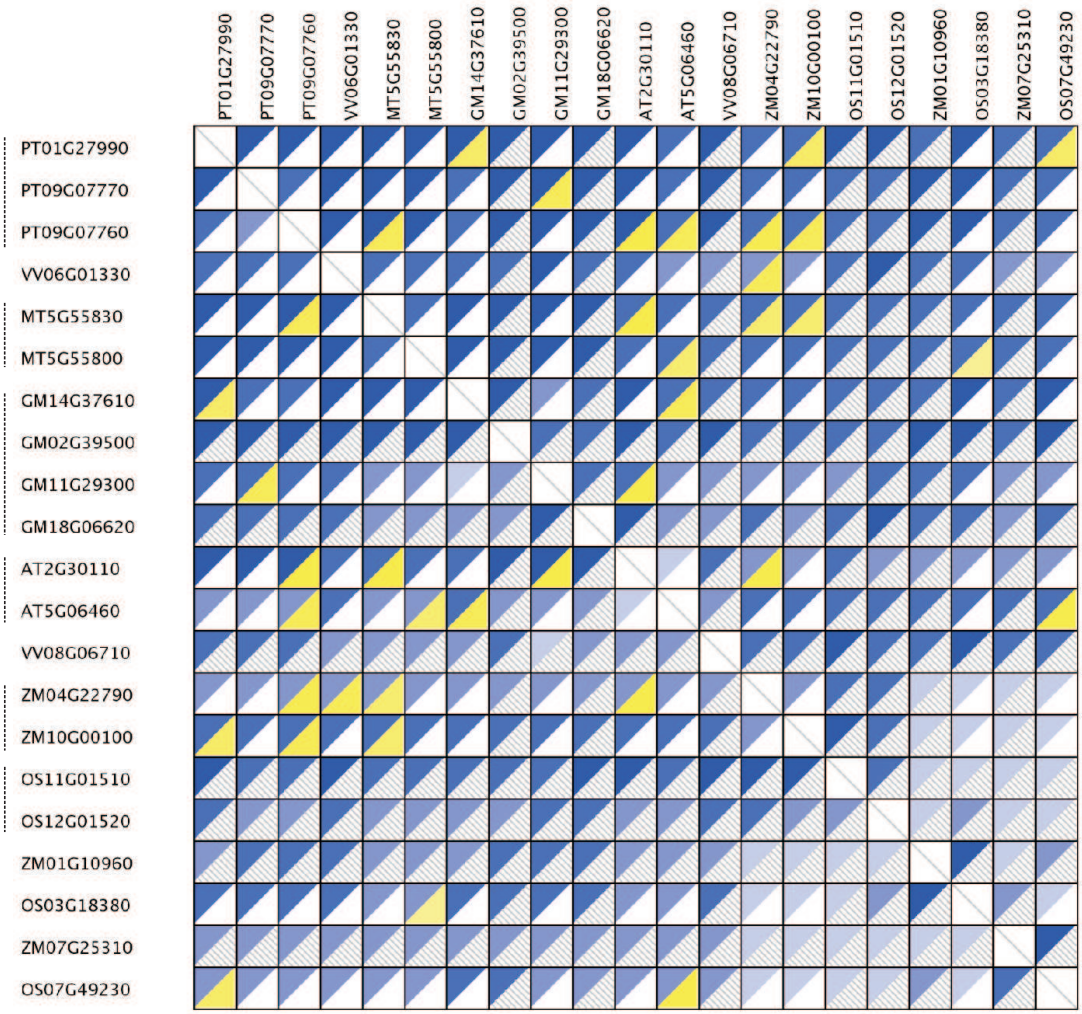


Figure 2.2: Plant orthologs with conserved co-expression (A) - Systematic evaluation of orthology and conserved co-expression using the ECC method for a set of 21 homologs (encoding ubiquitin-activating enzyme E1) from Arabidopsis, grape, Medicago, maize, poplar, rice and soybean (AT, VV, MT, ZM, PT, 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.

	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 coexpression measure (3)	PCC	cosine correlation coefficient	Highest Reciprocal Rank (based on PCC)	PCC	PCC

Clustering algorithm (4)	gene-centric	Confeito algorithm extracting highly interconnected subgraphs	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 coexpression clusters	list of coexpressed genes in other species based on individual query gene	filtering and quantification homology links between coexpression clusters	network alignment (mixed coexpression topology and homology; IsoRankN)	filtering and quantification homology links between coexpression 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	(51)	(48)	(50)	(82)	(49)

Algorithm available (6)	no	no	yes	no	no
Website cross-species coexpression clusters	`http://aranet.mpimp-golm.mpg.de/download/`	`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 coexpression 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

Table 2.1: Overview of cross-species coexpression studies in plants.

2.4 Detection of gene clusters and construction of coexpression 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 2.3). 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 (83). 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 a certain algorithm should be preferred over another (41, 84).

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 coexpression studies are cosine and Spearman's correlation coefficient (Table 2.1). 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 2.1). 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 2.3). 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.

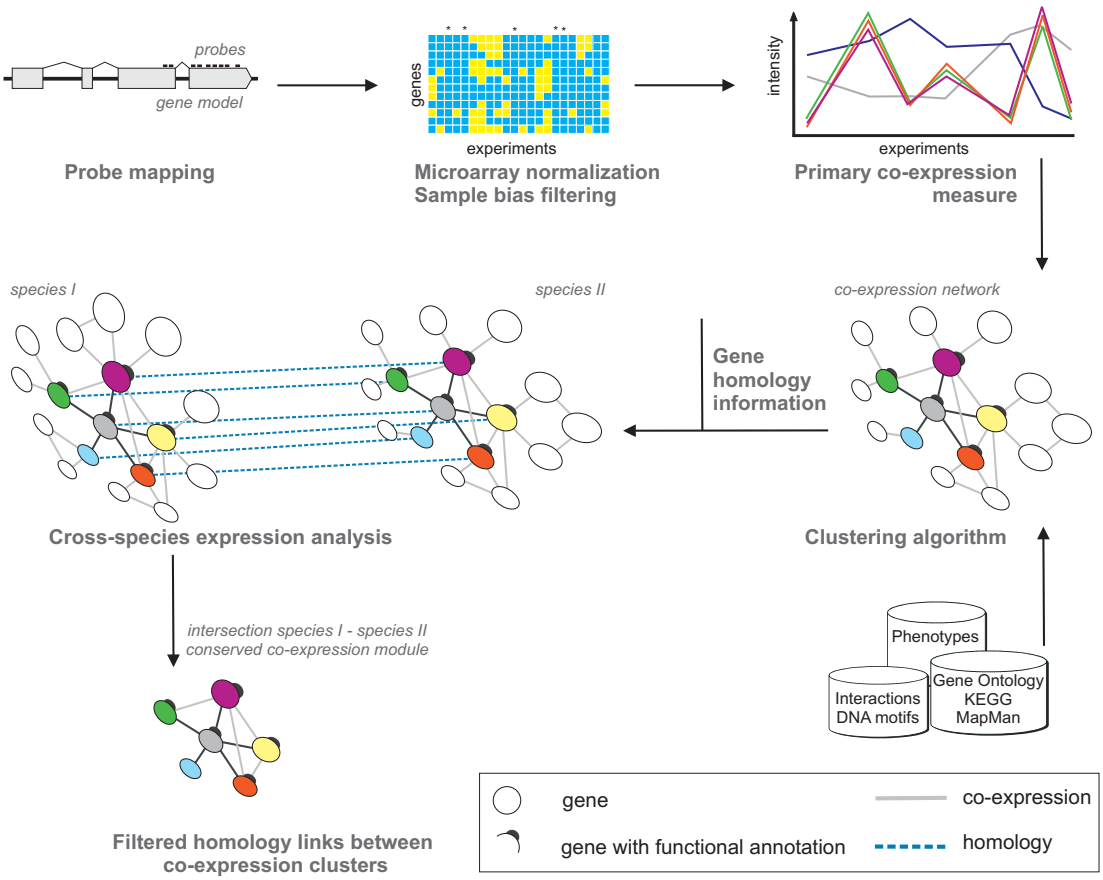


Figure 2.3: Workflow for cross-species expression network analysis - Asterisk above the gene-experiment matrix indicate potentially redundant experiments which can cause a sample bias when computing gene expression similarities. In the co-expression graph circles denote genes while lines indicate expression 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.

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) (10, 85), the Confeito algorithm (11), Weighted Gene Coexpression Network Analysis (WGCNA) (12), Random Matrix Theory (RMT) (86) and Heuristic Cluster Chiseling Algorithm (HCCA) (13) are examples of graph-based algorithms which have been applied for defining gene co-expression clusters in plants (Table 2.1).

2.5 Comparing coexpression networks across species

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 2.1) and to identify genes with conserved gene functions or conserved modules that participate in similar biological processes (6, 44, 45). Although different approaches are available to identify homologous and orthologous genes (18), 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 (87), the PFAM database provides information about protein domains and families (88). Although reciprocal best hits (RBH) provide a practical solution to identify orthologs between closely related species, OrthoMCL and Inparanoid (19, 20) 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 (89). Although phylogeny-based orthology predictions are available in a number of plant comparative genomics resources (90), 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 (89, 91, 92). 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 (92, 93), 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 analysis combined gene expression clusters per species with homology information to identify conserved gene expression (Table 2.1). Examples in plants include Coexpressed biological Processes (CoP) (48), Expression Context Conservation (ECC) (49), Plant Network (PLaNet) (50) and STARNET2 (51) (Table 2.1). 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 2.3). Gene expression is typically compared between species in a pairwise manner and, optionally, information about conserved genes in multiple species is combined (50). Although this approach provides a first glimpse on the coexpressed genes that are conserved between different species (94), recently developed methods also applied statistical tests to verify if the number of shared orthologs between two expression clusters is significant (49, 50, 95, 96). 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 2.4, 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 (95, 96). However, for species with many multi-gene families like plants (97), 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 2.1), 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 (49). As a consequence, these models correct for specific expression breadth[1] 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 (96). Although this approach explores the two-dimensional parameter landscape (Figure 2.4) to find the best coexpression module definition, it is still required to pre-specify a coexpression stringency value for seed identification.

Complementary to a two-step approach, which first defines expression clusters and then filters coexpressed edges in the networks using gene homology information, Ficklin and Feltus (82) used a global network alignment approach to combine the co-expression

[1]Expression breadth is the number of tissues (or experiments) where a gene is expressed.

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.

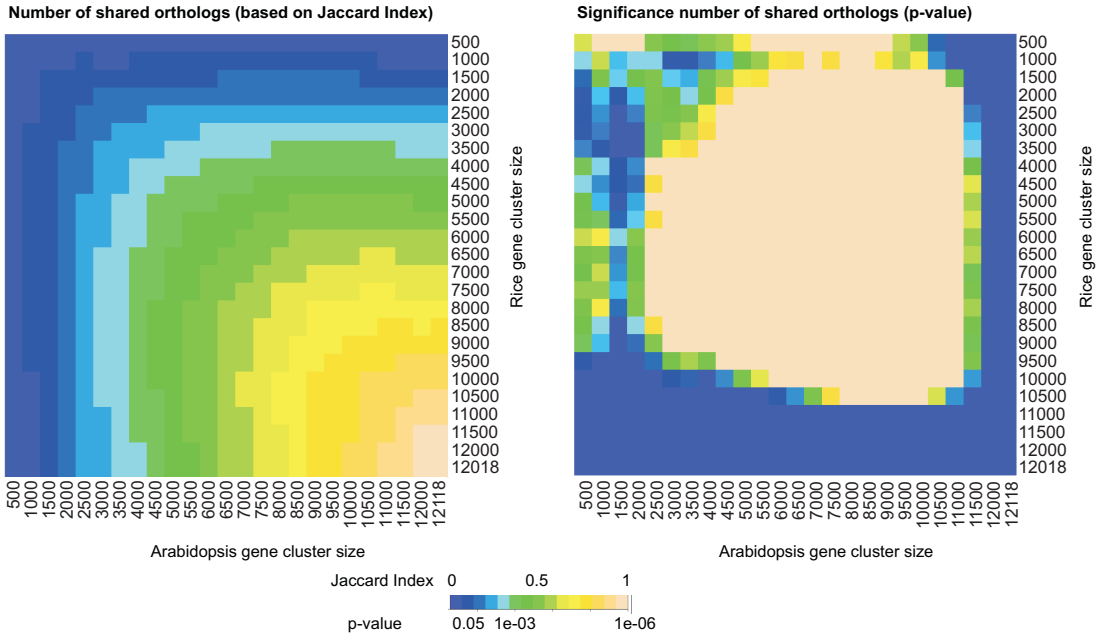


Figure 2.4: Significance testing of the number of shared orthologs during expression context conservation analysis - The Jaccard Index quantifies the number of shared orthologous gene families between two expression clusters (intersection number of shared families over union number of shared families) while the p-value indicates the significance of the observed number of shared orthologs (given the gene cluster sizes based on applied PCC threshold in Arabidopsis and rice, and the gene orthology information). Starting from the orthologous gene pair TSD2 (AT1G78240) and OS02G51860, gene expression clusters with different sizes were defined in Arabidopsis and rice, respectively. Expression clusters are expanded each time by adding 500 new genes with highest PCC similarity to the seed gene. As a consequence of increasing expression cluster sizes also the number of shared orthologs increases (left panel). However, based on permutation statistics, the p-value landscape (right panel) indicates that cluster sizes corresponding with a maximal Jaccard Index do not correspond with significant p-values.

2.6 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 (98), Reactome (99) or MapMan (100) has also been exploited (Table 2.1). 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 (101). 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 (49, 50, 82).

Graphviz and Cytoscape (102) are frequently applied software tools to graphically integrate expression networks, homology information and functional annotations (Table 2.1). Typically, genes are depicted by nodes while different edge attributes are used to represent expression similarity and homology information within and between species (Figure 2.5). 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[1] (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 2.1). Furthermore, meta-networks are an important instrument to identify regulatory interactions and cross-talk between different processes (50).

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 (49, 82) , 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.

[1] Network of networks

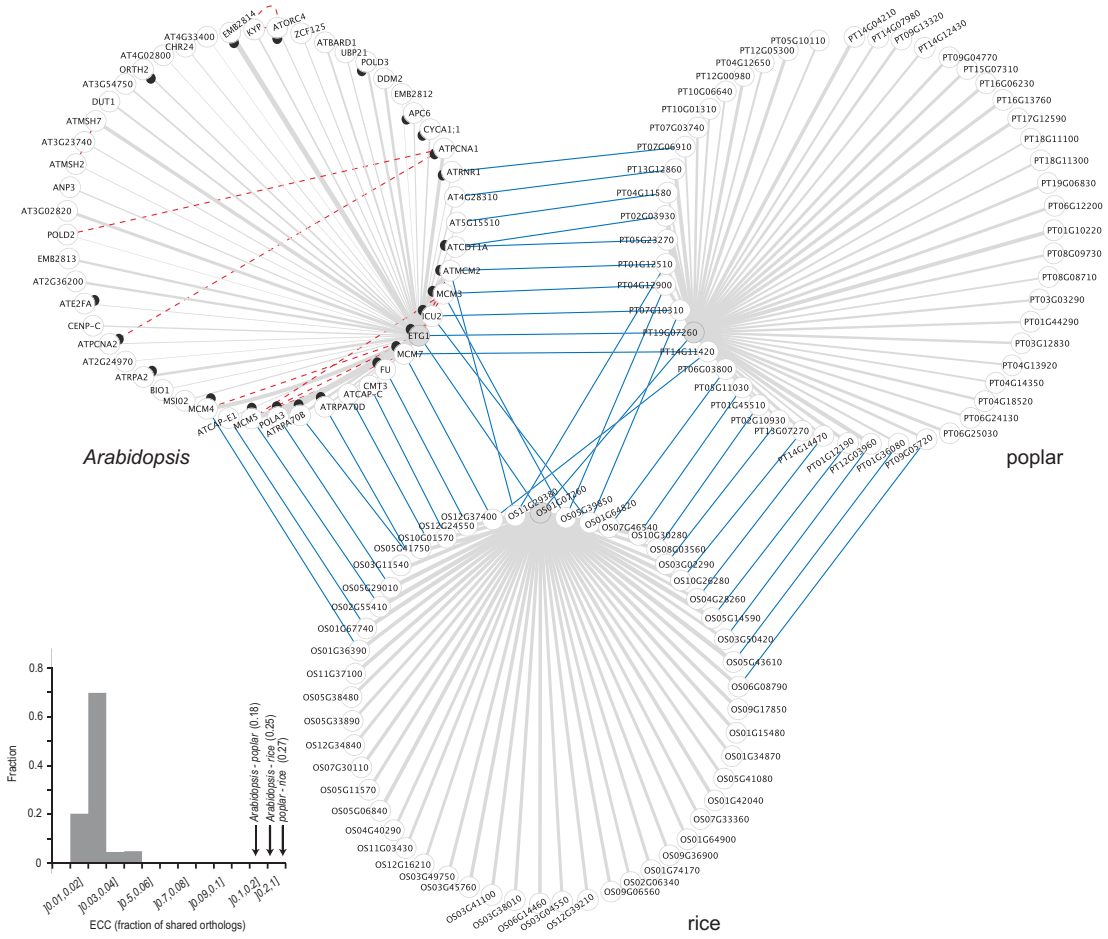


Figure 2.5: Plant orthologs with conserved co-expression (B) - Co-expression context analysis for the Arabidopsis ETG1 gene and its orthologs in poplar and rice (based on PLAZA 2.0 annotations). Grey edges represent co-expression links between ETG1 (query gene) and its top 50 coexpressed genes, weighted 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.

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.

2.7 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 2.5 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 (103), recent experimental work revealed it has an essential role in sister chromatin cohesion during DNA replication (104). 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 (38). 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 (85) 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 2.5). 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 2.5), 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 (105), comparative transcriptomics also offers a practical solution to identify functional homologs in multi-gene families (95). 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 2.2), 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.

2.8 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 (41), 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 (49, 50, 82). 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 (5). 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 (49).

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 (22, 38), 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 2.2), 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 2.1). 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 (106). 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.

2.9 Acknowledgments

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).

2.10 Author Contribution

This chapter is redrafted from an article published in the PCE (Plant, Cell and Environment) journal. The text is mostly written by me and Klaas Vandepoele. Michiel Van Bel and Ken S. Heyndrickx contributed in section 2.7, figure 2.5 and figure 2.2.

2.11 Supporting Information

The initial probeset definitions provided by for example Affymetrix frequently rely on draft gene annotations and therefore might suffer from cross-hybridization events (i.e. probes matching to different genes). In case where cDNA sequences or assembled unigenes were used to design the microarray, not all gene models identified using genome annotation are represented on the array (50). In addition, it is not always clear what

transcript, in case of alternative splicing, is actually measured using a specific probeset.

As a test case, remapping all 631,066 O. sativa Affymetrix GeneChip probes to the rice gene models (TIGR5 genome annotation (107)) showed that 43% of the original probesets contained one or more cross-hybridizing probes. These cross-hybridization effects can have an important role in imprecise profiling of low-copy-number molecules (59). Since it is, for all microarray studies, important to ascertain that a probe specifically measures the intended gene/transcript, the inaccuracies present in the original Affymetrix probesets bring up the need to develop a generic software tool to construct high-quality probesets. This can be done by mapping probes at the gene level or at the transcript level and only retaining uniquely mapping probes in the final probesets (108, 109). As an example, we created a new rice gene CDF file that allowed to reliably measure expression levels for approximately 75% of all non-transposable rice gene models (32,004 genes, requiring 8 or more probes per probeset). Although the creation of a transcript-specific CDF file, discriminating between different splice variants, is in theory possible, the initial gene-based probe design for most platforms results in only a small number of genes for which multiple transcripts can be measured (Figure 2.6). For instance, with a custom made rice transcript CDF file containing 5 or more non-cross-hybridizing probes per transcript, we could measure 29,976 rice transcripts covering 29,366 genes. Apart from the 28,160 genes matching a single transcript, only 589 genes were found for which two or more transcripts could be reliably studied.

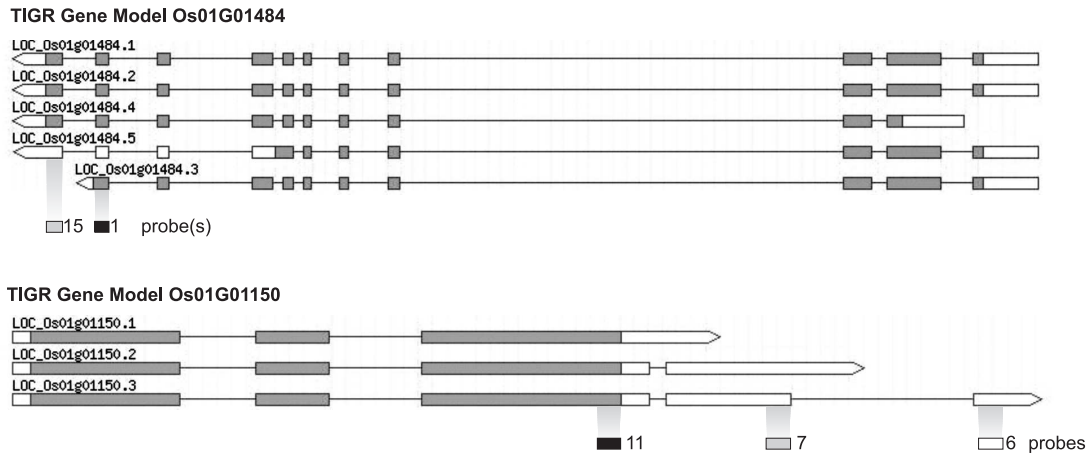


Figure 2.6: Probeset definitions at the gene and transcript level - Examples of two rice genes with alternative splicing and their probe organization at the gene and transcript level (TIGR5). Probe mappings which are shared over all transcripts are colored black, probes specific for one transcript are colored white while grey probes are only shared over a subset of transcripts. Black and white probes are most suited to construct probesets at the gene and transcript level, respectively.

3

Comparative Network Analysis Reveals That Tissue Specificity and Gene Function Are Important Factors Influencing the Mode of Expression Evolution in Arabidopsis and Rice

Sara Movahedi, Yves Van de Peer, and Klaas Vandepoele

3.1 Abstract

Microarray experiments have yielded massive amounts of expression information measured under various conditions for the model species Arabidopsis (*Arabidopsis thaliana*) and rice (*Oryza sativa*). Expression compendia grouping multiple experiments make it possible to define correlated gene expression patterns within one species and to study how expression has evolved between species. We developed a robust framework to measure expression context conservation (ECC) and found, by analyzing 4,630 pairs of

orthologous Arabidopsis and rice genes, that 77% showed conserved coexpression. Examples of nonconserved ECC categories suggested a link between regulatory evolution and environmental adaptations and included genes involved in signal transduction, response to different abiotic stresses, and hormone stimuli. To identify genomic features that influence expression evolution, we analyzed the relationship between ECC, tissue specificity, and protein evolution[1]. Tissue-specific genes showed higher expression conservation compared with broadly expressed genes but were fast evolving at the protein level. No significant correlation was found between protein and expression evolution, implying that both modes of gene evolution are not strongly coupled in plants. By integration of cis-regulatory elements, many ECC conserved genes were significantly enriched for shared DNA motifs, hinting at the conservation of ancestral regulatory interactions in both model species. Surprisingly, for several tissue-specific genes, patterns of concerted network evolution were observed, unveiling conserved coexpression in the absence of conservation of tissue specificity. These findings demonstrate that orthologs inferred through sequence similarity in many cases do not share similar biological functions and highlight the importance of incorporating expression information when comparing genes across species.

3.2 Introduction

Comparative sequence analysis provides valuable information about the functional parts encoded by a genome, mainly by exploring the conserved DNA sequences that code for proteins, RNAs, or regulatory elements (4). Interspecies comparisons have two major applications: conservation between species helps to detect and characterize functional elements, whereas differences can reveal biological adaptations linking genotype with phenotype (5). Recently, the increase in functional genomics data has transformed comparative approaches from basic sequence analysis to detailed studies of functional attributes such as gene expression or protein-protein interactions (6, 7). For instance, microarray experiments have yielded large amounts of genome-wide expression information under various conditions or in different tissues for several model species. Expression compendia grouping multiple microarray experiments make it possible to define correlated expression patterns between genes (42, 43). Genes within a coexpression cluster

[1] the rate of nonsynonymous substitutions (Ka)

are expected to have more similar functionality than those without expression similarity. In other words, functionally related genes are often coexpressed, both in closely related and more distantly related species (6, 44).

To compare expression data between different species, microarrays can be processed in three different manners (45). One approach combines multiple microarray experiments to identify differentially expressed genes in each species independently and then compare these genes among different species (46). This method requires for each species orthology information to link gene expression states across species and can be applied to both closely related or more distantly related species. Although orthologs are defined as homologs derived by a speciation event, it is important to note that sequence-based orthology inference does not necessarily imply functional equivalence (110). Another approach hybridizes samples from different but closely related species to the same microarray and requires similar experimental conditions as well as orthology information. Finally, separate arrays can be used to sample similar experimental conditions for different species, and all of them are analyzed together to investigate expression evolution of orthologs, paralogs, or specific functional categories (45). The latter approach was used, for example, to investigate species-specific gene duplications in human and mouse (47). In addition to comparison of expression profiles between orthologous genes, some studies also try to combine expression information with sequence evolution and gene function. Bergmann et al. (44) integrated expression data of six species with their genomic sequence information to identify coexpression conservation and to improve functional gene annotation. Based on graph theory, transcriptional networks were inferred revealing that highly connected genes are often conserved and have essential functional rules. Sets of genes that are conserved at both sequence and expression levels among multiple species are expected to play a key role in biological responses (6, 45).

Although comparative expression analysis is most straightforward when compatible expression data sets are used that cover equivalent conditions for all species, in this approach only a small fraction of all available data in different species can be utilized (5). Furthermore, how similar gene expression is modulated in distantly related species is still not understood, even when considering compatible conditions. To overcome these limitations, more advanced transcriptomics studies have shown that comparing coexpression, instead of the raw expression values, provides a valid alternative to

identifying regulatory modules and studying their evolution (6). For instance, gene expression between the four distantly related species *Caenorhabditis elegans*, *Drosophila melanogaster*, *Homo sapiens*, and *Saccharomyces cerevisiae* was compared by means of a coexpression meta-analysis (14). Pairs of species were considered, and for each gene the "expression context," which is based on the coexpression with all other genes that have unequivocal orthologous counterparts in both genomes, was compared. Significant expression context conservation (ECC) was found for many orthologs; but sequence and coexpression context evolution did not strongly correlate after duplication and speciation (14).

Rice (*Oryza sativa*) is one of the most important alimentary crops, with a relatively small genome size compared with many other cereals, and serves as a model for monocotyledons. Although the genome size of Arabidopsis (*Arabidopsis thaliana*), a model for dicotyledonous plants, is much smaller than that of rice (115 and 420 Mb, respectively), both species share a large number (56%−77%) of homologous genes (97, 111). Analysis of similarities and differences between the Arabidopsis and rice transcriptomes for similar organ types with custom-made oligomer microarrays revealed that similar portions were expressed in their corresponding organ types (65). In addition, evidence was found that a large fraction of rice genes lacking Arabidopsis homologs were expressed (65). Here, we developed a statistical framework to compare the expression of orthologous genes in rice and Arabidopsis. The importance of gene function and tissue specificity in correlation with the coexpression conservation was analyzed as well as the relationship between coding sequence and expression evolution.

3.3 Results

3.3.1 Relationship between Expression Similarity and Gene Function

To compare gene expression between the two model species Arabidopsis and rice, we assembled for both organisms, an expression compendium based on Affymetrix microarray experiments, retrieved from the National Center for Biotechnology Information Gene Expression Omnibus. Starting from a set of 322 Arabidopsis and 203 rice microarray slides, data normalization and averaging of replicates resulted in an initial expression data set of 129 Arabidopsis and 84 rice experiments. Subsequently, the effect of highly similar experiments per data set was reduced by the identification and

removal of highly redundant samples (see section 3.5) because they can introduce functional biases (38). After collapsing redundant conditions and removing transgenic or mutant experiments, we obtained a final data set covering 76 Arabidopsis and 63 rice experiments (Appendix B). By means of a custom-made chip description file (CDF) grouping only noncross-hybridizing probes in probe sets, the expression patterns of 19,937 Arabidopsis genes and 32,004 rice genes were monitored (see section 3.5).

As both compendia contained experiments covering different tissues, developmental stages, or (a)biotic treatments, we first determined whether biologically relevant information could be retrieved from both expression data sets. Starting from predefined gene sets that grouped genes based on Gene Ontology (GO) annotations, we used the expression coherence (EC) to quantify the level of expression similarity for functionally related genes in one species. EC is a measure for the amount of expression similarity within a set of genes and is high for a set of genes that converges into one cluster or a few tight coexpression clusters (112). Expression similarities between gene pairs were calculated with the Pearson correlation coefficient (PCC), and an EC value of 100% indicated that all genes were coexpressed with each other (see section 3.5). For both species, 16% to 25% of all 1,550 GO functional categories (with five or more genes) had high coexpression levels (EC $>$ 10%). As a control experiment, the influence of individual microarray experiments on the globally observed coexpression pattern was determined with jackknifing. The application of this bootstrapping procedure, which iteratively removes a subset of the initial data, showed that the observed EC values are robust for both species (Figure 3.2).

Many functional GO terms had significant EC in both species (Figure 3.1, black dots), and examples included general housekeeping functions related to DNA and RNA metabolism, ribosome biogenesis, translation, photosynthesis, tricarboxylic acid cycle, starch metabolic processes, and cell cycle (Supplemental Table S2 (See section 3.7)). Some GO terms only showed high EC values in one organism, and examples covered protein polymerization, defense response to fungus, and response to brassinosteroid stimulus in rice and cell recognition, phospholipid transport, and 1,3-b-glucan metabolism in Arabidopsis.

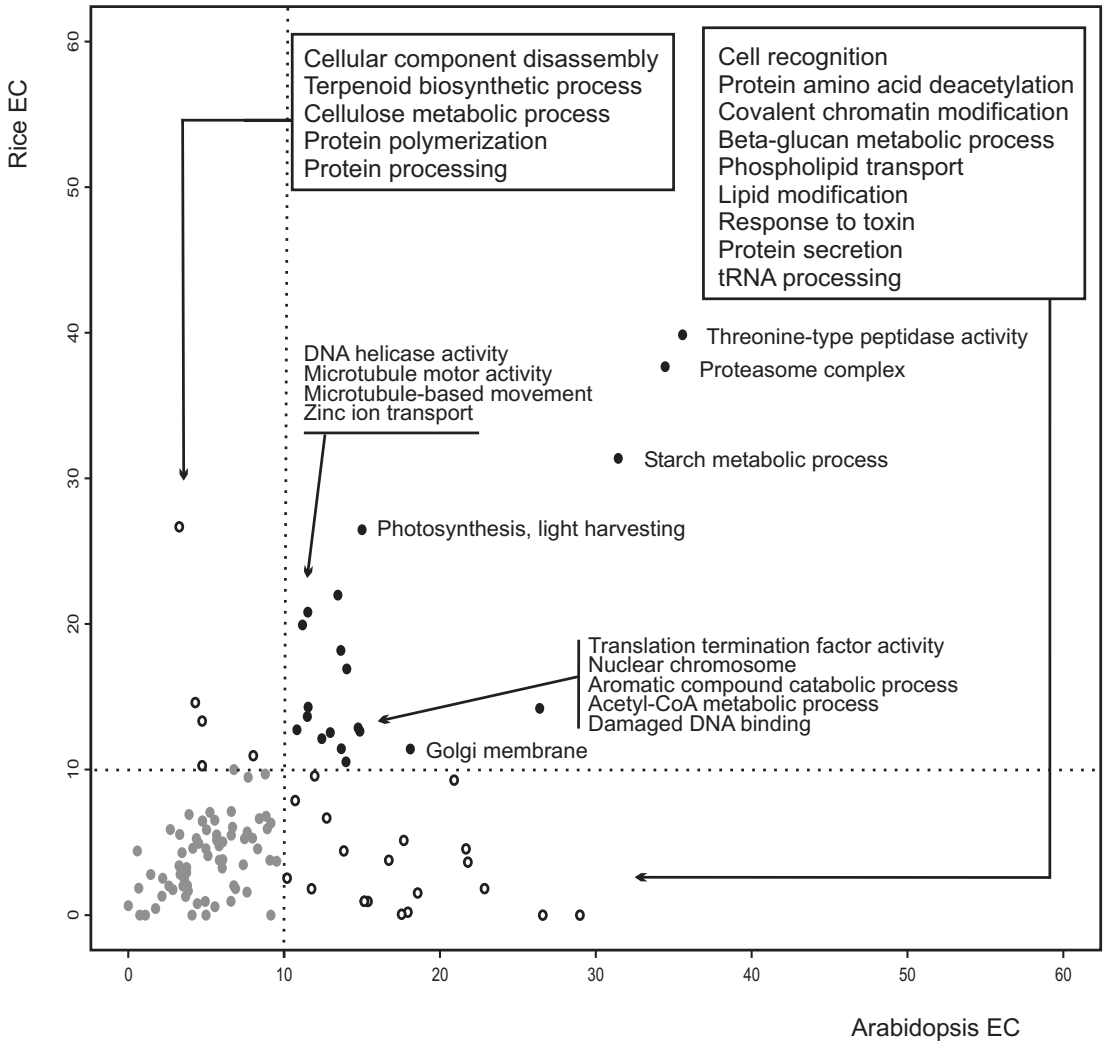


Figure 3.1: EC values for different GO categories - EC is shown for 114 nonredundant GO biological process terms in rice and Arabidopsis. GO terms with elevated coexpression in both organisms are shown as black dots, whereas those significantly coexpressed in only one species are shown as open dots. Only GO terms covering between 10 and 80 genes are shown, and redundant GO terms are omitted (for full list, see Supplemental Table S2).

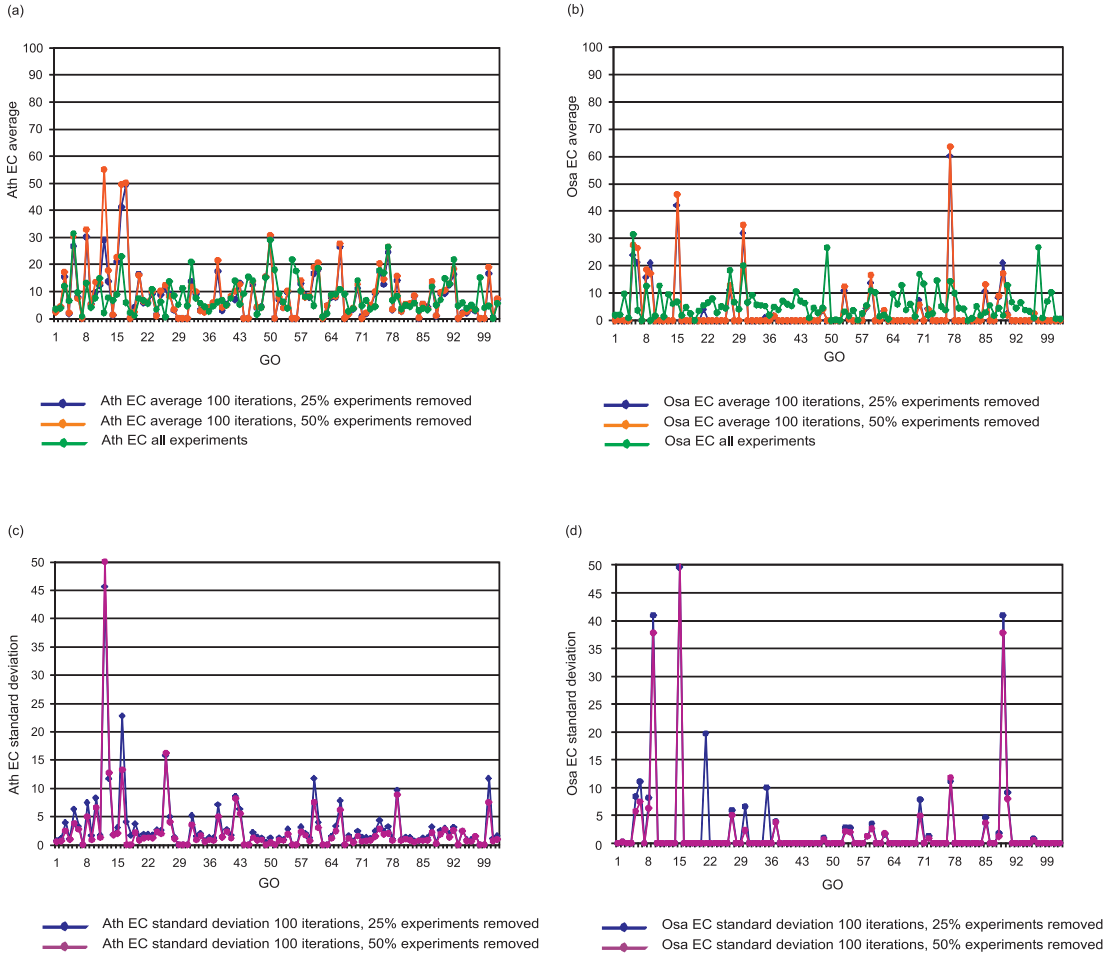


Figure 3.2: Effect of removing 25% and 50% random experiments from expression data on EC values with standard deviation and average measures - Only 102 GO labels in the BP hierarchy with gene range of [10,80] are studied. a) Arabidopsis EC average of 100 iterations for random (25% and 50%) experiment removal b) Rice EC average of 100 iterations for random (25% and 50%) experiment removal c) Arabidopsis EC standard deviation of 100 iterations for random (25% and 50%) experiment removal d) Rice EC standard deviation of 100 iterations for random (25% and 50%) experiment removal.

3.3.2 Measuring ECC between Arabidopsis and Rice Orthologs

Whereas EC values indicated that for some predefined functional categories differences in global coexpression levels existed, this measure did not report whether orthologous genes had similar expression patterns in different species or whether, for specific gene functions, the underlying coexpression network had diverged during evolution. To determine whether the expression profiles were conserved between two species, we developed an ECC score. According to Dutilh et al. (14), the expression context is based on the expression correlations between a query gene and all other genes in that species (or gene-centric coexpression cluster). The ECC was obtained by starting from a 1:1 orthologous seed gene pair, retrieving all coexpressed genes per species, and calculating how many orthologous genes were coexpressed in both species (Figure 3.3; see section 3.5). High ECC values indicated that in both species, the same genes were coexpressed, potentially reflecting the conservation of an ancestral coexpression module, whereas low ECC values suggested that since the divergence of both species, a substantial number of coexpression partners had been gained or lost. The OrthoMCL algorithm was applied to identify orthologous genes from the full set of Arabidopsis and rice proteins. In total, 7,911 orthologous gene families were found that covered approximately 12,000 genes for both species (see section 3.5). Based on the functional annotations for both species, all orthologous families were annotated with GO and Reactome (see section 3.5).

The biological relevance of ECC was evaluated by comparing the obtained ECC score for an orthologous gene pair against a null model in which neutral expression evolution was assumed. Application of an iterative sampling procedure generating 1,000 random pairs of gene-centric clusters (per species and per orthologous gene pair) and comparison with the expected ECC scores made it possible to classify the orthologous expression contexts as significantly conserved, diverged, or not significant (called ECC category in Figure 3.5). Whereas ECC scores nonsignificantly differing from the expected background distribution were simply considered as nonsignificant, significantly diverged ECC scores referred to orthologous gene pairs with fewer shared coexpression partners between both species than expected by chance, possibly reflecting positive selection. As the false discovery rate for ECC diverged genes was much higher than that of ECC conserved genes (25% and 0.82%, respectively), it should be treated with

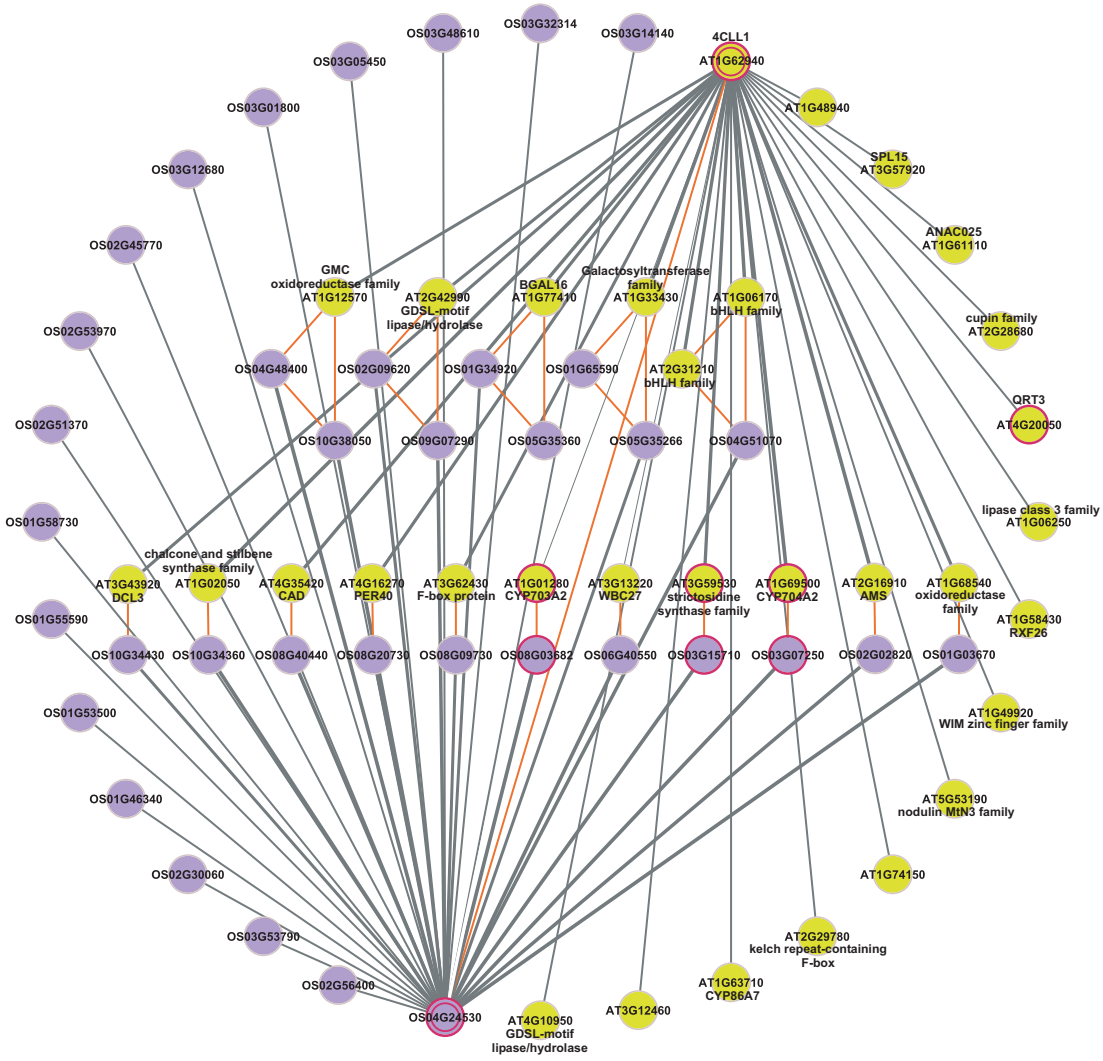


Figure 3.3: Calculation of ECC scores - Starting from an orthologous gene pair (rice gene OS04G24530 and Arabidopsis gene AT1G62940, marked with double circles), all coexpressed genes per species were retrieved (solid gray lines). The thickness of the line indicates the expression similarity measured with the PCC. All orthologous relationships are linked with orange lines and are used to determine the number of shared families between both coexpression clusters. Red circles represent GO functional annotations enriched in both clusters (GO:0009555 pollen development). The Jaccard Index of the depicted ECC conserved gene pair is 0.088 (16 shared families over 182 families in total). Note that for clarity, not all coexpressed genes and GO terms are depicted.

caution. Application of two different null models to estimate significance levels, one controlling tissue-specific expression and one correcting the degree of distribution in the network (or connectivity), yielded highly similar results (see section 3.5). Results from the connectivity model are reported throughout, whereas genes with nonconserved ECC refer to the categories diverged and nonsignificant.

ECC scores for all 4,630 1:1 orthologous rice-Arabidopsis gene pairs revealed that 77% had a conserved expression context, whereas 8.5% and 14.5% had a diverged and nonsignificant ECC, respectively (Figure 3.4). As expected, low ECC scores primarily included diverged and nonsignificant genes, whereas high values were mostly classified as significantly conserved (Appendix C). Functional biases determined with GO and MAPMAN enrichment analysis indicated that several biological processes linked with general housekeeping functions had highly conserved expression contexts (e.g. photosynthesis, plastid organization, DNA replication, RNA processing, cell division, and reproductive structure development). Interestingly, several functional categories were significantly underrepresented in ECC conserved genes, and examples include transcription factor (TF) activity, cell communication, ubiquitin-protein ligase activity, response to salt stress, and hormone stimulus. MAPMAN annotations indicated that different specific TF families (GRAS, Homeobox, WRKY, bZIP, and JUMONJI) were enriched in the set of nonconserved ECC genes.

3.3.3 ECC Varies for Different Functional Categories

The distributions of conserved, diverged, and nonsignificant ECC categories differed for various functional classification systems (Supplemental Table S4 (See section 3.7)). Average ECC scores indicated that genes assigned to KEGG/AraCyc pathways had a more conserved coexpression than those of the GO Biological Process or Molecular Function categories (average fractions of conserved ECC genes were 86%, 79%, and 76%, respectively). Categories with 50% or fewer conserved genes included regulation of signal transduction and cell communication, GTPase regulatory activity, starch metabolism, response to ethylene and GA stimulus, and response to starvation (Figure 3.4; Supplemental Table S4 (See section 3.7)). A high fraction of ECC diverged genes were found for receptor activity (31%), polysaccharide metabolism (36%), and starch metabolism (31%).

As a substantial number of genes involved in gene regulation, hormone response, and

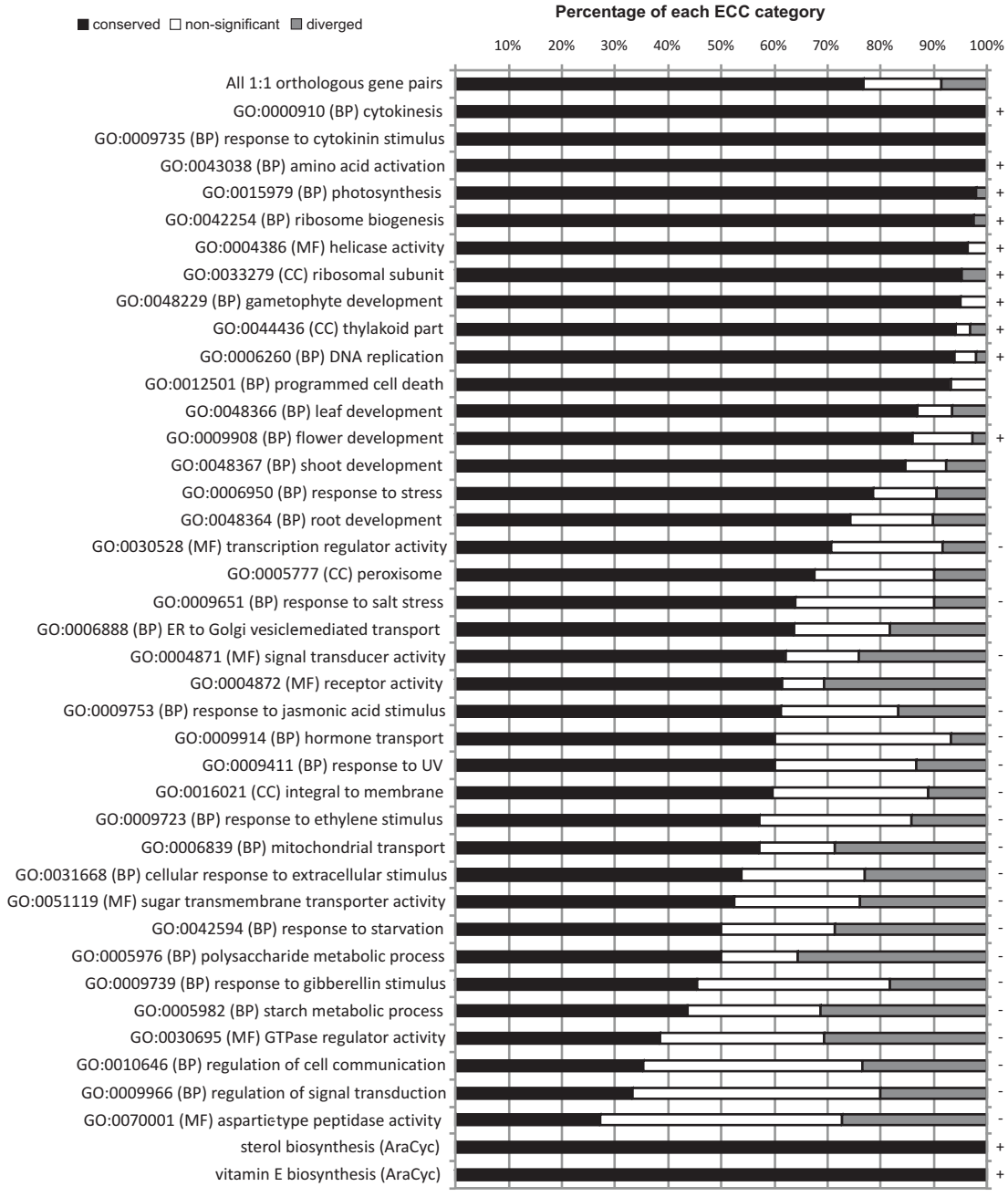
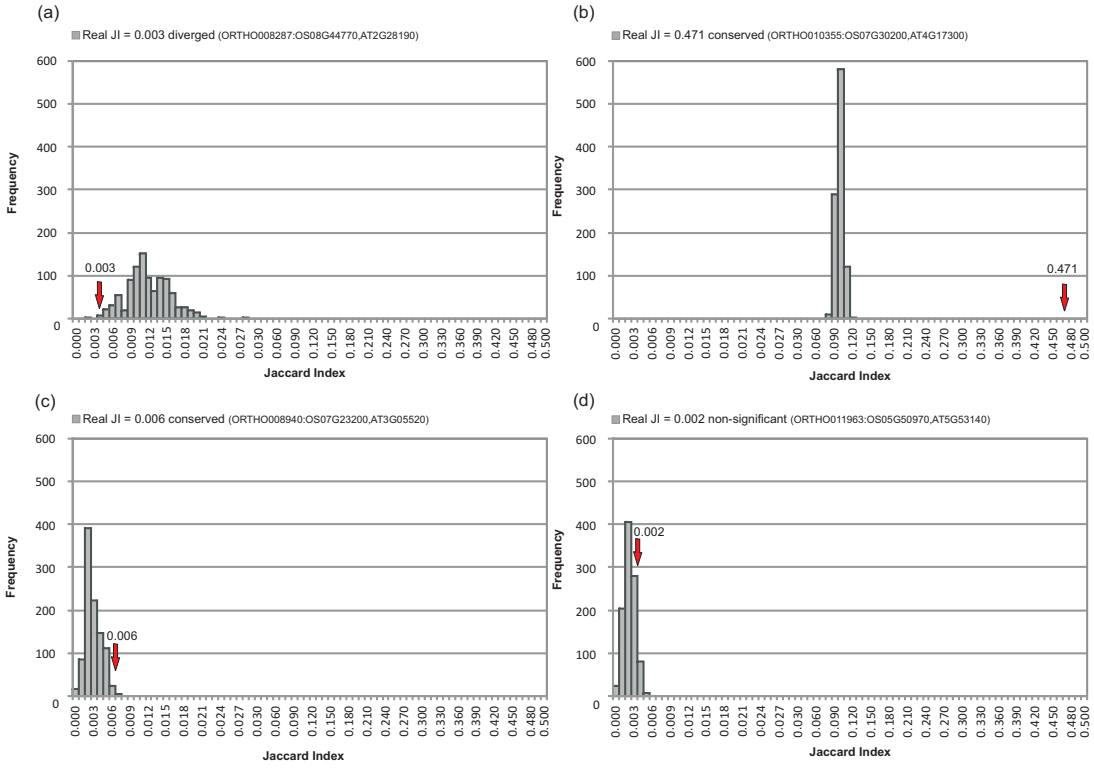


Figure 3.4: Comparison of ECC scores for different functional categories between Arabidopsis and rice - The fractions of genes with conserved, diverged, and nonsignificant ECC scores for different gene sets are shown. The first line reports the results for all 4,630 1:1 orthologous gene pairs, while the other lines refer to different functional sets delineated with GO and AraCyc (Reactome). The plus or minus sign indicates that the fraction of ECC conserved genes is significantly higher or lower compared with the overall ECC conservation level, respectively.



	Ortholog family	Osa gene	# Genes in Osa cluster	# Families in Osa cluster	Ath gene	# Genes in Ath cluster	# Families in Ath cluster	# Shared families between Osa and Ath clusters	Jaccard Index	ECC category
(a)	ORTHO008287	OS08G44770	844	816	AT2G28190	93	91	3	0.003	Div
(b)	ORTHO010355	OS07G30200	1494	1368	AT4G17300	1219	1114	795	0.471	Cons
(c)	ORTHO008940	OS07G23200	1490	1397	AT3G05520	18	18	9	0.006	Cons
(d)	ORTHO011963	OS05G50970	2830	2425	AT5G53140	15	15	5	0.002	Non

Figure 3.5: Examples of ECC scores assigned to different categories with random Jaccard Index scores - Plots are drown for a representative gene pair in each category: a) a diverged gene pair with low JI and b) a conserved gene pair with high JI c) a conserved gene pair with low JI and d) a non-significant gene pair with low JI. Grey boxes indicate random JI frequencies while red arrows show the real JI.

starch metabolism showed nonconserved coexpression contexts, we analyzed these categories in more detail. Focus on TFs revealed that developmental regulators were overall well conserved at the transcriptional level (e.g. ECC conserved TFs: 11 of 14 tissue, four of five leaf, seven of nine shoot, and nine of 12 flower development). Examples of highly diverged hormone-related TFs included HY5 (bZIP involved in light-regulated transcriptional activation), the auxin-responsive NAC domain-containing protein 9 (AT1G26870), MYB26 (AT3G13890; GA responsive), and the ethyleneresponse factor AT5G25190 (subfamily B-6 of the ERF/AP2 family). In contrast, TFs with highly conserved coexpression patterns covered OCP3, BLH1, and ATCDC5 (involved in response to fungus) as well as CLF, FMA, AGL16, LAS, and ATMYB5 (role in development). For all Arabidopsis genes discussed throughout this paper, a list with orthologous gene identifiers is provided (see section 3.5).

Several diverged stress-related genes encoded for DNA photolyases (CRY3 and PHR1), were responsive to DNA damage stimulus, or were involved in DNA repair (ATRAD51B /RAD51B and AT1G49980). GO categories such as "response to stimulus" or "response to stress" shared 78% conserved ECC genes, suggesting that general stress-related signaling was largely conserved between both species. Interestingly, the ECC conservation levels varied largely for some specific response categories: all responses to cytokinin stimulus geneswere ECC conserved (e.g. GCR1,ARR11/22, TSD2, ADA2B, MCB1, PAS1, AHK5, and CKI1), 62% of the responses to UV genes, but only 50% of the genes responsive to GA or Suc stimulus (Supplemental Table S4 (See section 3.7)).

Although starch metabolic genes had elevated levels of coexpression per species (Figure 3.1), many genes had nonconserved coexpression patterns. Starch metabolism is tightly coupled with photosynthesis, resulting in biosynthesis in transient starch granules during the day and nocturnal breakdown. Consequently, both starch synthases and different degrading enzymes, phosphatases, and transporters are required to maintain correct sugar levels in different plant tissues (113). Starch, stored in tuberous tissues or seeds, plays a central role as an energy source during germination. Whereas in most plants, ADPglucose, a substrate for starch synthases, is produced in the plastid through a ADPglucose pyrophosphorylase, cereals possess a cytosolic ADPglucose pyrophosphorylasesynthesizing ADPglucose in the developing endosperm that is then imported into the plastid for starch synthesis. Nonconserved ECC genes involved in starch metabolism included both genes involved in starch synthesis (ATSS3) and degradation

(isoamylases ISA2 and ISA3, a-amylase ATPU1, and b-amylase CT-BMY). Whereas most Arabidopsis starch-related genes were expressed in several tissues, including leaves and seeds, the corresponding rice genes were expressed in fewer tissues, with ECC diverged starch genes being primarily expressed in seed endosperm and embryos. These results suggest that, at the transcriptional level, the integration of light perception coupled with the complex regulation of starch metabolism in Arabidopsis and rice has diverged substantially since the divergence of dicots and monocots (114).

3.3.4 ECC Patterns for Evolutionarily Conserved Plant Genes

In addition to the study of expression conservation for different functional categories, we also delineated specific gene sets focusing on genome organization and phyletic distribution. Closely related species have extensive regions of shared gene content and order (or colinearity), but as the evolutionary distance between two species increases, colinear segments erode due to gene loss and rearrangements (115, 116). Starting from a set of 76 regions with conserved gene content and order between Arabidopsis and rice, 116 1:1 orthologous colinear gene pairs were extracted. Comparison of the degree of coexpression of neighboring genes in colinear regions revealed that only 5.3% and 5.9% of these genes had coordinated expression profiles in Arabidopsis and rice, respectively. Compared with the expected degree of coexpression (i.e. maximum 10% with the 90th percentile PCC threshold), these genes showed no strong evidence for large-scale coregulation (117). Of all orthologous colinear gene pairs, 69% had a conserved ECC.
Genes with homologs in all other plants are known as “core” genes and include essential genes covering the basic genetic toolbox in plants. The study of core plant genes combined with expression states provides a simple means to enlarge our understanding of the evolution of gene function versus regulation. From the 8,478 conserved Arabidopsis core genes (with homologs in all nine plant species present in the PLAZA comparative genomics platform; (89)), 1,787 genes were present in 1:1 orthologous families. Of these core genes, 80% (1,426 of 1,787) had conserved ECC patterns, revealing a 3% increase in conserved coexpression state compared with the average level in the complete data set ($P < 8.5e\text{-}05$; hypergeometric distribution).

3.3.5 Organization and Conservation of cis-Regulatory Elements in ECC Conserved Genes

To characterize the underlying mechanism of conserved coexpression patterns in Arabidopsis and rice, we analyzed the promoter regions of all genes with known cis-regulatory elements. First, known plant DNA motifs from AGRIS (118) and PLACE (119) were mapped onto the 1-kb upstream promoters followed by motif enrichment analysis per gene-centric cluster (85). To reduce the inclusion of false-positive motif instances, for a given gene-centric cluster, only motifs present in the seed gene promoter and significantly enriched in the corresponding coexpression cluster were retained for further analysis. A total of 308 and 276 DNA motifs were found to be enriched in Arabidopsis and rice, respectively, with an average number of 15 and 21 enriched motifs per promoter (covering 3,242 Arabidopsis and 3,267 rice genes).

For the 3,270 ECC conserved orthologous genes, all promoter motifs conserved between Arabidopsis and rice were determined to study whether conserved coexpression patterns correlated with shared cis-regulatory elements. As a control, we shuffled all enriched gene-motif annotations and compared the real motif conservation rates with the expected values. In total, 3,003 (84%) Arabidopsis genes were found with one or more conserved motif, whereas 161 DNA motifs were conserved in orthologous genes from both species. In contrast, the motif conservation rate in the nonconserved ECC genes and the control set was 75% and 73%, respectively. When genes with at least five or 15 conserved motifs were considered, the ECC conserved category contained 1,240 and 608 genes (i.e. 2- and 3.7-fold more than in the control data set, respectively).

At least 14 different regulatory elements were much more conserved between orthologous genes than expected by chance ($P < 0.05$), and several of them were related to (or showed GO enrichment for) conserved processes, such as photosynthesis (IBOXCORE, MYBST1, SORLIP1AT, and ABRELATERD1), ribosome biogenesis (SITEIIATCYTC), or basic transcriptional control (INRNTPSADB). Motifs enriched in more than 50 Arabidopsis genes, but not conserved in the orthologous rice genes, covered CBFHV (C-repeat-binding factors; dehydration-responsive element), ARR1AT (ARR1-binding element; response regulator), MYBGAHV (central element of the GA response complex), AP3SV40 (AP-3-binding site consensus sequence), and SV40COREENHAN (SV40 core enhancer).

3.3.6 Influence of Tissue Specificity, Protein Evolution, and Connectivity on ECC

Tissue-specific gene expression plays a fundamental role in multicellular systems, underlying development, function, and maintenance of diverse cell types within an organism. Here, we used τ (Tau in the figures) to quantify the level of differential expression across conditions (accounting for the quantitative variations in transcript levels) and to identify tissue specificity (40). τ ranges from 0 to 1, and high values correspond with low expression breadth (i.e. expressed in a few tissues). First, we investigated how tissue-specific or constitutive expression was linked with ECC and protein evolution (Figure 3.9). For orthologs with different expression patterns, the fraction of genes with conserved ECC rose with increasing τ values (Figure 3.6). Both for Arabidopsis and rice, the median t values were significantly higher for ECC conserved than ECC nonconserved genes (Arabidopsis $\tau = 0.184$ and 0.169, respectively [$P <$ 2.6e-05]; rice $\tau = 0.200$ and 0.175, respectively [$P <$ 2.2e-10]). For all orthologous gene pairs, the rate of nonsynonymous substitutions (Ka) was used to measure evolution at the protein-coding level. For both species, comparison of τ as a function of Ka showed that tissue specificity correlated with increasing rates of protein evolution (Figure 3.7).

Ka distributions for genes with different ECC patterns did not differ significantly in protein evolution for genes with conserved and nonconserved coexpression contexts (median Ka values of 0.323 and 0.329, respectively [$P = 0.28$]). Plotting the distribution of ECC categories for genes with increasing rates of protein evolution confirmed the absence of a strong association between expression and protein evolution (Figure 3.8). The ratio of the number of nonsynonymous substitutions to the number of synonymous substitutions (Ka/Ks) was used to evaluate whether the selective pressure that acted on a protein-coding gene varied per ECC category. Although significant, the difference in the level of purifying selection between different ECC types was small (median Ka/Ks values for conserved and nonconserved genes were 0.087 and 0.084, respectively [$P = 0.023$]).

Finally, we evaluated whether the connectivity of a gene, defined as the number of coexpressing partners, had an influence on the evolution of orthologs or tissue specificity. As shown in Figure 3.10, both in Arabidopsis and rice highly connected genes were enriched for ECC conserved genes, while no clear trend was observed for τ or Ka (data not shown). Control for rates of protein evolution (by binning genes based on

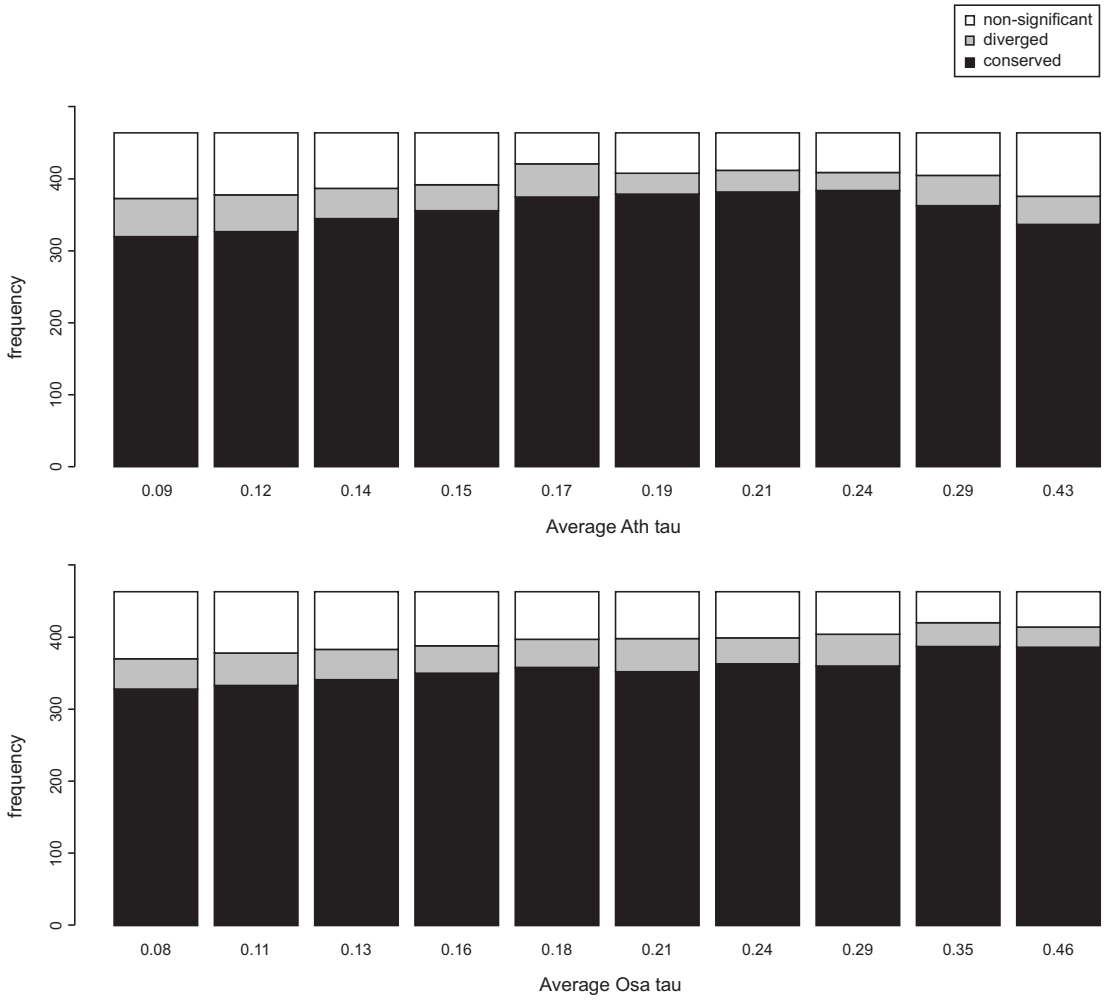


Figure 3.6: ECC as a function of Tau - Tau values for all 1:1 orthologous families are divided in 10 equal bins. For each bin the average Tau value is reported and bars show the fraction of conserved, diverged and non-significant ECC categories (black, grey and white, respectively). The top and down panel shows results for Arabidopsis and rice, respectively.

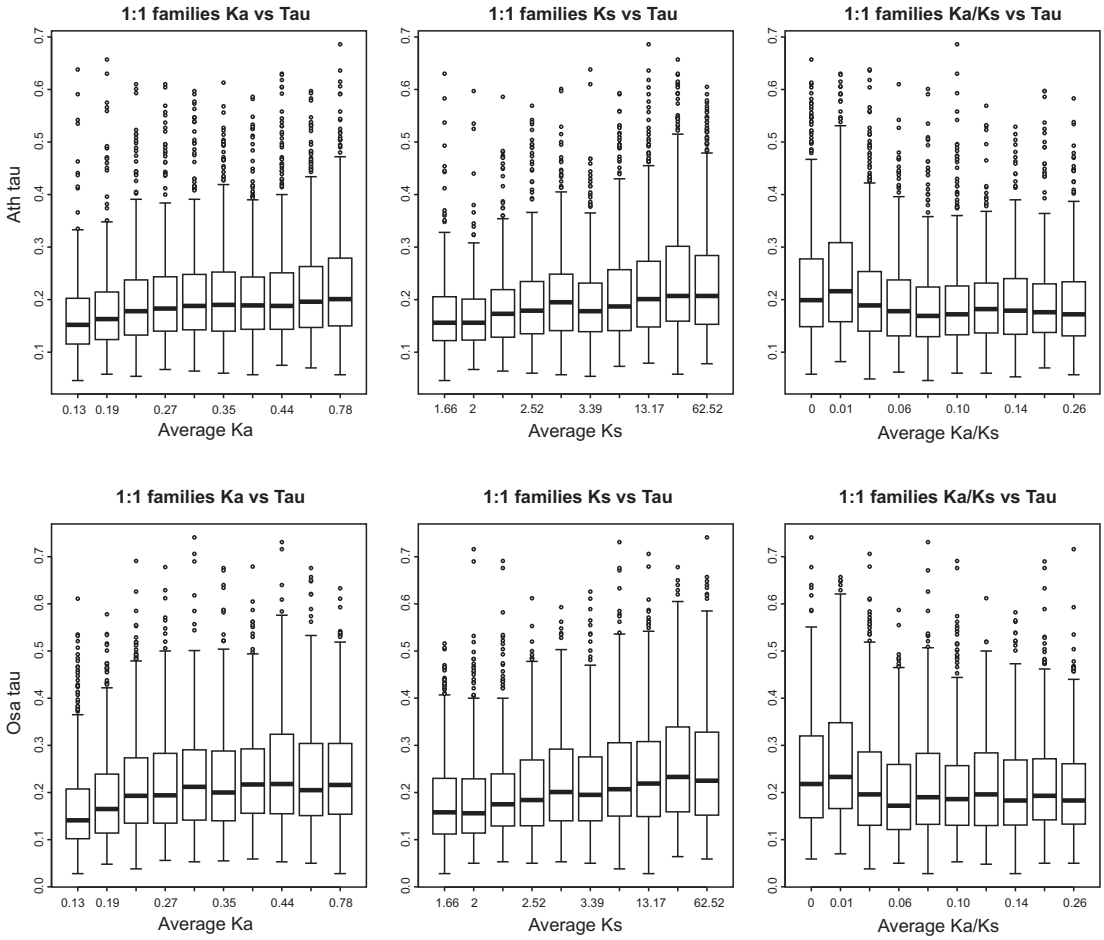


Figure 3.7: Tau as a function of Ka, Ks, and Ka/Ks - Ka, Ks and Ka/Ks values of all 1:1 orthologous families divided in 10 equal bins. For each bin the average value is shown and the box plots indicate the distribution of Tau values per bin (central bold line indicates median while the lower and upper lines refer to the 25th and 75th percentile, respectively). The upper and lower panel shows results for Arabidopsis and rice, respectively.

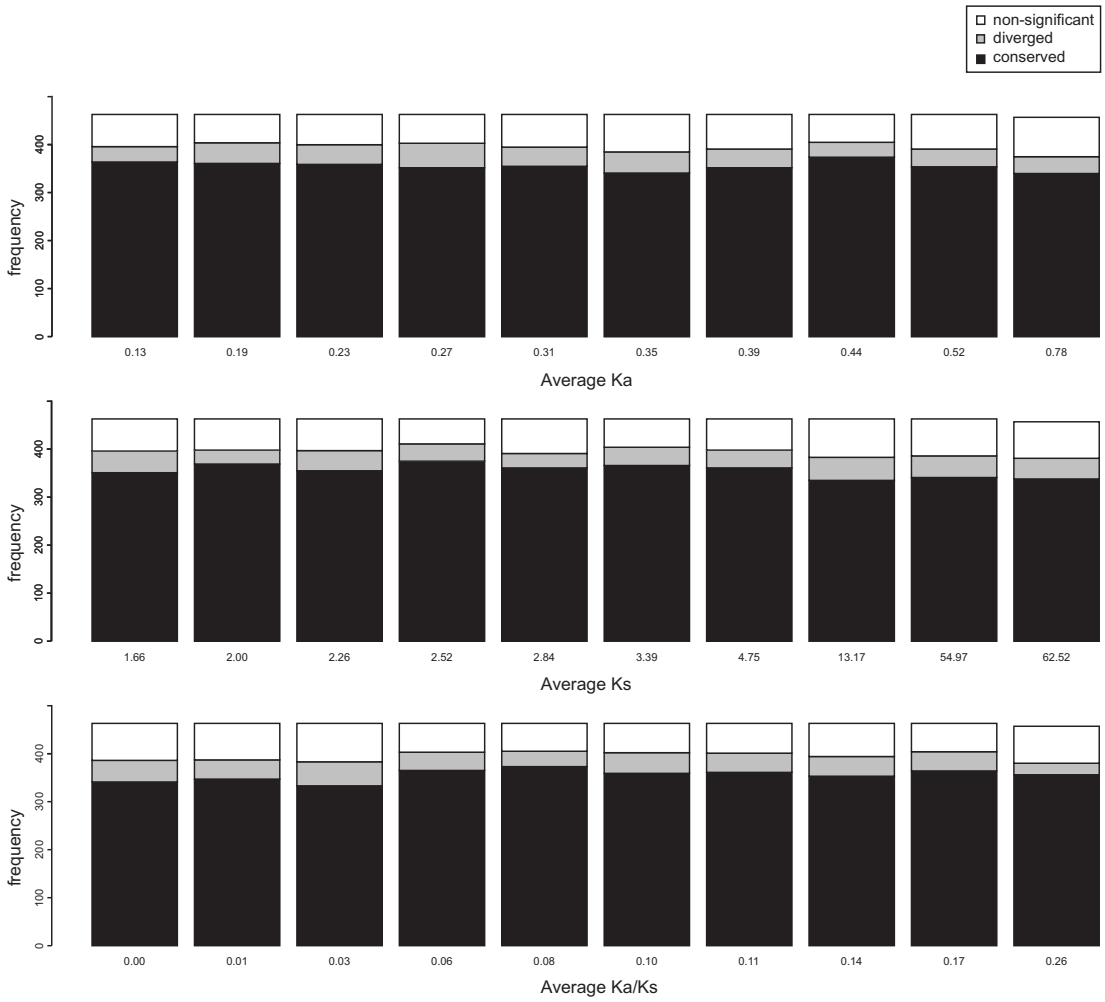


Figure 3.8: ECC as a function of Ka,Ks,and Ka/Ks - Ka (Ks,Ka/Ks) values of all 1:1 orthologous families are divided in 10 equal bins with the average Ka (Ks,Ka/Ks) value reported per bin. Bars show the fraction of conserved, diverged and non-significant ECC categories (black, gray and white, respectively).

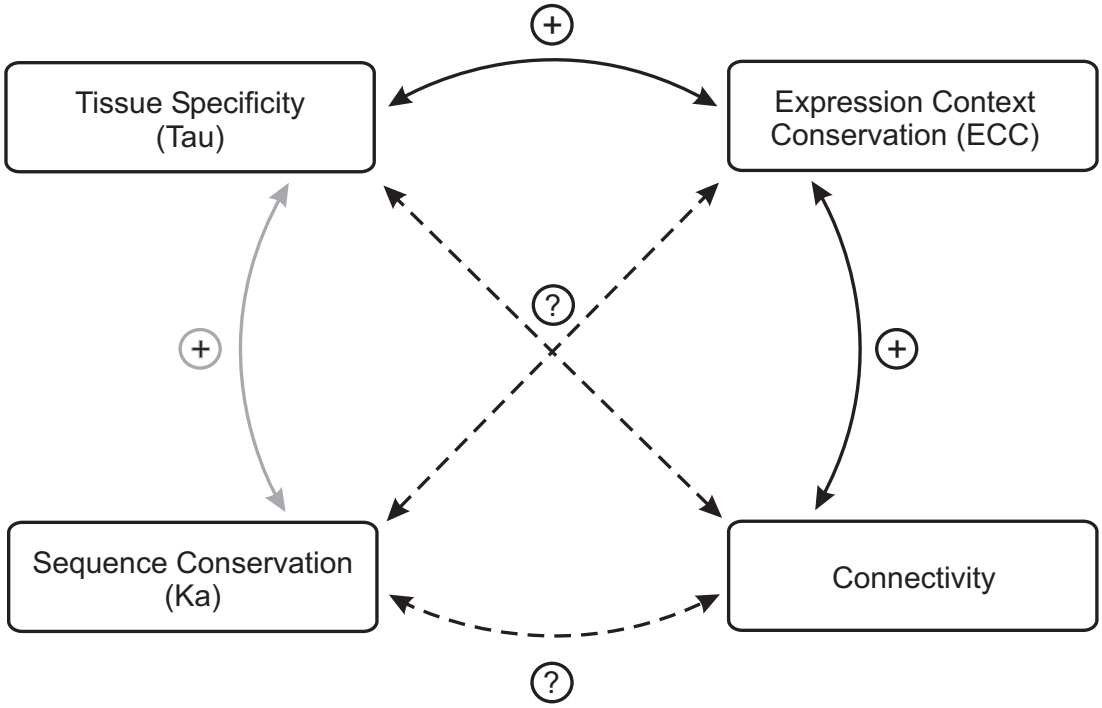


Figure 3.9: Summary of the correlations between expression and sequence evolution, connectivity, and tissue specificity - Plus signs denote positive correlations, whereas question marks and dotted lines indicate that no positive or negative correlation is found. Correlations deemed significant with the Mann-Whitney U test are highlighted as solid black arrows.

Ka) confirmed that, overall, ECC conserved genes were more tissue specific than their nonconserved counterparts (Figure 3.11). The presence of a TATA promoter motif did not correlate with the level of coexpression conservation.

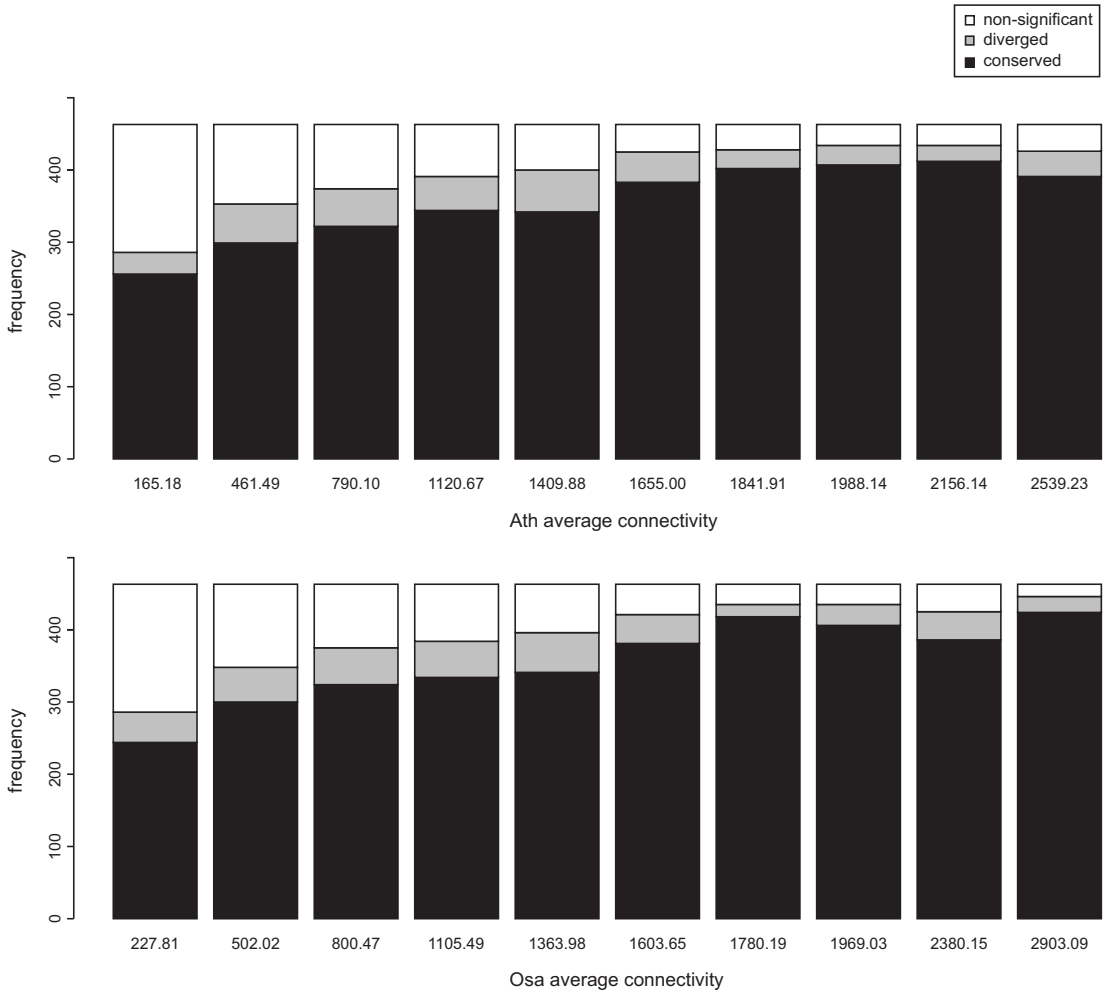


Figure 3.10: ECC as a function of connectivity - Connectivity values, measured as the number of coexpressed genes, of all 1:1 orthologous families are divided in 10 equal bins and average connectivity values are reported per bin. Bars show the fraction of conserved, diverged and non-significant ECC categories (black, gray and white, respectively). The upper and lower panel shows results for Arabidopsis and rice, respectively.

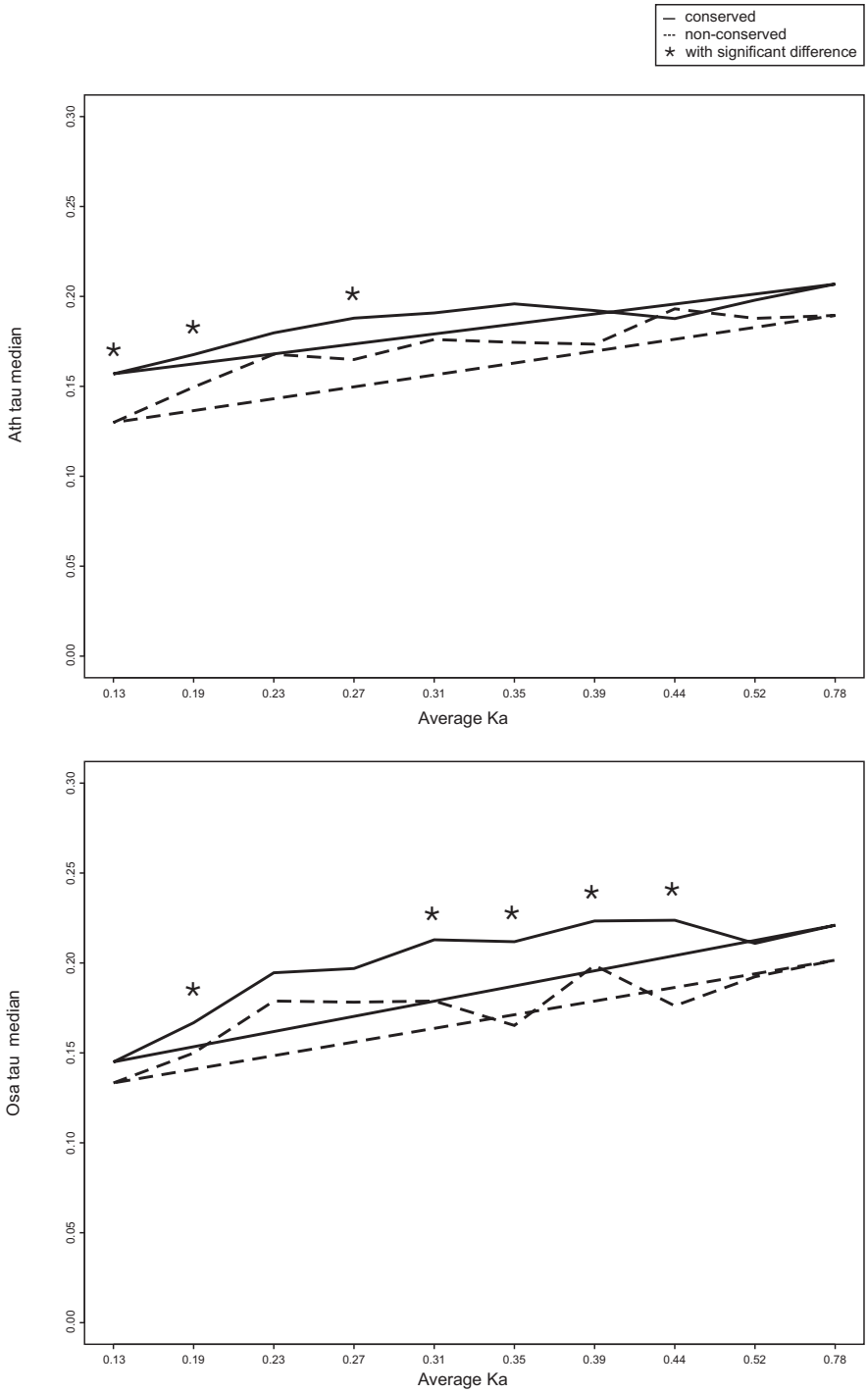


Figure 3.11: Tau as a function of Ka and ECC - Ka values of all 1:1 orthologous families are sorted and binned into 10 equal bins and each bin is presented with the average Ka values of that bin. The solid and dashed lines display median tau values of conserved and non-conserved (diverged and non-significant) families respectively. Bins are signed with a star if the median tau values of conserved and nonconserved categories are significantly different.

3.3.7 Concerted Network Evolution for Genes without Conserved Tissue Specificity

As in vertebrates (120), in Arabidopsis and rice many genes show tissue specificity as well as expression conservation (e.g. AT3G04700, AT3G17060, GATL4, AMS, and AG in flower; AT4G31830, AT2G28420, AT1G05510, AT3G12960, and ATOEP16-2 in seed; AT1G30870 in root). However, based on a set of tissue-specific Arabidopsis genes (75), several orthologs (five of 14) were broadly expressed in rice. By including additional Arabidopsis and rice genes with high τ values in our expression compendia, we found several orthologs without conserved tissue specificity (malate synthase [AT5G03860], seed; C3HC4-type RING finger [AT2G20650], root xylem; selenium-binding protein [AT3G23800], root cortex; TET4, stage 8 seeds without siliques; Figure 3.12). Whereas for a large number of genes the loss of tissue specificity in one species coincided with nonconserved coexpression patterns (Table 3.1; type II), more than 20 genes were identified with tissue-specific expression in only one species but also with ECC (type I). In this case, the tissue-specific gene under investigation as well as the genes showing a conserved expression adopted, in a concerted manner, a different expression pattern in the other species. Figure 3.13 depicts a flower-specific (stage 15 stamen pollen) Arabidopsis Ser carboxypeptidase (SCPL38) with functional enrichment for reproduction and flower development. Although the orthologous gene had a broad expression pattern in rice (thin borders of rice genes in Figure 3.13), the coexpression with several known flower development genes such as SEP1-3, AG, and AP3 was conserved in both species. Another example is the concerted network for Arabidopsis IQD10, a gene involved in phloem or xylem histogenesis but with ubiquitous expression in rice (Figure 3.14). Finally, a small set of genes was specifically expressed in both species, but for different tissues (type III). Although all these genes have a nonconserved ECC, in at least one case (AT1G70500, pectin lyase) the coexpression contexts in both species had a similar GO enrichment (carbohydrate metabolism), suggesting a similar molecular function.

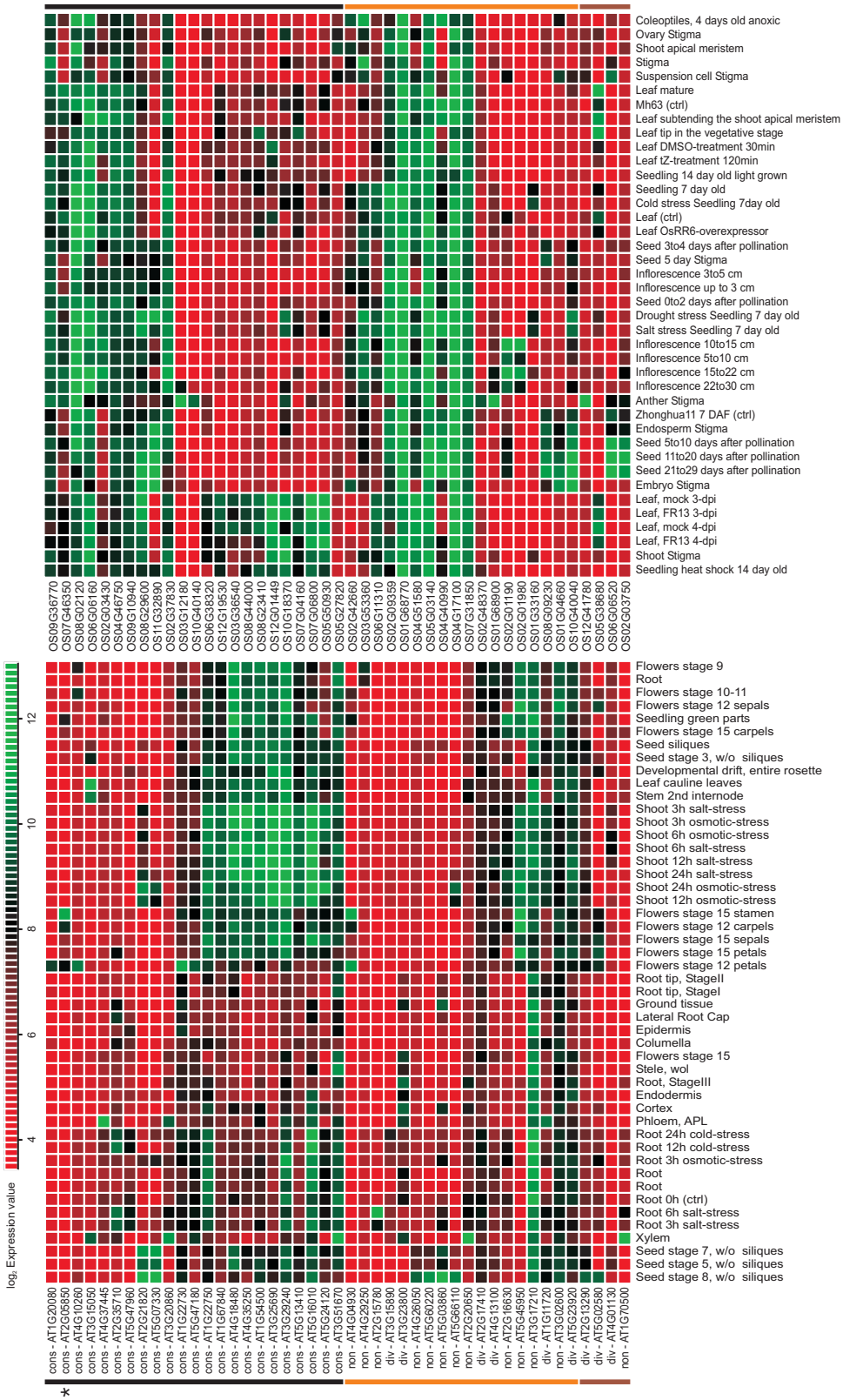


Figure 3.12: Expression evolution of tissue-specific genes - Expression heat map of orthologous Arabidopsis genes (left) and rice genes (right) lacking conservation of tissue specificity. Green values indicate expression above background, whereas the prefixes "cons," "div," and "non" indicate the ECC category. Black, orange, and brown bars correspond to type I, II, and III genes, respectively (for details, see Table 3.1).

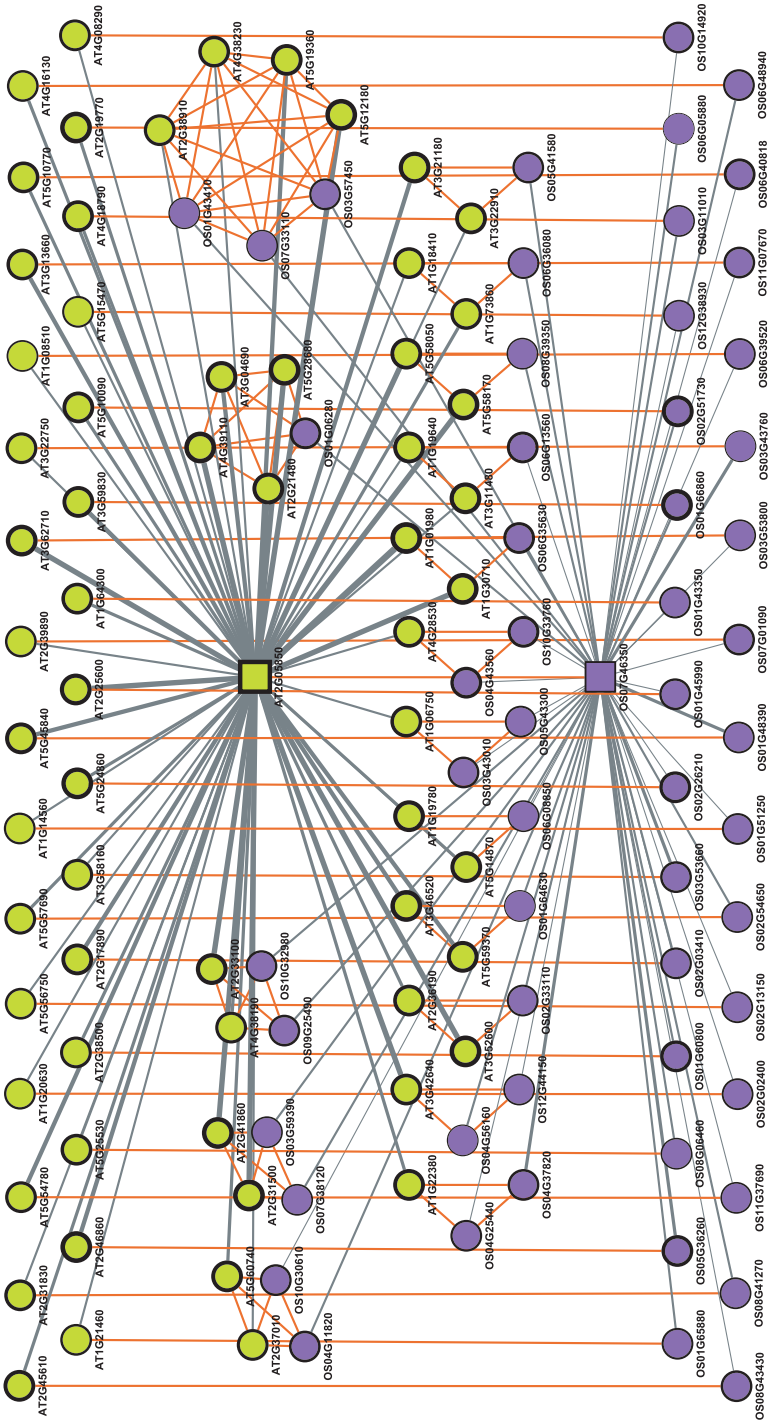


Figure 3.13: Expression evolution of tissue-specific genes - Expression network of concerted expression divergence for the flower-specific Arabidopsis gene SCPL38. Gray and orange lines show coexpression and orthology relationships, respectively, whereas the thickness of the gray lines indicates the expression similarity. Green and purple nodes denote Arabidopsis and rice genes, respectively, whereas the orthologous seed genes are drawn as boxes (and indicated by the asterisk in the heat map shown in 3.12). Node border thickness gives the tissue-specific expression measured with τ.

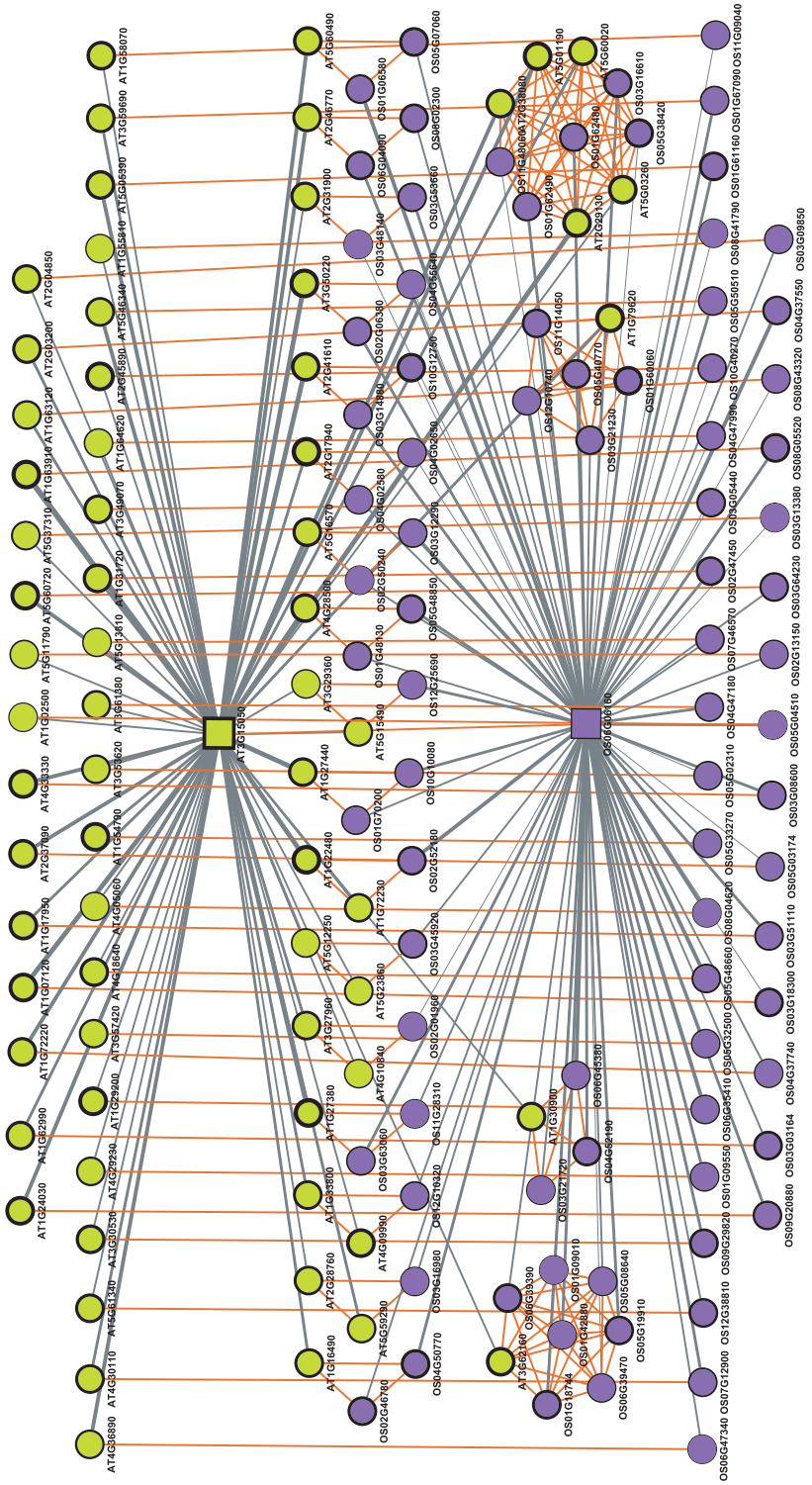


Figure 3.14: Concerted expression network evolution for the tissue specific Arabidopsis gene AT3G15050 - Grey and orange lines show coexpression and orthology relationships while the thickness of the line indicates the expression similarity. Green and purple nodes denote Arabidopsis and rice genes whereas the orthologous seed genes are in square shape. Node border thickness indicate the tissue specific expression measures used Tau.

Ath Gene (1)	Osa Gene (1)	Ath Ebreadth (Tau)	Osa Ebreadth (Tau)	ECC category	Ath tissue (2)	Osa tissue	Type (3)	Description (4)
AT1G20080	OS09G36770	1 (0.492)	58 (0.130)	cons	flower	-	I	SYTB
AT1G22730	OS03G12180	19 (0.309)	2 (0.706)	cons	-	anther-stigma	I	MA3 domain-containing protein
AT1G22750	OS06G38320	63 (0.136)	2 (0.407)	cons	-	leaf	I	co-factor metabolism*
AT1G54500	OS08G23410	55 (0.191)	6 (0.425)	cons	-	leaf, seedling	I	rubredoxin family protein
AT1G67840	OS12G19530	50 (0.156)	4 (0.468)	cons	-	leaf	I	chloroplast sensor kinase (CSK)
AT2G05850	OS07G46350	3 (0.593)	36 (0.232)	cons	flower	-	I	serine carboxypeptidase-like 38 (SCPL38)
AT2G21820	OS08G29600	8 (0.601)	21 (0.328)	cons	seed, shoot osmotic-stress	-	I	reproductive structure development*
AT2G35710	OS04G46750	6 (0.405)	61 (0.137)	cons	root	-	I	glycogenin glucosyltransferase (glycogenin)-related
AT3G15050	OS06G06160	4 (0.558)	59 (0.152)	cons	leaf, xylem, stem	-	I	IQ-domain 10 (IQD10)
AT3G20860	OS02G37830	3 (0.451)	61 (0.122)	cons	xylem, phloem, root salt-stress	-	I	member of the NIMA-related serine/threonine kinases (ATNEK5)
AT3G25690	OS12G01449	52 (0.247)	9 (0.626)	cons	-	leaf,shoot-stigma,seedling	I	actin binding protein required for normal chloroplast positioning (CHUP1)
AT3G29240	OS10G18370	67 (0.156)	9 (0.504)	cons	-	leaf,shoot-stigma, seedling	I	chlorophyll metabolic process*
AT3G51670	OS05G27820	69 (0.172)	1 (0.402)	cons	-	shoot	I	SEC14 cytosolic factor family protein / phosphoglyceride transfer family protein

AT4G10260	OS08G02120	3 (0.460)	61 (0.131)	cons	flower	-	I	pfkB-type carbohydrate kinase family protein
AT4G18480	OS03G36540	55 (0.204)	2 (0.521)	cons	-	leaf	I	CHLI subunit of magnesium chelatase which is required for chlorophyll biosynthesis (CHLI1)
AT4G35250	OS08G44000	51 (0.250)	3 (0.413)	cons	-	leaf	I	vestitone reductase-related
AT4G37445	OS02G03430	1 (0.547)	25 (0.306)	cons	phloem, APL	-	I	disaccharide biosynthesis*
AT5G07330	OS11G32890	6 (0.618)	13 (0.519)	cons	seed, shoot osmotic-stress	-	I	response to abscisic acid stimulus*
AT5G13410	OS07G04160	50 (0.169)	6 (0.429)	cons	-	leaf,shoot-stigma,seedling	I	immunophilin / FKBP-type peptidyl-prolyl cis-trans isomerase family protein
AT5G16010	OS07G06800	64 (0.230)	6 (0.584)	cons	-	leaf,shoot-stigma,seedling	I	3-oxo-5-alpha-steroid 4-dehydrogenase family protein / steroid 5-alpha-reductase family protein
AT5G24120	OS05G50930	35 (0.323)	6 (0.574)	cons	-	leaf,shoot-stigma,seedling	I	sigma factor E (SIGE)
AT5G47180	OS10G40140	13 (0.261)	1 (0.676)	cons	-	anther-stigma	I	vesicle-associated membrane family protein / VAMP family protein
AT5G47960	OS09G10940	3 (0.415)	59 (0.089)	cons	root	-	I	Encodes a small molecular weight g-protein (ATRABA4C)
AT1G11720	OS08G09230	34 (0.224)	7 (0.548)	div	-	seed	II	ATSS3 (starch synthase 3)
AT2G15780	OS06G11310	2 (0.604)	31 (0.291)	non	root	-	II	glycine-rich protein
AT2G16630	OS02G01190	33 (0.301)	6 (0.620)	non	-	inflorescence	II	proline-rich family protein
AT2G17410	OS02G48370	22 (0.113)	1 (0.514)	div	-	anther-stigma	II	ARID/BRIGHT DNA-binding domain-containing protein

AT2G20650	OS07G31850	6 (0.471)	63 (0.059)	non	xylem, cortex, root salt-stress	-	II	zinc finger (C3HC4-type RING finger) family protein
AT3G02600	OS01G04660	73 (0.172)	4 (0.492)	non	-	seed, embryo-stigma	II	phosphatidic acid phosphatase (LPP3)
AT3G15890	OS02G09359	1 (0.416)	59 (0.132)	div	root	-	II	protein kinase family protein
AT3G17210	OS01G33160	76 (0.123)	3 (0.537)	non	-	salt-drought-cold stress seedling	II	HS1 (HEAT STABLE PROTEIN 1)
AT3G23800	OS01G68770	7 (0.437)	63 (0.164)	div	root, xylem, phloem, cortex	-	II	selenium-binding protein 3 (SBP3)
AT4G04930	OS02G42660	4 (0.591)	21 (0.350)	non	flower	-	II	sphingolipid delta4-desaturase, involved in sphingolipid biosynthesis (DES-1-LIKE)
AT4G13100	OS01G68900	16 (0.200)	2 (0.519)	div	-	anther-stigma	II	zinc finger (C3HC4-type RING finger) family protein
AT4G26050	OS04G51580	1 (0.416)	44 (0.229)	non	seed	-	II	leucine-rich repeat family protein
AT4G29250	OS03G53360	1 (0.402)	21 (0.363)	non	flower	-	II	transferase family protein
AT5G03860	OS04G40990	5 (0.566)	30 (0.311)	non	seed, cortex	-	II	protein with malate synthase activity (MLS)
AT5G23920	OS10G40040	60 (0.174)	9 (0.516)	div	-	seed, embryo-stigma	II	unknown
AT5G45950	OS02G01980	44 (0.390)	3 (0.676)	non	-	inflorescence	II	GDSL-motif lipase/hydrolase family protein
AT5G60220	OS05G03140	1 (0.538)	63 (0.080)	non	seed	-	II	member of TETRASPANIN family (TET4)

AT5G66110	OS04G17100	3 (0.537)	63 (0.129)	non	seed,shoot osmotic-stress	-	II	metal ion binding
AT1G70500	OS02G03750	2 (0.630)	4 (0.528)	non	xylem, root	seed	III	polygalacturonase, putative / pectinase, putative
AT2G13290	OS12G41780	5 (0.283)	1 (0.565)	div	seed	anther-stigma	III	glycosyl transferase family 17 protein
AT4G01130	OS06G06520	5 (0.344)	5 (0.640)	div	shoot	seed	III	acetylesterase, putative
AT5G02580	OS05G38680	7 (0.479)	9 (0.557)	div	flower	leaf	III	unknown

(1) Ath: Arabidopsis. Osa: rice

(2) - indicates expression in a large number of tissues

(3) Type I: ECC conserved, tissue specific expression in only one species; Type II: ECC non-conserved, tissue specific expression in only one species; Type III: ECC non-conserved, tissue specific expression in both species but different tissue

(4) Descriptions indicated by asterisk denote GO enrichments based on the coexpression context of the tissue specific gene with unknown function

Table 3.1: Genes showing non-conserved tissue specific expression in Arabidopsis and rice.

3.4 Discussion

Whereas comparative sequence analysis is a powerful tool to study genome evolution, to discover conserved orthologous genes, and to characterize species- specific gene families, the integration of functional genomics data provides an additional layer of information to study gene function and regulation. Recently, expression data, functional gene annotations, protein-protein interaction data, knockout phenotype information, and cis-regulatory elements have been combined to delineate coexpressed modules, to predict new gene functions, and to identify transcriptional regulatory interactions (13, 38, 85, 121, 122, 123). Although most coexpression approaches have been used to predict different types of gene-gene interactions in model species such as Arabidopsis and rice, interspecies comparisons have identified examples of conserved coexpression modules in plants (65, 66).

To determine which factors influence expression evolution in plants, we performed a comparative transcriptomics analysis with large-scale expression data from Arabidopsis and rice. ECC represents an unsupervised approach to systematically compare gene expression networks between two species, including a statistical framework to quantify coexpression conservation. Although this approach does not directly compare expression profiles from identical experiments between species (5), comparison of coexpression clusters provides a valuable alternative to studying expression between less and more distantly related species lacking perfectly matched data sets. In contrast to two recent methods that also compare conserved expression patterns to identify functional analogy among homologous genes (50, 95), the presented ECC method includes different null models to reliably estimate the significance levels of conserved coexpression that control network properties, such as connectivity or tissue specificity. Overall, conserved coordinated expression occurred in 77% of the analyzed genes of Arabidopsis and rice. This degree of expression similarity is higher than the 60% reported for Arabidopsis and Populus orthologs that have broadly similar expression patterns during leaf development (66). Whereas the set of genes with expression conservation delineated here is robust when alternative models are used to estimate significance levels (77% and 75% ECC conservation with the connectivity and tissue-specific null models, respectively), the study of Street et al. (66) monitored conservation of leaf expression in the absence of a statistical framework. Although only genes with orthologs in both species have

been retained, almost 25% of all analyzed genes had no significant coexpression conservation, indicating that a substantial fraction of coexpression links have been rewired during plant evolution.

The observed conservation of many developmental regulators generates valuable information for the transfer of biological knowledge between different species in plant biotechnology. In contrast, many genes that respond to a specific stress stimulus or are involved in signal transduction had low conservation levels, suggesting a link between regulatory evolution and adaptation to lifestyle or environment. As the degree of coexpression conservation varied for different functional categories, the relationship between protein evolution (Ka), tissue specificity (τ), and ECC was analyzed in more detail considering different GO categories (Supplemental Table S5 (See section 3.7)). Notably, several GO terms deviated significantly from the general trends. Genes evolving rapidly at the expression level but slowly at the protein level (low fraction of ECC conserved genes and low Ka) include categories such as cellular response to extracellular stimulus, oxygen and reactive oxygen species metabolism, response to salt stress, and mitochondrial and carbohydrate transmembrane transport. Expression divergence of duplicate genes under environmental stress has also been found to be significantly greater than that under developmental stress (124). In addition, the levels of expression divergence between gene duplicates were the highest in extracellular transport, signal transduction, stress response, and transcription and were the lowest in the cellular and developmental processes, such as energy pathway, protein metabolism, intracellular transport, DNA and RNA metabolism, and cell organization and biogenesis. These functional categories are highly congruent with the ECC results (Figure 3.4) and indicate that expression divergence in response to external processes not only acts on recent duplicates but also on orthologous genes that have been present for hundreds of millions of years within the genomes of flowering plants. Moreover, both the diverged and conserved gene functions highlight the robustness and stochasticity of gene regulatory networks in the control of gene expression. Ninety-three percent of a set of 147 essential Arabidopsis embryo-defective genes (125) were ECC conserved, confirming the robustness of regulatory developmental programs across species (126).

Although it is plausible to assume that conserved expression contexts are the output of ancestral regulatory interactions that have been conserved in extant species (94), the

absence of large-scale TF-target data makes it difficult to construct genome-wide regulatory networks in plants and to directly study their evolution. Therefore, information about known cis-regulatory elements was integrated to annotate coexpression contexts. Tight regulatory promoter conservation (i.e. five or more conserved motifs) explains the observed coexpression conservation for 41% of all analyzed genes, revealing a 2-fold enrichment compared with a control data set. Corroborating the relationship between expression evolution and adaptation, several motifs enriched only in the context of Arabidopsis have been annotated as responsive elements, suggesting that different cis-regulatory elements and/or other TFs regulate these genes in each organism. The contribution of cis-regulatory elements to conserved transcriptional control should be interpreted cautiously, because the knowledge of plant promoter elements is far from complete.

Although "ancient" core plant genes with a broad phyletic distribution displayed a small, but significant, increase in the fraction of ECC conserved genes (80% versus 77% considering all genes), we found no strong evidence for the hypothesis that genes with conserved gene organization are highly conserved in expression (69% ECC conservation). These results imply that colinear genes between Arabidopsis and rice are, at the regulatory level, not more conserved than genes that have been rearranged since the divergence of both species (127). Despite the coexpression of neighboring genes in rice (65), currently there is little evidence that general coregulatory mechanisms, complementary to, for example, bidirectional promoters, act on a global scale in plant genomes. Altogether, our findings indicate that orthologs inferred through sequence similarity in many cases do not resemble similar biological gene functions and highlight the importance of incorporating expression information when homologous genes between different species are analyzed.

The combined information about tissue-specific gene expression (τ) and protein sequence evolution (Ka) indicates that high tissue specificity is linked with conserved ECC and high rates of protein evolution (Figure 3.9). Reversely, this pattern suggests that genes expressed in many tissues or conditions, potentially with a pleiotropic function, are strongly constrained at the sequence level. Network connectivity analysis confirmed that genes expressed coordinately with many other partners have more conserved patterns of expression evolution. A large-scale comparative expression study in mammals confirmed that highly tissue-specific genes tend to evolve rapidly at the

sequence level but slowly at the expression level (40). The high divergence at the protein level seems to coincide with the assumption that tissue-specific genes have less pleiotropic effects (128, 129), whereas carefully tuned expression of, for example, tissue-specific TFs is essential for cell differentiation and the proper execution of developmental programs. The functional importance hypothesis (130), in which highly expressed genes are functionally more important and therefore more conserved in their coding sequences, corresponds with the high levels of sequence conservation for genes with low τ values. Also in yeast, genes with many protein-protein interactions or functional protein sites negatively correlate with coding-sequence divergence, confirming the strong constraint to maintain physical interactions (131). We observed no significant correlation between protein and expression evolution, indicating that both modes of gene evolution are not coupled in plants. Similarly, after comparison of expression similarity with coding sequences for more than 10,000 human-mouse orthologs, no strong correlation between expression and sequence evolution was found in mammals (40). Although contradictory conclusions have been drawn in different species (40, 132, 133, 134, 135), also in yeast no evidence has been found for a strong correlation between these two modes of gene evolution (131).

Changes in the expression pattern of a tissue-specific gene will have important consequences for the interacting genes and the biological function of the underlying network. Complementary with a report of concerted expression divergence after large-scale duplications in Arabidopsis (136), we have identified more than 20 genes in which the expression contexts are conserved and tissue specificity is not. Although in the case of gene duplication, evolution might work on the initially redundant version of a specific pathway or a set of genetically interacting genes, the pattern of concerted expression evolution[1] for different single-copy tissue-specific markers in Arabidopsis and rice is intriguing. In yeast, cross-species promoter analysis has shown how the regulation of ribosomal proteins evolved via intermediate redundant programs in which the concurrent emergence of cis-regulatory elements was followed by the loss of more ancient elements (137). A similar mechanism might provide a mechanistic explanation for the concerted tissue-specific divergence (Table 3.1), but the knowledge about regulatory control in plants is currently insufficient to determine the role of cis- versus

[1]Concerted evolution is the tendency of the different genes in a gene family or cluster to evolve in concert. This means that each gene locus in the family comes to have the same genetic variant.

trans-regulatory changes in this network rewiring. The developed ECC framework provides a practical approach to compare expression patterns and molecular phenotypes between species. Nevertheless, the detailed characterization of target genes for orthologous regulators in different species will be required to obtain a more detailed view of the regulatory evolution of signaling networks and the rate of expression evolution in different plant families.

3.5 Materials and Methods

3.5.1 Expression Data

We obtained 203 and 322 Affymetrix CEL files for rice (*Oryza sativa*) and Arabidopsis (*Arabidopsis thaliana*), respectively, from the National Center for Biotechnology Information Gene Expression Omnibus database monitoring the transcriptional activity in different tissues and developmental stages (75, 138, 139), root cell types (140, 141), and under different stress conditions (142, 143, 144). The ATH1 microarray slides were processed with a custom-made CDF measuring 19,937 genes as described before (85). Briefly, all raw data were processed with the robust multichip analysis algorithm implementation (58) in Bioconductor and with custom-made CDFs (background adjustment, quantile normalization, and finally summarization). A remapping of all 631,066 rice probes to the rice gene models (The Institute for Genome Research 5) showed that 43% of the original probe set (as defined by Affymetrix CDF ricecdf) contained one or more cross-hybridizing probes (S. Movahedi and K. Vandepoele, unpublished data). These cross-hybridization effects could lead to significant errors, especially under conditions where changes in expression are not dramatic. For all microarray analyses, it is important to ascertain that a probe really measures the intended gene. Therefore, this inaccuracy in the original rice Affymetrix probe set urged us to design a new set of probe set (or CDF)- containing probes unique at the gene level, as described for Arabidopsis (108). The new rice gene CDF allowed us to reliably measure expression levels for 32,004 rice gene models and is available upon request. The mean intensity values were calculated for the replicated slides. To identify and remove redundant experiments, we first clustered all experiments with hierarchical clustering (considering all genes and the PCC as a measure) and retained only one experiment as representa-

tive for a set of similar experiments (38).

3.5.2 Clustering of Expression Data

Expression similarities between gene pairs (per species) were calculated with the PCC. To identify coexpressed genes, a similarity threshold was determined for both expression compendia. Based on the similarity between expression profiles for 1,000 random genes (approximately 1,000 × 999 × 0.5 gene pairs), a PCC threshold of 0.41 and 0.48, corresponding with the 90th percentile of this distribution, was set for rice and Arabidopsis, respectively. Although the absolute PCC threshold differed for both species (because of differences in size and composition of the expression data set), these percentile-based thresholds returned a similar relative cutoff for both species. Application of different percentile thresholds to predict functional enrichments with gene-centric coexpression clusters for genes with known annotations (excluding electronic GO functional assignments) showed that the 90th percentile threshold recovered the largest number of correct gene functions (results for Arabidopsis/rice with a random set of 100 reference genes are as follows: 85th percentile, 32%/22% correct predictions; 90th percentile, 33%/24% correct predictions; 95th percentile, 33%/23% correct predictions).

The EC for a set of N genes was calculated as the fraction of all possible N × (N - 1) × 0.5 gene pairs with a PCC higher than the threshold value defined for that species. To determine the stability of the coexpression on the removal of subsets of experiments from the original expression compendium, a jackknife procedure was applied. Based on 100 iterations, randomly 25% and 50% of the original experiments were removed, percentile-based PCC thresholds were defined for the retained experiments, and the EC per GO category was calculated.

To create gene-centric coexpression clusters, each gene was considered as a seed gene and all genes with a PCC value greater than the determined PCC threshold were assigned to the cluster. Therefore, the number of clusters is equal to the number of genes available in the expression data set, and clusters are potentially overlapping on a genome-wide scale. Data files containing the Arabidopsis and rice gene coexpression clusters as well as the orthologous gene families are available at http://bioinformatics.psb.ugent.be/supplementary_data/Movahedi_ECC.

3.5.3 Identification of Orthologous Gene Families

To identify orthologous genes, we clustered similar protein sequences from rice and Arabidopsis by means of the OrthoMCL algorithm starting from an all-against-all sequence similarity search with BLASTP (E < 1e-05, with a default inflation factor of 1.5). Arabidopsis and rice protein sequences were downloaded from The Arabidopsis Information Resource 7 and The Institute for Genome Research 5, respectively. Among 33,195 families, 7,911 contained at least one gene from each species also present in the expression data. These families covered 12,019 rice genes and 12,419 Arabidopsis genes and could be subdivided into three subcategories: 1:1 orthologous groups (both species contained one gene; 4,630 families), 1:n groups (one of the species contained a single gene and the other contained multiple genes; 2,194 families with 3,748 and 4,031 rice and Arabidopsis genes, respectively), and n:m groups (n,m > 1; 1,087 families with 3,641 and 3,758 rice and Arabidopsis genes, respectively). A list of Arabidopsis gene symbols, Arabidopsis Genome Initiative locus identifiers, and orthologous rice genes is available in Supplemental Table S6 (See section 3.7). Inparalogs are gene duplicates postdating speciation. Recent benchmark studies have shown that the OrthoMCL gene-clustering method performs well in modeling recent and ancient duplication events compared with more advanced methods based on phylogenetic inference (89, 145). For every functional category (see below), all genes with one or more orthologous gene in the other species were retained for the EC analysis. Genes from 1:1 orthologous families were used as seeds to study the ECC (retaining all genes with orthologs in both plants to calculate the gene-centric coexpression clusters). Because for both species not all genes are present on the Affymetrix microarrays, the reported ECC scores are not exact but represent estimates based on the orthologs available in the CDFs.

3.5.4 Functional Annotation

Genes and orthologous gene families were functionally annotated with GO and Reactome terms. Whereas GO labels were retrieved from the GOWeb site (versions of July 21, 2009, for Arabidopsis and April 18, 2009, for rice), Reactome data were downloaded from `http://arabidopsisreactome.org/` (99). The gene-GO annotations were extended to include parental GO terms (i.e. a gene assigned to a given GO category was automatically assigned to all the parent categories as well) by propagating all GO

annotations up to all possible edges of the GO graph (with the Perl GO::Parser and GO::Node modules). GO annotations were assigned to all 1:1 orthologous gene pairs starting from the original gene-GO annotation files. ECC scores for different functional categories excluding electronic evidence tags (IEA, ISS, NAS, NR, and ND) are reported in Supplemental Table S7 (See section 3.7). To calculate the EC for different GO categories (Figure 3.1; Supplemental Table S2 (See section 3.7)), the original gene annotations were used.

The statistical significance of the functional enrichment for GO and MAPMAN (146) annotations was evaluated with the hypergeometric distribution adjusted by the false discovery rate correction for multiple hypotheses testing by means of the p.adjust stats package in R (Benjamini and Hochberg, 1995). Corrected values of $P < 0.05$ were considered significant. MAPMAN results are reported in Supplemental Table S8 (See section 3.7). Colinear regions between Arabidopsis and rice were computed with i-ADHoRe and retrieved from PLAZA, a comparative genomics resource (89).

3.5.5 Calculation ECC Scores

In a first step, for all genes a set of coexpressed genes was defined (genecentric cluster) by retrieving the neighboring genes in the coexpression network. Next, for gene-centric clusters from 1:1 orthologs, the number of orthologs was determined with coexpression in both species (Figure 3.1). The overlap of conserved shared orthologous families in both coexpression clusters was quantified with the Jaccard Index, which measures the ratio of the number of shared families over the total number of families found in the two coexpression clusters. A permutation test was applied to determine whether the observed ECC score for that 1:1 orthologous gene pair could be considered conserved, diverged, or not significant (i.e. significantly larger, smaller, or not different compared with a background distribution of ECC values [$P < 0.05$]). For each 1:1 orthologous gene pair A and R, 1,000 pairs of gene-centric clusters were randomly sampled, maintaining the gene-centric cluster sizes (A) and (R), respectively. The expected ECC scores, reflecting a neutral null model of expression evolution, were used to estimate a P value for coexpression conservation or divergence. Two null models were used, one controlling tissue-specific expression and one correcting the degree distribution in the network (or connectivity) (see section 3.6). One model controlled the expression breadth (with τ) in determining the ECC significance, and the other controlled the

degree distribution (or connectivity) of the different genes in the network. Briefly, during the permutation-based significance estimation using these models, we randomly sampled genes with properties similar to those of the genes present in the real network (i.e. expression breadth or connectivity). First, all genes were grouped in 50 bins based on τ (or connectivity); subsequently,we sampled from a specific bin to control expression breadth (or the number of coexpressed genes). Orthologous pairs with higher (or lower) Jaccard Index than that of the random values were considered as conserved (or diverged), and genes not belonging to any of these categories were classified as non-significant (Figure 3.5). To correct for multiple comparisons, again the false discovery rate method was applied with p.adjust. Classification of the Jaccard Index with other percentile-based PCC thresholds for expression similarity yielded highly similar results.

3.5.6 Calculation of Ka and Ks

The coding sequences for the gene pairs were aligned with ClustalW version 1.83 (147) with the protein sequences as alignment guides. From this alignment, unambiguously aligned positions were retained for further analysis (89). Synonymous and nonsynonymous substitutions (Ks and Ka) were estimated with codeml (part of the PAML package; (148)) with the following parameter settings: verbose 0, noisy 0, runmode -2, seqtype 1, model 0, NSsites 0, icode 0, fix_alpha 0, fix_kappa 0, and RateAncestor 0. Note that values of Ks > 5 should be considered as unreliable.

3.5.7 Statistical Analysis

Differences between ECC conserved and nonconserved gene sets were tested with the nonparametric Mann-Whitney U test (also known as the Mann-Whitney-Wilcoxon test) with the "wilcox.test" function in the "stats" R package version 2.9.1. To determine whether the fraction of ECC conserved genes was significantly larger for a specific functional category than that of the overall ECC conservation considering all genes, the hypergeometric distribution was used (Supplemental Tables S4 and S6 (See section 3.7)). Differences between Ka (or τ) values between all 1:1 orthologs and genes annotated with a specific functional category were also evaluated with the Mann-Whitney U test, as were differences between Ka (or τ) values for ECC conserved and nonconserved gene sets within a specific functional category (Supplemental Table S5 (See section 3.7)).

		Tau null model			
		Conserved	Diverged	Non-significant	total
Connectivity null model	Conserved	3461	0	97	3558
	Diverged	0	385	8	393
	Non-significant	18	110	551	679
	total	3479	495	656	4630

Table 3.2: Comparing Connectivity and Tau (τ) null models

3.6 Supplementary Notes

3.6.1 Connectivity null model and Tau (τ) null model

ECC scores for all 4630 1:1 orthologous rice - Arabidopsis gene pairs revealed that, using connectivity null model, 3558 (77%) pairs had a conserved expression context, whereas 393 (8.5%) and 679 (14.5%) pairs had a diverged and non-significant ECC, respectively.

ECC scores for all 4630 1:1 orthologous rice - Arabidopsis gene pairs revealed that, using Tau (τ) null model, 3479 (75%) pairs had a conserved expression context, whereas 495 (11%) and 656 (14%) pairs had a diverged and non-significant ECC, respectively (see table 3.2 for a comparision with connectivity null model).

3.6.2 ECC control experiment 1: Using a different set of experiments

We determined for the same set of genes the ECC scores using two other expression compendia (now including 491 Arabidopsis and 80 rice non-redundant experiments). Overall, we observed that 90.4% of the genes initially scored as ECC conserved received the same classification using these alternative expression compendia, confirming the reliability of the reported ECC results.

3.6.3 ECC control experiment 2: Random removal of expression cluster members

We performed an additional control experiment where we calculated ECC based on coexpression clusters with random removal of 12% of the genes in 100 iterations.

Initially, 100 random 1:1 orthologous gene pairs were picked and to keep the distribution the same as connectivity null model, these 100 random pairs include 77 conserved, 8

diverged and 15 non-significant random pairs. In the next step, for each selected pair, we randomly removed 12% of the cluster members and recalculated the ECC score for 100 iterations. Comparing these results to the original ECC scores, we found out on average in 96.3% of the cases the same ECC category was obtained, confirming the robustness of our approach.
In conclusion, since the missing genes could have a small effect on the reported ECC scores, the reported ECC scores are not exact because some orthologous genes might be missing from the CDF files.

3.7 Additional data files

The following additional data are available with the online version of this paper or at `http://bioinformatics.psb.ugent.be/supplementary_data/samov/apr2011/`.

Tables (Movahedi_et_al_STables.xls)
Supplemental Table S1. Overview of microarray experiments included in the expression compendia for Arabidopsis and rice
Supplemental Table S2. Expression Coherence (EC) for GO categories with five or more genes
Supplemental Table S3. ECC distributions for all 1:1 orthologous genes
Supplemental Table S4. Distributions of ECC conserved, diverged, and non-significant genes for different GO and Reactome categories
Supplemental Table S5. ECC conservation, and Ka for Arabidopsis genes in different functional categories
Supplemental Table S6. Overview 1:1 orthologous families
Supplemental Table S7. Distributions of ECC conserved, diverged, and non-significant genes for different GO categories, excluding electronic annotations
Supplemental Table S8. MAPMAN enrichment for different gene sets

Supplementary Data
Arabidopsis expression clusters (tab-delimited compressed file)
Rice expression clusters (tab-delimited compressed file)
Arabidopsis:Rice orthologous families (tab-delimited compressed file)

3.8 Acknowledgments

We thank Martine De Cock for help in preparing the manuscript. Received April 8, 2011; accepted May 11, 2011; published May 13, 2011.

3.9 Author Contribution

This chapter is redrafted from an article published in Plant Physiology. The text is written by me and Klaas Vandepoele, with several helpful suggestions from Yves Van de Peer.

4

Comparative analysis of tissue specific gene networks in seven plant species

4.1 Abstract

The availability of gene expression data for different tissues enables us to identify the genes which are specifically expressed in one tissue or only a few tissues. However the accuracy of defining tissue specific genes using only publicly available data expression profiles, depends widely on the number and variety of experiments and the applied identification method. In this study a new approach is presented to define high quality tissue specific genes in seven different plant species; *A.thaliana* (Arabidopsis), *Z.mays* (Maize), *M.truncatula* (Medicago), *P.trichocarpa* (Poplar), *O.sativa* (Rice), *G.max* (Soybean) and *V.vinifera* (Grape) using Affymetrix microarray expression profiles which cover over 4500 experiments and 135700 genes. A combination of mathematical approaches and statistical models provide us with a final set of 70427 tissue specific genes in 40 different tissues. A down stream analysis on the relationship between leaf tissue specific genes coexpression clusters, within a species and in comparison with other species, for a total number of 11108 strictly selected genes is also performed, which is discussed in detail.

4.2 Introduction

Animals, plants and also some fungi are differentiated by tissue types. Each tissue has its specific composition and function, therefore tissue specific genes has always been a study of interest. So far there have been many studies done on tissue specific genes each focusing on a different aspect. Although many of these studies consider a single species, recently a lot of comparative interspecies research have been also conducted. It has been shown that the gene expression and tissue specificity determine the rate of gene sequence evolution (128, 129, 149, 150). Liao et al. investigated a negative correlation between the rate of gene (or protein) sequence evolution and the gene expression level. They also found that tissue specific genes evolve slowly at the expression level while their evolution is rapid at the sequence level (151). Housekeeping genes have always been an interesting case of study since they often express in almost every tissue, but do they evolve slower than tissue specific genes? To answer this question Zhang et al performed a study on sequence evolution of 1581 human-mouse orthologous gene pairs which their expression profiles where available for brain, kidney, muscle, prostate, lung, liver, and vulva tissues. Comparing tissues specific genes with housekeeping genes, they observed significantly smaller values of Ka/Ks for housekeeping genes and therefore concluded that housekeeping genes evolve more slowly than tissue specific genes and undergo stronger purifying selection (149). The results of Zhang confirmed an earlier study which was performed by Duret et al. on the effect of gene expression patterns on mutation rates and selection intensity in mammalian genes. Duret et al. analyzed 2,400 human/rodent and 834 mouse/rat orthologous genes, measured the relationships between substitution rates and tissue distribution of gene expression in 19 tissues from three development states and revealed that tissue specific genes are evolved faster at protein level than broadly expressed genes such as housekeeping genes (129). Around the same time, Pal et al. analyzed a combination of DNA sequence, gene expression and genome duplication in yeast *Saccharomyces cerevisiae* and concluded highly expressed genes evolve slowly at protein level in yeast (150). Odom et al. studies tissue specific genes from another aspect. They investigate conserved transcription factor binding sites of tissue specific genes between human and mouse by mapping the binding of four tissue-specific transcription factors involved in liver development and regulation

to 4000 orthologous pairs of mouse and human genes. Results show a significant divergence between human and mouse tissue-specific transcriptional regulation (152).
Recently with the huge amount of gene expression experiments there are thousands of experiments available for a single species, for instance in this study there are 2417 experiments and 1417 experiments available for Arabidopsis and soybean respectively. Therefore it happens that a tissue specific gene is highly expressed in many leaf tissue experiments but no other experiment, while on the other hand a gene can be highly expressed in a few experiments but from different tissues. Considering this issue and other similar problems, in the proposed method many different factors have been taken in to account in order to increase the quality of defined tissue specific genes. We also consider the cases which a gene has low expression breadth (i.e. expressed in a few tissues) but on the other hand is highly expressed in different tissues.
To summarize, in this study a new approach is proposed to first define best tissue specific gene candidates starting from Affymetrix gene expression profiles and then leaf tissue specific genes are selected for an in-depth study. For instance an improved version of ECC method is undertaken to determine if any evidence of coexpression conservation can be observed among tissue specific genes.

4.3 Results

In our attempt to study tissue specific genes it is important to first define a list of reliable tissue specific genes and then use this data to continue with down stream analysis such as studying the distribution of tissue specific genes over orthologous families or performing an in-depth study on a specific tissue category (e.g. leaf tissue specific genes).

4.3.1 Detection of tissue specific genes

Function and expression of tissue specific genes are preferred in only one tissue. Identification of these genes helps better understanding of their role in an organism. It is also interesting to compare tissue specific genes in different organisms which is the main idea of this work. Here, τ (Tau) measure is used to quantify the level of differential expression across conditions (see section 4.5.7). τ ranges from 0 to 1 and high values

correspond with high gene expression in a few (different or similar) tissues.
In order to define tissue specific genes three steps have been applied:

1. Which **genes** are tissue specific? Step 1. Tau values
2. Which **tissues** are specific? Step 2. Z-Scores
3. **Final set** of tissue specific genes? Step 3. Filter genes depending on the count of experiments/tissues

Here we explain each step in more details:

Step 1. Gene selection: defining a tau threshold using random expression profile

At the first step we want to select a primary set of tissue specific genes. Therefore for each gene τ values is calculated to measure the level of differential expression across conditions. Then a threshold is needed to select higher τ values which correspond with high expression in a few tissues. Trying to set a threshold for τ, expression profile of each species is first shuffled and then tau is calculated for this new expression profile. The 99% threshold obtained from these new τ values is applied on original τ values (see figure 4.1). This results in a primery set of 3904 tissue specific genes (See table 4.1).

Step 2. Experiments selection: applying z-score measure

In this step for the list of selected genes in step.1, which we are pretty confident on their tissue specificity, we try to define the type of tissue specificity. Indeed we need a method which can be applied on this large number of genes considering that each species has different number of experiments. It is expected that each of the selected genes differentially express in only one experiment or a few experiments of the same tissue. It is possible to select the highest z-value and define the tissue specificity but the study of hundreds of random genes showed us in many cases a tissue specific gene is differentially expressed in more than one experiment, although these few experiments mostly present the same tissue. Consequently instead of picking the highest expression value we use z-score measure (see section 4.5.8) to fairly select most of the highly expressed experiments. For each of the selected genes, z-score of its expression values

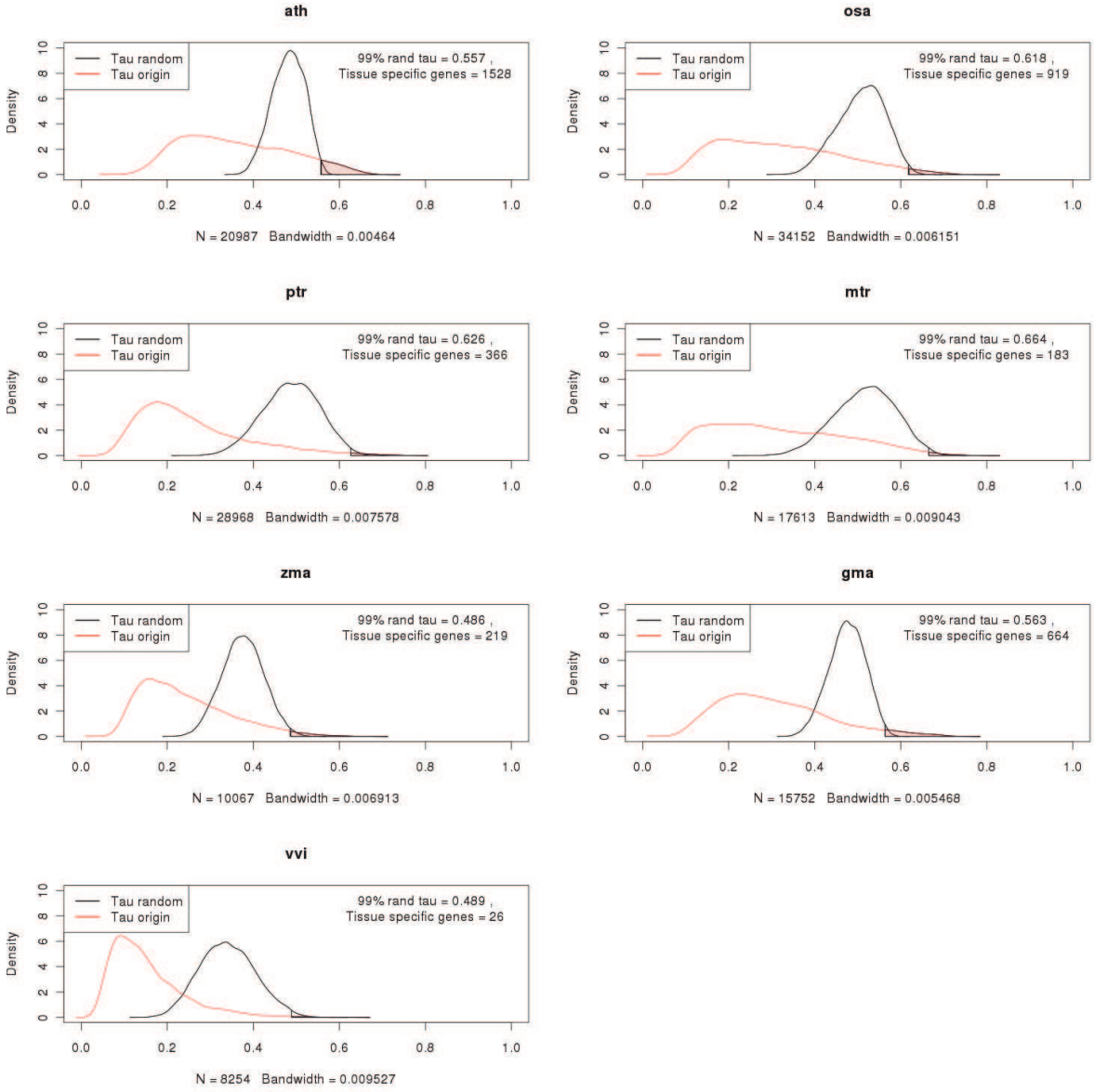


Figure 4.1: Defining tau values threshold - This plot illustrates how the 99th percentile of tau values generated from a shuffled expression profile is applied as a threshold on original tau values for each of the seven species. Red lines represent tau values distribution in original expression profile and black lines indicate tau values distribution in shuffled expression profile. In the top right corner of each plot you can see the tau value corresponding to 99th percentile of the tau value distribution in the shuffled expression profile. Applying this tau value as a threshold on original tau values will result in a number of tissue specific genes which is also indicated in the top right corner of each plot. N is the number of values in the data set and bandwidth is the default smoothing parameter.

is calculated (see figure 4.2 for a sample comparing raw expression values and z-core transformed expression values of a poplar leaf tissue specific gene (PT19G04750)), and then experiments with z-score higher or equal to two (or 97.7% of cumulative percentage) and z-score higher than maximum z-score divided by two[1], are selected:

$$(z\_score \geq 2)\ AND\ (z\_score > (maximum\ z\_score/2)) \tag{4.1}$$

The thresholds applied on z-score are defined after analyzing hundreds of randomly selected genes from different species. The definition of expression 4.1 is applicable on expression profiles with experiment range of 27 (for grape) to 491 (for Arabidopsis). The more experiments are available the more accurate predictions are made.

Selected experiments using this method can define the type of each gene tissue speci-

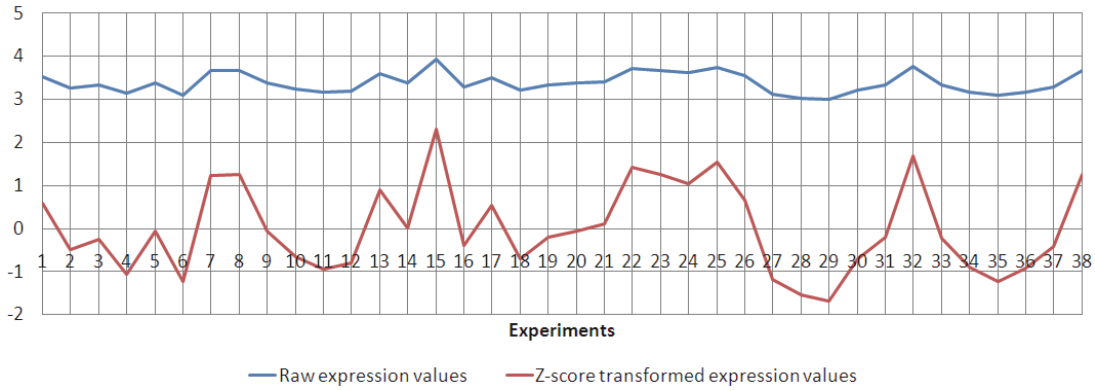


Figure 4.2: z-score transformed expression values - Raw expression values of a poplar leaf tissue specific gene (PT19G04750) are shown in blue and standardized (z-score transformed) values are shown in red. Following the method explained in section 4.3.1 this gene is highly expressed in 15th experiment (GSM377365.CEL-GSM377366.CEL-GSM377367.CEL) which is a leaf tissue specific experiment.

ficity but if the selected experiments predict many unrelated tissues for a gene that gene is out of our interest. It does not mean that these cases are not tissue specific, because already a strict threshold has been applied on τ values, it just means for these genes a more stringent method is needed to accurately define their type of tissue specificity. This issue is addressed in the next step.

[1]This condition is useful for the expression profiles with large number of experiments

Species	Step 1	Step 2	**Step 3**
A. thaliana (Arabidopsis)	1528	1528	**1067**
Z. mays (Maize)	219	219	**219**
M. truncatula (Medicago)	183	183	**183**
P. trichocarpa (Poplar)	366	366	**366**
O. sativa (Rice)	919	919	**902**
G. max (Soybean)	664	664	**386**
V. vinifera (Grape)	25	25	**25**

Table 4.1: Number of tissue specific genes after applying each step (see section 4.3.1).

Step 3. Selection of final tissue specific gene sets

In the last step we want to identify and remove the genes which we could not predict a clear type of tissue specificity for.

To achieve this goal, we apply one or more labels on each experiment depending on its description available at NCBI GEO database. These labels contain 40 keywords such as shoot, leaf, pollen, root, seed, flower, etc. An in house Perl script runs over thousands of experiment descriptions and points out the labels. Detected labels can provide useful information for each experiment which helps defining type of tissue specificity for each tissue specific gene, but on the other hand they can including some misleading information as well. For example for an experiment description like "leaf tissue of a flowering plant" two different labels are detected, "leaf" and "flower", while only "leaf" label shows the real tissue type. In order to reduce the amount of unclear predictions we only keep tissue specific genes with less than or equal to six specific experiments (see step.2) or less then or equal to five key labels (see expression 4.2).

$$(experiments.num \leq 6)\ OR\ (labels.num \leq 5) \tag{4.2}$$

A study on the distribution of tissue specific genes over different number of experiments and labels showed us these thresholds result in the most reasonable number of tissue specific genes in total and in each species (see figure E.1, Appendix E).

After the final filtration in step.3, 70427 tissue specific genes are defined (See table 4.1 for the number of tissue specific genes in each organism). In table 4.2 it is shown for each label how many tissue specific genes are defined and these labels come from how many different experiments.

Tissue labels	Tissue specific genes	Experiments	Num.Species	Species
aleurone	102	1	1	GM
caryopsis	580	1	1	OS
catkin	5690	1	1	PT
seedling	89	1	1	AT
coleoptiles	1354	2	1	OS
columella	471	2	1	AT
mycelia	607	2	1	GM
plumule	483	2	2	GM,OS
vascular	98	2	1	GM
endothelium	119	3	1	GM
pulvinus	414	3	1	ZM
stele	98	3	1	AT
stigma	1659	3	2	AT,OS
xylem	3086	3	2	AT,PT
zoospores	704	3	1	GM
endodermis	108	4	1	AT
hilum	219	4	1	GM
phloem	326	4	1	AT
parenchyma	324	5	2	AT,GM
stamen	3578	5	2	AT,OS
suspensor	684	5	2	AT,GM
anther	4060	6	2	AT,OS
integument	354	6	1	GM
epidermis	373	8	2	AT,GM
panicle	9274	10	1	OS
inflorescence	1181	11	2	AT,OS
pollen	3143	11	3	AT,GM,OS
hypocotyl	1197	12	2	AT,GM
apex	2367	15	4	AT,GM,PT,ZM
meristem	1339	15	4	AT,GM,MT,ZM
embryo	2325	16	4	AT,GM,OS,ZM
rosette	769	18	1	AT
flower	3134	31	4	AT,GM,MT,OS
endosperm	5167	32	4	AT,GM,OS,ZM
stem	9997	32	6	AT,GM,MT,OS,PT,ZM
cotyledon	2282	38	2	AT,GM
shoot	8525	60	6	AT,GM,OS,PT,VV,ZM
root	16675	134	6	AT,GM,MT,OS,PT,ZM
leaf	37534	155	7	AT,GM,MT,OS,PT,VV,ZM
seed	14148	166	6	AT,GM,MT,OS,PT,ZM

Table 4.2: The number of defined tissue specific genes for each label and the number of experiments presenting these labels.

4.3.2 Comparing defined tissue specific genes with available datasets

The method introduced in section 4.3.1 provided us a final set of 1067 Arabidopsis tissue specific genes. Here we compare this set of Arabidopsis tissue specific genes with Schmid et al. (75) list of Arabidopsis tissue specific genes which is based on a study in 2005. Their dataset contain 204 Arabidopsis tissue specific genes covering eight different tissues. Among these genes, 189 genes are present in our expression profile and 81 genes overlap with our defined Arabidopsis tissue specific genes. Comparing the type of tissue defined for these 81 genes shows 79% of our predictions (64 genes) match with Schmid et al. results (see table 4.3 for details).

Gene ID	Schmid tissue	Movahedi tissue	Overlap
AT1G01280	flower	anther,flower,leaf	1
AT1G03890	seed	embryo,seed,leaf,endosperm	1
AT1G06990	flower	flower,stamen,leaf	1
AT1G07340	flower	pollen,anther,flower	1
AT1G08065	flower	pollen,anther,flower,endosperm,stamen	1
AT1G16770	seed	embryo,seed	1
AT1G17810	seed	embryo,leaf,seed	1
AT1G18970	root	root,seed	1
AT1G21860	pollen	seed,leaf	0
AT1G21890	stem	hypocotyl,inflorescence,root	0
AT1G21970	seed	suspensor,cotyledon,seed,endosperm	1
AT1G22600	seed	endosperm,leaf,seed	1
AT1G22760	flower	pollen,flower,anther,stamen	1
AT1G25270	seed	cotyledon,seed	1
AT1G25330	flower	stigma,stem,meristem,inflorescence,flower	1
AT1G26250	root	rosette,root,leaf,seed	1
AT1G26710	flower	inflorescence,flower	1
AT1G28375	flower	inflorescence,flower,stamen	1
AT1G30795	flower	stigma,stem,root,flower,meristem,inflorescence	1
AT1G48910	seed	shoot,seed,endosperm	1
AT1G48940	flower	pollen	0
AT1G50650	seed	cotyledon,endosperm,seed	1
AT1G53010	pollen	stigma,pollen,root,columella	1
AT1G54240	flower	stigma,pollen,anther,leaf	0
AT1G56360	flower	pollen,flower	1
AT1G62210	seed	seed	1
AT1G63520	stem	xylem,stele,root,seed,leaf	0

AT1G64590	root	root,seed	1
AT1G68380	seed	cotyledon,seed	1
AT1G70720	flower	inflorescence,flower,stamen	1
AT1G72100	seed	embryo,endosperm,seed,leaf	1
AT1G72290	flower	stigma,stem,meristem,inflorescence,flower,leaf	1
AT1G80090	seed	embryo,leaf,seed	1
AT1G80330	seed	cotyledon,endosperm,seed	1
AT2G01920	pollen	stigma,pollen	1
AT2G17090	pollen	stigma,pollen	1
AT2G23550	seed	cotyledon,seed	1
AT2G42000	seed	seed	1
AT2G42860	seed	cotyledon,seed	1
AT3G05260	seed	seed,endosperm,leaf	1
AT3G05920	root	seedling,endodermis,root,seed,leaf	1
AT3G08490	stem	xylem,stele,root,seed	0
AT3G13820	pollen	stigma,pollen,root,leaf	1
AT3G22500	seed	seed	1
AT3G23770	flower	flower	1
AT3G43160	senescence	root,leaf	0
AT3G49450	pollen	stigma,pollen	1
AT3G50170	apex	shoot,root,columella	0
AT3G52310	senescence	root,flower,columella	0
AT3G52810	flower	pollen,flower,anther,stamen	1
AT3G57620	flower	pollen,anther,flower,leaf	1
AT3G58740	seed	cotyledon,endosperm,seed	1
AT3G60730	seed	embryo,seed,endosperm	1
AT3G62280	root	shoot,root,epidermis	1
AT4G07820	root	seedling,root,seed	1
AT4G22100	seed	seed	1
AT4G23290	leaf	shoot,rosette,root,seed,leaf	1
AT4G26050	seed	root,seed	1
AT4G33870	flower	cotyledon,anther,leaf,endosperm	0
AT5G16460	seed	embryo,endosperm,seed,leaf	1
AT5G17200	flower	shoot,leaf	0
AT5G17830	flower	anther,leaf	0
AT5G20340	pollen	stigma,pollen	1
AT5G23960	flower	stigma,inflorescence,flower	1
AT5G26100	pollen	stigma,pollen	1
AT5G28470	flower	pollen,flower,stamen,leaf	1
AT5G35380	pollen	stigma,pollen	1
AT5G40940	flower	anther,root	0
AT5G41090	flower	pollen,anther,root,endosperm,leaf	0
AT5G41890	flower	anther,leaf	0
AT5G43340	flower	pollen,anther	0
AT5G44330	flower	anther,root,columella,leaf	0

AT5G48210	flower	inflorescence,flower,anther	1
AT5G49070	flower	flower,root,leaf	1
AT5G49130	flower	pollen,stigma,anther,flower,stamen	1
AT5G55020	pollen	stigma,pollen,phloem,parenchyma,inflorescence,leaf	1
AT5G55750	seed	seed	1
AT5G56310	pollen	stigma,pollen,root,leaf	1
AT5G61110	flower	shoot,anther,flower,leaf	1
AT5G61605	flower	shoot,inflorescence,flower,stamen	1
AT5G65550	seed	seed	1

Table 4.3: Comparison of Arabidopsis tissue specific genes with Schmid et al. dataset.

4.3.3 The distribution of tissue specific genes over orthologous families

In this study it is also investigated how the 70427 tissue specific genes (see Step.3 in 4.3.1) are distributed over the orthologous families. Results show 15816 out of 57910 orthologous families do not have any tissue specific genes (see supplementary table). The rest of the 42094 families show different patterns which is discussed following.

A few samples are shown in table 4.4. There exist families with more than one gene but not all of the genes are tissue specific, such as ORTHO017489 with three genes of the same species which only one gene is *seed* tissue specific, or ORTHO001428 with eight genes in seven different species which again only one gene is *stem* tissue specific. According to table 4.2 *stem* tissue experiment is present in the expression profile of six species, therefore the difference between tissue specificity of genes in ORTHO001428 family can not be due to absence of respective experiment.

On the other hand, there have been observed families like ORTHO005033 with four genes from four different species which three of the genes are *leaf* tissue specific. The only non tissue specific gene in this family can be diverged since there exist *leaf* experiments for all the species under study (see table 4.2).

In a family with more than one tissue specific gene, tissue specificity is expected to be conserved. As you see in the table 4.4 ORTHO008094 has three genes of the same species and all of the genes are tissue specific in the same tissue (*catkin*) or ORTHO016893 with three genes from two different species which all the genes are *endosperm* tissue specific, however the results show this is not always the case. For

	Tissue type						All		Tissue specific	
family ID	catkin	endosperm	leaf	root	seed	stem	species	genes	species	genes
ORTHO017489	0	0	0	0	1	0	1	3	1	1
ORTHO008094	3	0	0	0	0	0	1	3	1	3
ORTHO009609	0	0	0	1	0	1	2	2	2	2
ORTHO016893	0	3	0	0	0	0	2	3	2	3
ORTHO005033	0	0	3	0	0	0	4	4	3	3
ORTHO001428	0	0	0	0	0	1	7	8	1	1

Table 4.4: A few samples of tissue specific genes distribution in orthologous families. Last four columns indicate the number of species and genes present in a specific orthologous family and the proportion of tissue specific genes. The first six columns show the distribution of tissue specific genes of one family over different tissue categories.

instance ORTHO009609 has two genes from two different species which both are tissue specific but one of them is *root* tissue specific and the other one is *stem* tissue specific.

4.3.4 Study of leaf tissue specific genes

In this section leaf tissue specific genes of seven different plant species are studied for their coexpression, functional enrichment and coexpression conservation. As a study sample we are particularly interested in **leaf** tissue specific genes because earlier in this chapter we could identify thousands of leaf tissue specific genes which cover all the seven plant species of interest (see table 4.2). There are totally 37534 leaf tissue specific genes and 155 correlated leaf tissue experiments included in this data set.

4.3.4.1 List of pure leaf tissue specific genes

In order to improve the quality of leaf tissue specific genes in this study, we keep the genes which only show tissue specificity in leaf tissue and not any other tissue. The final set of under study leaf tissue specific genes has 11108 genes and 90 correlated leaf experiments (see table 4.5 for details). Figure 4.3 shows the expression heatmap of *Z.mays* (Maize) leaf tissue specific genes and correlated leaf tissue experiments. The GO functional enrichment of 11108 carefully selected leaf tissue specific

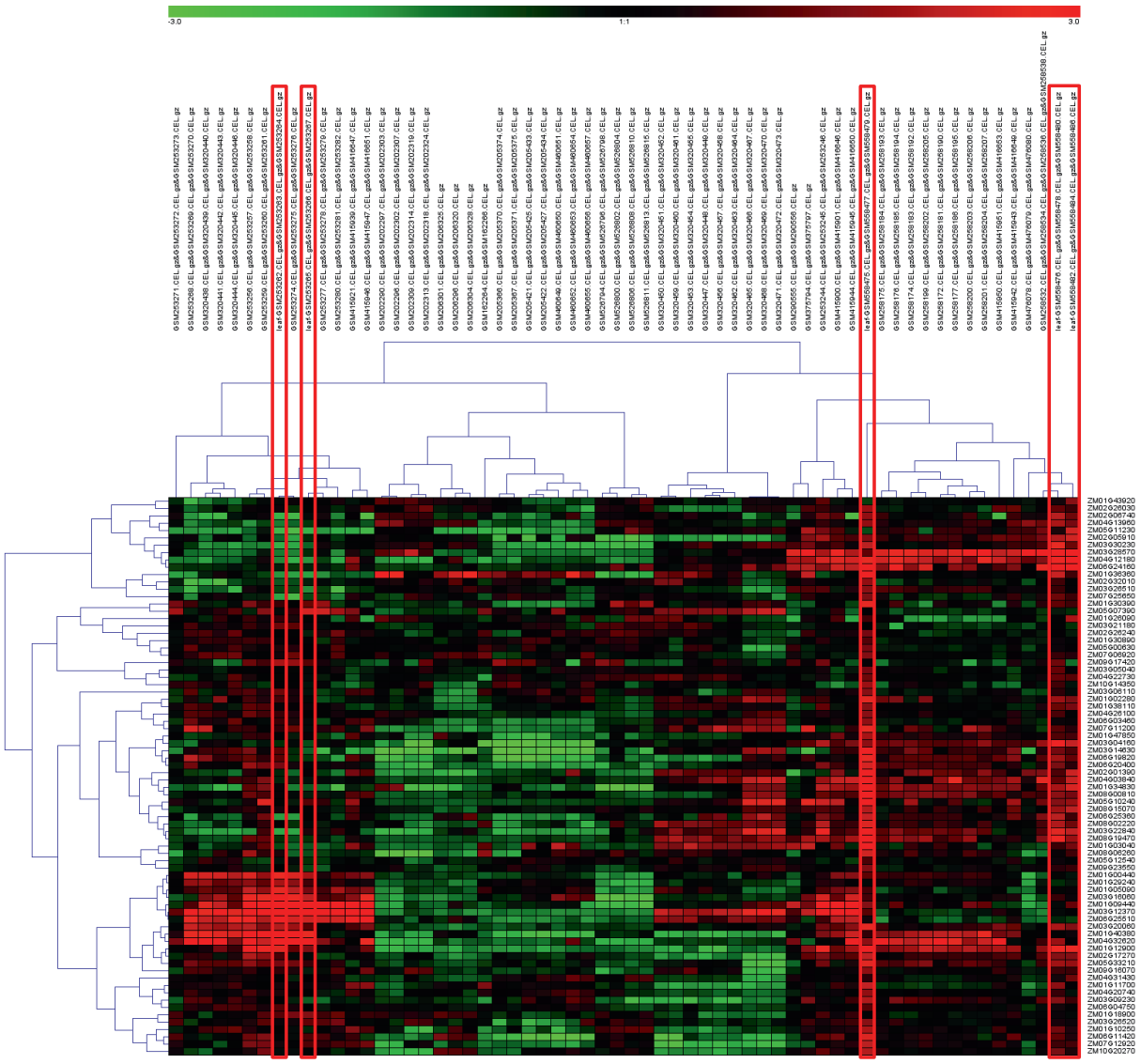


Figure 4.3: Maize leaf tissue specific genes expression profile - In this plot *Z.mays* (Maize) leaf tissue specific genes expression heatmap is shown and correlated leaf experiments are defined with red lines.

Species	experiments	genes
A. thaliana (Arabidopsis)	40	500
Z. mays (Maize)	5	76
M. truncatula (Medicago)	5	1178
P. trichocarpa (Poplar)	7	982
O. sativa (Rice)	11	6872
G. max (Soybean)	15	119
V. vinifera (Grape)	7	1382
total	90	11108

Table 4.5: Number of genes defined as **leaf** tissue specific genes and their correlated experiments after selection of pure leaf tissue specific genes (see section 4.3.4.1).

genes is shown in table D.1 (Appendix D) (see section 4.5.4 for the details of functional enrichment method). As it is listed in the table D.1 among the most significant GO enrichments some leaf related GO terms such as chloroplast (GO:0009507), plastid (GO:0009536), cytoplasm (GO:0005737), chloroplast part (GO:0044434), thylakoid (GO:0009579), thylakoid part (GO:0044436), chloroplast thylakoid (GO:0009534), plastid thylakoid (GO:0031976) and plastid stroma (GO:0009532) are observed repeatedly.

4.3.4.2 Clustering coexpressed leaf tissue specific genes

In this section, leaf tissue specific genes coexpression clusters are built to facilitate studying their relationship and improve leaf tissue specific genes network representation. In section 4.3.4.1 we came up with 11108 final genes. This large amount of genes will result in a very vague network. In order to solve this problem, coexpressed leaf tissue specific genes are clustered together per species and then the networks are drawn. It is important for this study to have high quality non-overlapping clusters, therefore Clustering Affinity Search Technique (CAST) (10) is used for this purpose (see section 4.5.6). Clusters only include highly coexpressed genes. Among the total number of clusters, 257 clusters are single clusters (i.e. including only one gene). In table 4.6 the number of clusters defined for each species is shown.

Species	Clusters num.	Max. cluster size	Single clusters	ECC cons. clusters
A. thaliana (Arabidopsis)	20	138	8	7 (35%)
Z. mays (Maize)	34	11	22	11 (32%)
M. truncatula (Medicago)	303	148	120	90 (30%)
P. trichocarpa (Poplar)	143	66	26	67 (47%)
O. sativa (Rice)	147	5631	36	48 (33%)
G. max (Soybean)	33	14	16	18 (55%)
V. vinifera (Grape)	127	213	29	57 (45%)

Table 4.6: In this table for each species you can see the number of leaf tissue specific gene clusters, size of the largest cluster, the total number of clusters with only one gene and the number of clusters which show ECC conservation with at least one other leaf tissue specific coexpression cluster from a different species.

4.3.4.3 Intraspecies leaf tissue specific genes network

In this section, for each species, the correlation of coexpressed leaf tissue specific genes are studied by using orthologous gene families. The intraspecies network for *Vitis vinifera* is shown in figure 4.4 as a representative. In this network nodes are gene clusters and the size of the node commensurates with the cluster size. Nodes are connected to each other only if both clusters have at least one orthologous family in common. Arabidopsis and Maize do not have any pair of clusters which share an orthologous family. Soybean has only one pair of clusters connected to each other with only one shared family (GM_2 is connected to single cluster GM_GM07G11910, because of shared family ORTHO004941 which includes genes GM09G30300 from first cluster and GM07G11910 from the second cluster). In the rest of the species there are also not many clusters connected, and if there is a connection, only one family is shared.

4.3.4.4 Interspecies leaf tissue specific genes network

In this section, we would like to investigate how leaf tissue specific genes from different species correlate with each other. Therefore, interspecies leaf tissue specific gene networks including all the seven plant species are made to see if more information can be obtained. In this network, nodes are leaf gene clusters as explained in section 4.3.4.2. The more genes in a cluster the larger the node size is. A gene cluster (node A) from one species is connected to a gene cluster (node B) from another species if they share at least one orthologous gene family. The edge line width is correlated with the Jaccrad

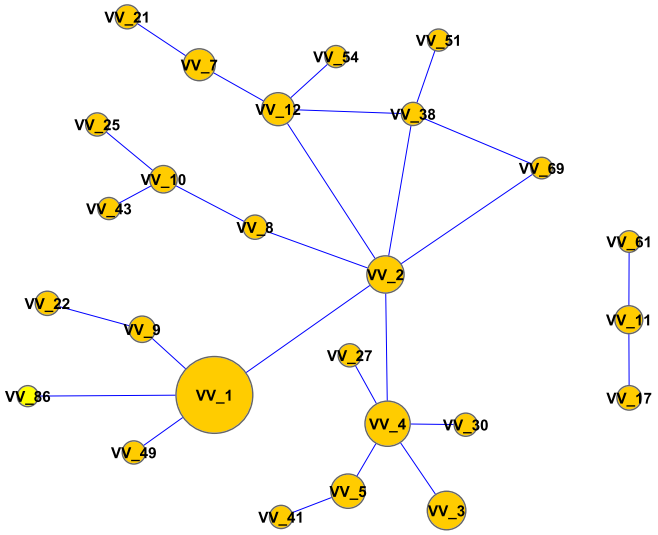


Figure 4.4: *Vitis vinifera* (Grape) intraspecies leaf tissue specific genes network (based on shared orthologous families) - Each node indicates one of the leaf tissue specific coexpression clusters of grape. Size of the nodes is correlated with the cluster size. An edge is drawn between two clusters only if they share at least one othologous family.

Index (JI) (see equation 4.3, figure 4.5 and figure 4.6).

$$JI = \frac{Shared\ orthologous\ families\ between\ two\ gene\ clusters}{Total\ number\ of\ unique\ orthologous\ families\ in\ both\ gene\ clusters} \quad (4.3)$$

4.3.4.5 Coexpression conservation of leaf tissue specific genes

To explore tissue specificity correlation with coexpression conservation, we calculated ECC (see section 4.5.5) for all coexpression cluster pairs with shared orthologous families between different species. Results show in total 247 conserved cluster pairs (P<0.01) which is illustrated in figure 4.7.

We further studied the most significant conserved clusters (P<0.001). For instance, rice coexpression cluster OS_2 with 413 genes and Medicago coexpression cluster MT_1 with 148 genes share 33 orthologous families (ECC conserved P<0.001). GO functional enrichment analysis of these coexpression clusters shows genes in rice coexpression cluster OS_2 are enriched in ligase activity forming aminoacyl-tRNA and related compounds, photosynthesis, ncRNA metabolic process, tRNA metabolic process, cytoplasm and plastid. Similarly, genes in Medicago coexpression cluster MT_1 are enriched in

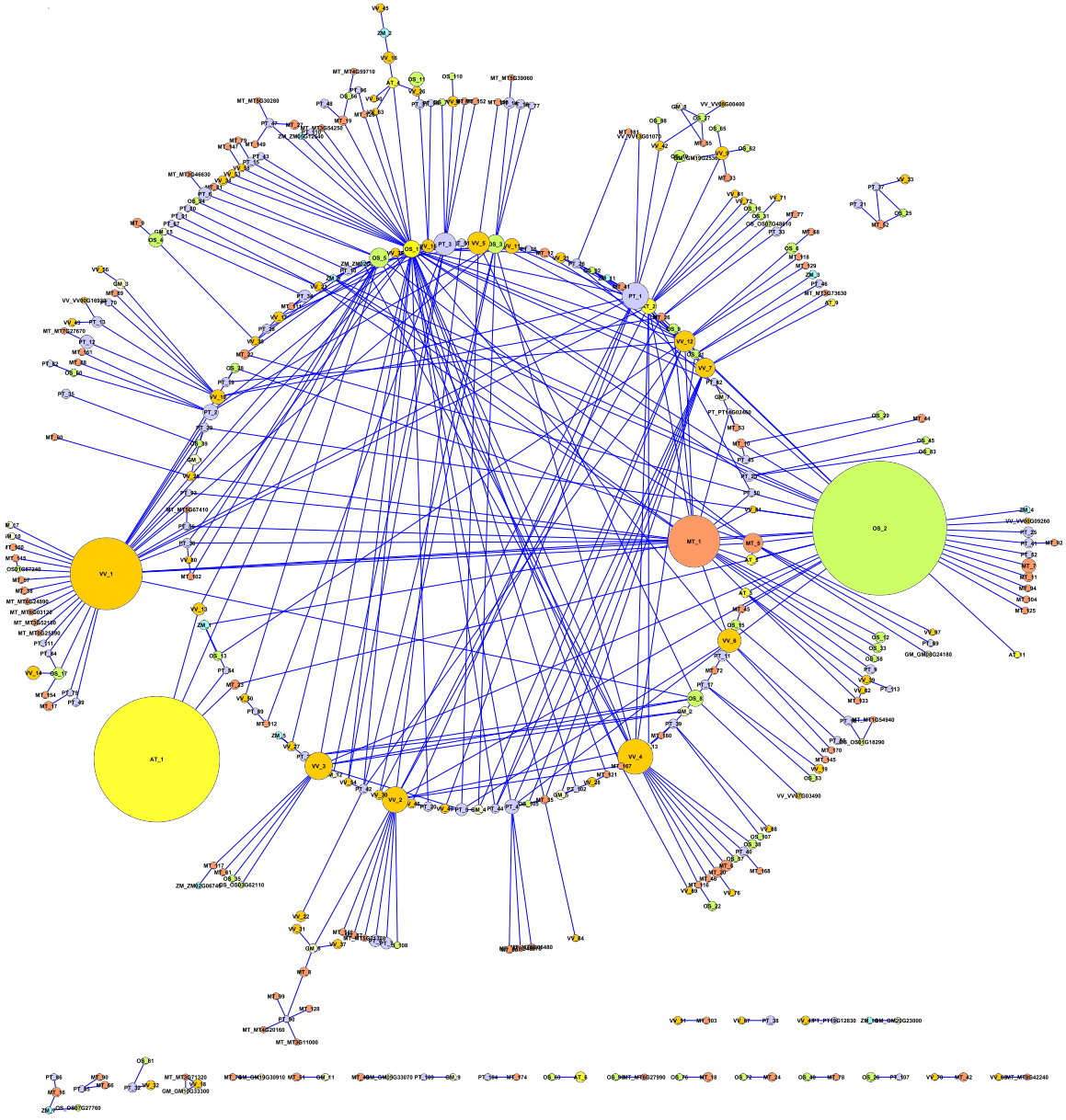


Figure 4.5: Interspecies leaf tissue specific genes network (based on shared orthologous families) - In this network each node is a coexpression cluster of leaf tissue specific genes. Node colors represent different species. Clusters with more genes have larger nodes. Nodes (gene clusters) are connected to each other with an edge only if they share at least one orthologous family. See Figure 4.6 for a zoomed out view.

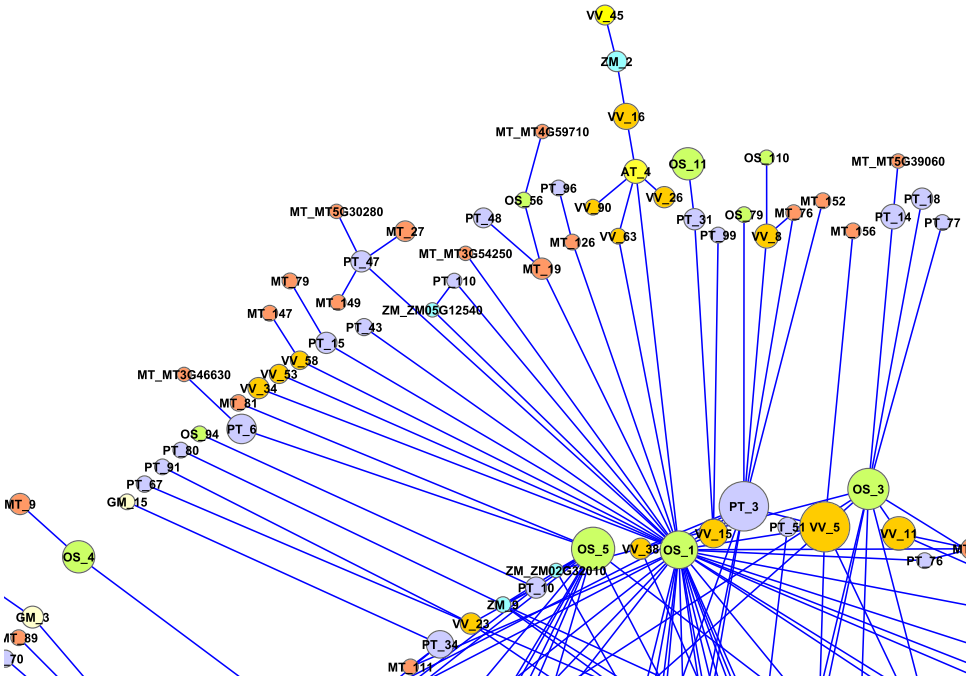


Figure 4.6: Interspecies leaf tissue specific genes network, a subgraph of figure 4.5 - In this network each node is a coexpression cluster of leaf tissue specific genes. Node colors represent different species. Clusters with more genes have larger nodes. Nodes (gene clusters) are connected to each other with an edge only if they share at least one orthologous family.

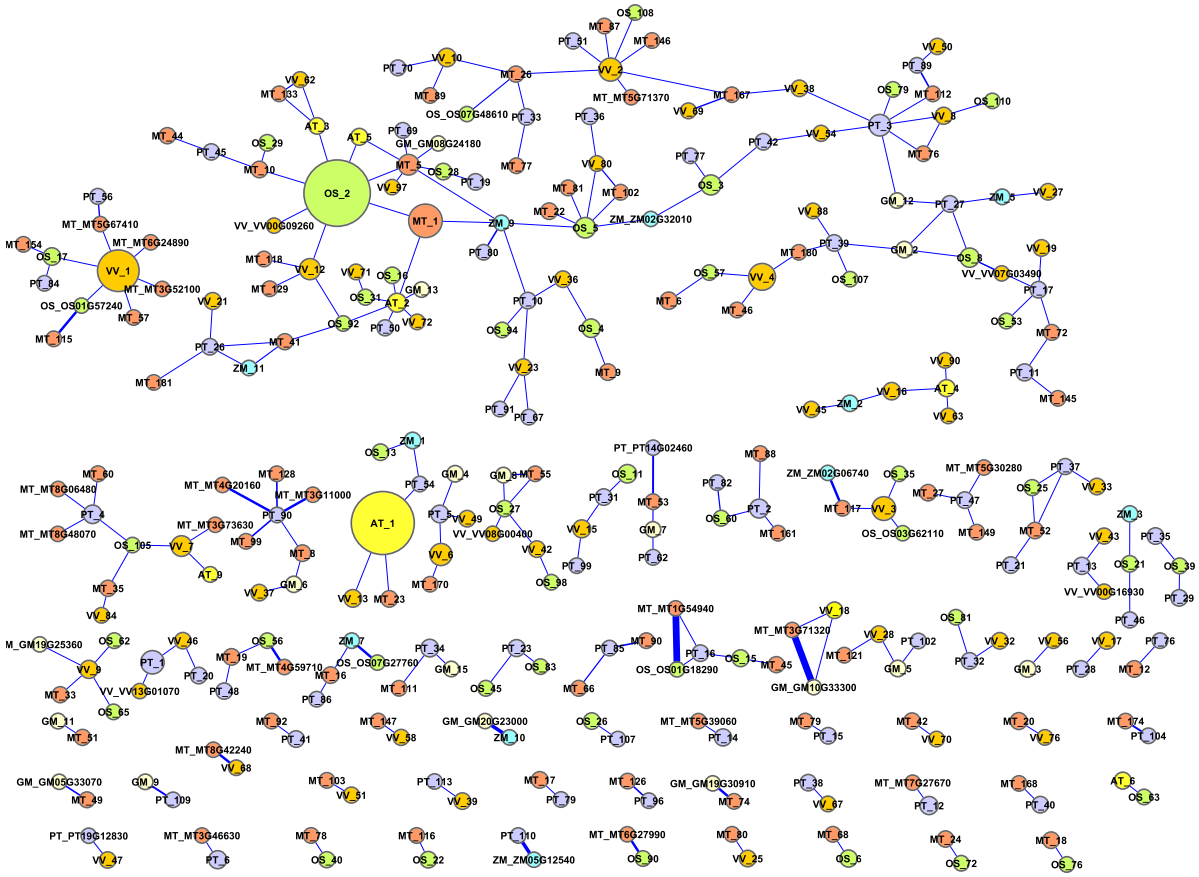


Figure 4.7: ECC conserved leaf tissue specific gene clusters - In this network each node is a coexpression cluster of leaf tissue specific genes. Each node color represents a different species. Clusters with more genes have larger nodes. Nodes (gene clusters) are connected to each other with an edge only if they are ECC conserved and consequently have at least one shared orthologous family. The thickness of the edge lines is proportional with the jaccard index (JI).

Cluster	GO type	GO term	Enrichment	P-value	Description
OS_2	MF	GO:0016876	3.98	6.23E-6	ligase activity, forming aminoacyl-tRNA and related compounds
OS_2	BP	GO:0015979	3.13	2.07E	photosynthesis
OS_2	BP	GO:0034660	3.99	4.15E-14	ncRNA metabolic process
OS_2	BP	GO:0006399	3.93	1.86E-10	tRNA metabolic process
OS_2	CC	GO:0005737	1.22	3.27E-65	cytoplasm
OS_2	CC	GO:0009536	2.54	2.50E-122	plastid
MT_1	MF	GO:0008236	4.36	7.41E-4	serine-type peptidase activity
MT_1	BP	GO:0015979	4.40	1.00E-4	photosynthesis
MT_1	BP	GO:0009536	4.42	1.50E-47	plastid
MT_1	BP	GO:0009507	4.51	9.13E-49	chloroplast
MT_1	CC	GO:0005737	3.09	7.95E-33	cytoplasm

Table 4.7: GO enrichment results for ECC conserved cluster pair OS_2 and MT_1. OS_2 is a leaf tissue specific coexpression cluster in rice and MT_1 is a leaf tissue specific coexpression cluster in Medicago. These two clusters show a significant coexpression conservation.

serine-type peptidase activity, photosynthesis, cytoplasm, plastid and chloroplast (see table 4.7).

As a second example, coexpression cluster pair Medicago MT_5 with 44 genes and Arabidopsis AT_5 with 14 genes have three shared orthologous families and are also ECC conserved (P<0.001). GO enrichment analysis for these coexpression clusters reveals similar functionality, as genes in Arabidopsis coexpression cluster AT_5 are enriched in chloroplast thylakoid membrane and plastid thylakoid membrane, likewise genes in Medicago coexpression cluster MT_5 are enriched in catalytic activity, chloroplast and plastid (see table 4.8).

4.4 Discussion

In this study gene expression profile evolution of seven different plant species is examined across various tissues. Expression data of 135793 genes in 1077 different conditions are investigated for seven plant species in order to find a list of reliable tissue specific genes. Our proposed method uses a combination of Tau value (40), z-score and a few other parameters to identify tissue specific genes. As a result 70427 tissue specific genes covering 40 different categories are defined (see section 4.3.1). Comparing our results with available tissue specific datasets, 79% of our defined Arabidopsis tissue specific

Cluster	GO type	GO term	Enrichment	p-value	Description
AT_5	CC	GO:0009535	4.21	0.02	chloroplast thylakoid membrane
AT_5	CC	GO:0055035	4.21	0.02	plastid thylakoid membrane
MT_5	MF	GO:0003824	1.70	1.36E-4	catalytic activity
MT_5	MF	GO:0008135	7.09	1.42E-4	translation factor activity,nucleic acid binding
MT_5	MF	GO:0003747	7.75	0.0038	translation release factor activity
MT_5	BP	GO:0034623	8.64	6.59E-6	cellular macromolecular complex disassembly
MT_5	BP	GO:0043624	8.64	6.59E-6	cellular protein complex disassembly
MT_5	BP	GO:0006415	8.64	6.59E-6	translational termination
MT_5	CC	GO:0009507	3.89	2.29E-7	chloroplast
MT_5	CC	GO:0009536	3.80	3.86E-7	plastid

Table 4.8: GO enrichment results for ECC conserved coexpression cluster pair (AT_52 and MT_5). AT_52 is a leaf tissue specific coexpression cluster in Arabidopsis and MT_5 is a leaf tissue specific coexpression cluster in Medicago. These two leaf tissue specific gene clusters show a significant coexpression conservation.

genes perfectly match with results of Schmid et al. (75).

Nie et al. compare gene expression of five common tissues between birds, mammals and amphibians and show that the expression patterns across tissues are highly similar for orthologous genes (153). While contrary to our expectations our study implies that genes in the same orthologous family do not always follow the same pattern of tissue specificity. Orthologous families can hold genes with similar tissue specificity from different species, but sometimes even the orthologous genes from the same species show different tissue specificity patterns (see section 4.3.3).

Our deep study on 11108 carefully selected leaf tissue specific genes revealed that coexpressed genes of the same species rarely share orthologous families. Which means orthologous leaf tissue specific genes of the same species are highly coexpressed. A Gene Ontology (GO) enrichment analysis showed that leaf tissue specific genes are enriched with GO terms reflecting the physiological functions of leaf tissue (see section 4.3.4.1 and 4.3.4.3).

It is shown that the evolutionary rate of expression profile is anti-correlated with tissue specificity (151). In this study also interesting cases of conservation in coexpressed leaf tissue specific genes have been observed among different species (see section 4.3.4.5). The GO enrichment analysis of ECC conserved leaf tissue specific genes reveals that these genes share many common GO terms and therefore are involved in similar bio-

Species	GEO-GPL id	Total samples	CEL samples
A. thaliana (Arabidopsis)	GPL198	6461	5769
Z. mays (Maize)	GPL4032	301	293
M. truncatula (Medicago)	GPL4652	94	94
P. trichocarpa (Poplar)	GPL4359	251	249
O. sativa (Rice)	GPL2025	737	636
G. max (Soybean)	GPL4592	3169	3157
V. vinifera (Grape)	GPL1320	210	178

Table 4.9: Number of Affymetrix microarray samples available on GEO database for each species. As you see for some samples there is no CEL file available yet.

logical functions. A similar study has been done by Piro et al. on human in order to find disease genes. They provide an atlas of tissue specific conserved coexpression for functional annotation and disease gene prediction (154).

The importance of tissue specific genes in different fields of biological sciences gives this study tremendous potential to improve. Fortunately, with the rapidly growing application of NGS technologies and the exceeding amount of generated expression profiles, a good opportunity has been provided to define and study tissue specific genes more effectively.

4.5 Materials and Methods

4.5.1 Expression data

We obtained Affymetrix CEL files for *A. thaliana*, *O. sativa*, *P. trichocarpa* (poplar), *G. max* (soybean), *V. vinifera* (grape), *Z. mays* (maize) and *M. truncatula* (Medicago) from the NCBI Gene Expression Omnibus (GEO) database (see table 4.9, January 2011), monitoring the transcriptional activity in different tissues and developmental stages and under different stress conditions.

The microarray slides were processed with a custom-made CDF file as described before in chapter 3 (5 probes per probeset and minimum 10 percent transcript coverage). Background adjustment, quantile normalization and summarization of the raw data was done with the RMA/justRMA algorithm implementation (58). To identify and remove

Species	Genes	Origin experiments	Reduced experiments
A. thaliana (Arabidopsis)	20,987	2409	491
Z. mays (Maize)	10,067	121	62
M. truncatula (Medicago)	17,613	63	30
P. trichocarpa (Poplar)	28,968	145	38
O. sativa (Rice)	34,152	329	80
G. max (Soybean)	15,752	1417	349
V. vinifera (Grape)	8,254	62	27

Table 4.10: Number of genes and experiments in the expression profiles. Last column indicates the number of remained experiments after removing redundant experiment (see section 4.5.1).

redundant experiments, first all experiments are clustered based on 95 percentile PCC threshold and only one experiment is retained as a representative for a set of similar experiments (38) (see table 4.10).

4.5.2 Clustering of expression data

In this study genes with similar expression pattern are grouped together using gene centric clustering algorithm (see chapter 2) and Pearson Correlation Coefficient (PCC) distance method. PCC values are calculated only for reduced expression profiles (see Expdata7). Based on the similarity between expression profiles for 1,000 random genes ($\approx 1,000 \times 999 \times 0.5$ gene pairs), a PCC threshold corresponding with the 90th percentile of this distribution, was set for each of the seven species (see table 4.11). Although the absolute PCC threshold differs for the seven species because of differences in size and composition of the expression data set, these percentile-based thresholds returns a similar relative cutoff for all. To create gene-centric coexpression clusters, for each seed gene, all coexpression partners with a PCC larger than the PCC threshold were retained. Application of gene-centric clustering for all the genes in the expression profile resulted in potentially overlapping clusters on a genome-wide scale.

4.5.3 Identification of orthologous gene families

To identify orthologous genes PLAZA 2.0 (89) is used. PLAZA clusters similar protein sequences from each species by means of the OrthoMCL algorithm starting from an

Species	80%	85%	**90%**	95%	99%
A. thaliana (Arabidopsis)	0.235	0.294	**0.370**	0.488	0.659
Z. mays (Maize)	0.374	0.460	**0.554**	0.663	0.807
M. truncatula (Medicago)	0.415	0.540	**0.678**	0.809	0.929
P. trichocarpa (Poplar)	0.277	0.340	**0.429**	0.500	0.707
O. sativa (Rice)	0.270	0.334	**0.443**	0.605	0.804
G. max (Soybean)	0.409	0.478	**0.551**	0.654	0.796
V. vinifera (Grape)	0.392	0.464	**0.546**	0.650	0.802

Table 4.11: PCC threshold corresponding with the 80th to 99th percentile of random distribution.

Species	total genes	source
A. thaliana (Arabidopsis)	20194	TAIR9
Z. mays (Maize)	9846	Maize B73 version 4.53
M. truncatula (Medicago)	17613	Mt3.0
P. trichocarpa (Poplar)	28968	JGI 2.0
O. sativa (Rice)	32456	TIGR6.1
G. max (Soybean)	15752	JGI 1.0
V. vinifera (Grape)	8249	Genoscope v1

Table 4.12: The distribution of genes in orthologous gene families.

all-against-all sequence similarity search with BLASTP (E-value < 1e-05) which only the top 1250 blast hits are retained. The resulted 114438 orthologous families are filtered for the genes which are present in the expression data. After filtration, 57910 orthologous families are left containing at least one gene from one of the seven species also present in the expression data. These families cover 133078 genes in total (see table 4.12 for details).

4.5.4 Functional enrichment

For the functional enrichment analysis of different gene groups, GO enrichment tool available in PLAZA workbench is used (89). PLAZA defines the significance of over- or under-representation using the hypergeometric distribution and applies the Bonferroni method to correct for multiple testing.

4.5.5 Calculation of ECC scores

To check for the conservation of two coexpression clusters the ECC method, as explained in section 3.5.5, is used. The algorithm can be fed with a pair of gene clusters from two different or similar species. The orthologous families information is also needed which is prepared earlier as explained in section 4.5.3. For the coexpression clusters gene centric clustering method can be used (see section 4.5.2), but for the special case of comparing tissue specific gene clusters we apply CAST clustering method (see section 4.3.4.2 and 4.5.6)(10). The number of permutations are 1000 times and the tau null model is applied because of the study content. A gene pair could be considered conserved, diverged, or not significant if it is respectively, significantly larger, smaller or not different compared to a background distribution of ECC values (p-value < 0.05).

4.5.6 Definition of CAST clustering method

CAST (Cluster Affinity Search Technique) (10) is a practical and fast algorithm for calculation gene coexpression clusters. In CAST method the average distance of an entry gene with all cluster members is calculated and if the value was less than a threshold the new gene is added to that cluster (see equation 4.4). In this study the correlation of genes is calculated using Pearson Correlation Coefficient (PCC) distance method (see section 4.5.2 for more information) and the average distance threshold is set to 90th percentile (see table 4.11).

Distance between gene i and cluster C:

$$d(i, C) = \textit{average distance between gene } i \textit{ and all genes in } C \tag{4.4}$$

Gene i is added to cluster C if $d(i, C) < \theta$, (θ: distance threshold)

4.5.7 Definition of Tau (τ)

Tau (τ) is a measure to quantify the level of differential expression across conditions (accounting for the quantitative variations in transcript levels) and to identify tissue specificity (40). τ ranges from 0 to 1, and high values correspond with low expression breadth (i.e. expressed in a few tissues). τ can be calculated from the following equation (4.5):

$$Tau\_value_{(i,n)} = \frac{\sum_{j=1}^{n} (1 - \left| \frac{ExpValue_{(i,j)}}{maxExpValue_{(i)}} \right|)}{n-1} \tag{4.5}$$

i: gene i ($1 \leq i \leq$ number of genes in expression profile)

n: number of experiments/tissues in expression profile

4.5.8 Definition of z-score

Z-scores also known as z-values, standardized variables, or normal scores tells how a single data point compares to normal data. A z-score says not only whether a point is above or below average, but also how unusual the measurement is. A z-score can be calculated from the following equation (4.6):

$$z = \frac{(x - \mu)}{\sigma} \tag{4.6}$$

x: a raw score to be standardized

σ: the standard deviation of the population

μ: the mean of the population

4.5.9 Visualization methods

Maize leaf tissue specific genes expression heatmap 4.3 is generated using Genesis software (155). Genesis normalize and visualize the expression profile with the Pearson Correlation Coefficient (PCC) distance measure and hierarchical clustering algorithm. For generating leaf tissue specific gene cluster networks (4.4,4.5,4.6,4.7), in the first step an in-house PERL script algorithm makes the necessary input interaction files as well as node and edge attribute files, and then these files are imported into Cytoscape software (156). The circular layout method is used for generating graph networks visualization.

4.6 Acknowledgments

I thank Michiel Van Bel for providing orthologous families dataset, Klaas Vandepoele for the fruitful discussions and critical reading of the text, and Yves Van de Peer for general support.

4.7 Author Contribution

This chapter is an article in preparation. I have performed all the analysis described and written the text, with helpful suggestions from Klaas Vandepoele.

5

Discussion

In the past few years we have witnessed rapid progresses in plant genome projects. These advances result in many complex datasets such as genome sequences of many species and microarray expression profiles of different tissues. These data challenges the field of Bioinformatics to find an answer for their underlying scientific problems. In this dissertation, we proposed two different methods to measure coexpression context conservation and identification of tissue specific genes.

Comparative (co-)expression analysis in plants

Gene co-expression networks have different applications, such as gene annotation, functional prediction, meta analysis, identification of biomarkers in cancer genetics, genomic data integration, and many more. Therefore they are one of the most promising applications for analyzing gene expression data. Subsets of highly connected genes may reveal relationships among biological processes and molecular functions. Recently, several studies have compared correlated gene expression patterns between species through the integration of gene homology and expression information. Many plant studies identified conserved expression modules with enriched gene functions in dicots and monocots (49, 50, 82). Other studies predict biological adaptations linking genotype with phenotype for the orthologous genes lacking expression conservation (5).
Our review based on recent comparative transcriptomics studies in plants, clarifies the route to systematically compare microarray expression data across species. The challenge starts with retrieval, normalization and annotation of microarray expression

information. Depending on the microarray technology this process can be different but regardless of the applied methodology all available microarray data for a specific organism is combined to build large-scale expression compendia which summarize expression profiles in tens or hundreds of different conditions (72). In order to compare genome-wide expression profiles between different species mostly a clustering algorithm is applied on the expression compendium and highly coexpressed genes are grouped together per species. Construction of coexpression networks facilitates studying groups of genes which share similar expression patterns. Afterwards many comparative expression studies compare gene coexpression clusters across species using homologous or orthologous genes. Defining sequence-based orthologs is a powerful approach to identify genes with conserved modules that participate in similar biological processes (6, 44, 45). To study the biological processes behind conserved coexpression modules, different functional annotation systems such as Gene Ontology (GO), information from KEGG pathways (98), Reactome (99) or MapMan (100) is used to identify enriched gene functions within conserved modules. With the application of next-generation sequencing to quantify plant transcriptomes (RNA-Seq) it is expected that new opportunities become available to study and compare expression profiles between species.

Comparative analysis of coexpression networks in Arabidopsis and Rice

The availability of massive amounts of microarray expression data under various conditions for Arabidopsis (*Arabidopsis thaliana*) and rice (*Oryza sativa*) encouraged us to study and compare the expression profiles of these two species. Starting from the systematically microarray expression data comparison across species, as explain above, we continued with a robust framework to measure expression context conservation (ECC). The analysis of 4,630 pairs of 1:1 orthologous Arabidopsis and rice genes demonstrate that 77% of the gene pairs show conserved coexpression patterns. Conservation of many developmental regulators provides valuable information for the transfer of biological knowledge between different plant species. In contrast, many genes that respond to a specific stress stimulus or are involved in signal transduction show low conservation levels, suggesting a link between regulatory evolution and adaptation to lifestyle or environment.

Furthermore the relationship between ECC, tissue specificity, and protein evolution

were analyzed. Tissue specific genes show higher conservation at the expression level comparing with broadly expressed genes, but evolve fast at the protein level. Other earlier studies have shown similar results comparing tissue specific genes and housekeeping genes (129, 149, 151). Protein and expression evolution does not show any significant correlation.

Comparative analysis of tissue specific gene coexpression networks in seven plant species

In the past few years large amount of microarray expression profiles have been generated for different plant species. We decided to take advantage of this opportunity and study tissue specific genes in different plant species. We study more than 130,000 genes in over 4500 conditions and present a method to define high quality tissue specific genes in seven different plant species; *A.thaliana* (Arabidopsis), *Z.mays* (Maize), *M.truncatula* (Medicago), *P.trichocarpa* (Poplar), *O.sativa* (Rice), *G.max* (Soybean) and *V.vinifera* (Grape). Results suggest 70,427 tissue specific genes in 40 different categories.
Afterward, a list of 11,108 leaf tissue specific genes were selected as a sample and the relationship between leaf tissue specific genes coexpression clusters within a species and in comparison with other species was studied in greater details. Results show orthologous leaf tissue specific genes of the same species are highly coexpressed. A few interesting cases of conservation in coexpressed leaf tissue specific genes among different species was also found. ECC conserved leaf tissue specific genes were enriched in similar GO terms when applying GO enrichment analysis.

Future aspects

DNA microarrays with the ability of simultaneously investigating thousands of transcripts, identifying differential expressed genes and detecting evolution of gene regulation, have been the technology of choice for large scale studies of gene expression level since mid 1990s. However this technology has some limitations. DNA microarrays do not translate all mRNAs and are not proper for detecting RNA splice patterns. Over the past few years NGS technologies have been used for different high-throughput

studies and could overcome many limitations of hybridization-based methods like microarrays. RNA-seq technology does not need bacterial cloning of the cDNA input. It can measure the expression of genes, transcripts and exons, and provide information on alternative splicing, gene fusions, post-transcriptional mutations and SNP variations (56, 157, 158, 159). Although currently analyzing the exceeding amount of new generated expression profiles is still a big challenge, there are lots of new bioinformatics tools and methods developed to perform comparative gene coexpression network analysis more accurately and efficiently (160).

With the outstanding level of sensitivity and high-throughput nature of NGS technology, RNA-seq is quickly replacing microarray based approaches for studying gene expression data with higher quantity, quality and accuracy. One of the most important advantages of RNA-seq over microarrays is the possibility of evaluating global gene expression patterns. In a RNA-seq experiment a large amount of mRNAs are converted to a set of randomly broken cDNA fragments with adaptors ligated to one end (single-end) or both ends (pair-end). Depending on the technology used (Roche 454, Illumina or SOLiD) final reads can have a length between 30 to 400 bp (161). To have an idea how gene expression analysis is done using RNA-seq data, here we briefly discuss general steps which are normally taken. First reads go through a quality control (QC) pipeline and low quality reads are removed or trimmed. Then reads are mapped to a reference genome or transcriptome using one of the different softwares available such as SOAP (162), TopHat (163), BWA (164), Bowtie (165) and GMAP/GSNAP (166, 167). In the third step, transcript/gene reads are normalized using one of the different statistical methods. For instance ERANGE (168) uses RPKM, edgeR uses TMM and Myrna uses upper quartile normalization method. Then in the last step statistical tests are performed to define differentially expressed (DE) genes. Some of the most popular tools which identify differentially expressed genes are DESeq (169), DEGseq (170), Cufflinks (171), Myrna (172) and edgeR (173) (174). Most of the times the study of DE genes is continued by applying known system biology approaches such as functional enrichment analysis, pathway analysis and gene network analysis in order to gain more information on the biological relevance.

References

[1] http://www.ebi.ac.uk/. 2

[2] ME Abalaka and D. Danboyi. **Mathematical and theoretical biology: An interdisciplinary scientific research field that links biology, medicine and biotechnology**. 2011. 2

[3] T.K. Attwood, D.J. Parry-Smith, and HM Wain. *Introduction to bioinformatics*. Longman, 1999. 2

[4] R. C. Hardison. **Comparative genomics**. *PLoS Biol*, **1**(2):E58, 2003. 3, 14, 38

[5] I. Tirosh, Y. Bilu, and N. Barkai. **Comparative biology: beyond sequence analysis**. *Curr Opin Biotechnol*, **18**(4):371–7, 2007. 3, 6, 14, 15, 33, 38, 39, 67, 107

[6] J. M. Stuart, E. Segal, D. Koller, and S. K. Kim. **A gene-coexpression network for global discovery of conserved genetic modules**. *Science*, **302**(5643):249–55, 2003. 3, 5, 6, 15, 25, 38, 39, 40, 108

[7] L. J. Jensen, T. S. Jensen, U. de Lichtenberg, S. Brunak, and P. Bork. **Co-evolution of transcriptional and post-translational cell-cycle regulation**. *Nature*, **443**(7111):594–7, 2006. 3, 38

[8] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell*. Garland Science, 4th edition, 2002. 3

[9] A.K. Jain and R.C. Dubes. *Algorithms for clustering data*. Prentice-Hall Inc., 1988. 3

[10] A. Ben-Dor, R. Shamir, and Z. Yakhini. **Clustering gene expression patterns**. *J Comput Biol*, **6**(3-4):281–97, 1999. 3, 25, 92, 103

[11] Y. Ogata, N. Sakurai, H. Suzuki, K. Aoki, K. Saito, and D. Shibata. **The prediction of local modular structures in a co-expression network based on gene expression datasets**. *Genome Inform*, **23**(1):117–27, 2009. 3, 25

[12] P. Langfelder and S. Horvath. **WGCNA: an R package for weighted correlation network analysis**. *BMC Bioinformatics*, **9**:559, 2008. 3, 25

[13] M. Mutwil, B. Usadel, M. Schutte, A. Loraine, O. Ebenhoh, and S. Persson. **Assembly of an interactive correlation network for the Arabidopsis genome using a novel heuristic clustering algorithm**. *Plant Physiol*, **152**(1):29–43, 2010. 3, 25, 67

[14] B. E. Dutilh, M. A. Huynen, and B. Snel. **A global definition of expression context is conserved between orthologs, but does not correlate with sequence conservation**. *BMC Genomics*, **7**:–, 2006. 4, 6, 40, 44

[15] S. Movahedi, M. Van Bel, K.S. Heyndrickx, and K. Vandepoele. **Comparative co-expression analysis in plant biology**. *Plant, Cell & Environment*, 2012. 4, 5

[16] W.M. Fitch. **Distinguishing homologous from analogous proteins**. *Systematic Biology*, **19**(2):99–113, 1970. 4

[17] T. Gabaldón et al. **Large-scale assignment of orthology: back to phylogenetics**. *Genome Biol*, **9**(10):235, 2008. 4

[18] E. V. Koonin. **Orthologs, paralogs, and evolutionary genomics**. *Annu Rev Genet*, **39**:309–38, 2005. 4, 25

[19] L. Li, Jr. Stoeckert, C. J., and D. S. Roos. **OrthoMCL: identification of ortholog groups for eukaryotic genomes**. *Genome Res*, **13**(9):2178–89, 2003. 4, 25

[20] G. Östlund, T. Schmitt, K. Forslund, T. Köstler, D.N. Messina, S. Roopra, O. Frings, and E.L.L. Sonnhammer. **InParanoid 7: new algorithms and tools for eukaryotic orthology analysis**. *Nucleic acids research*, **38**(suppl 1):D196–D203, 2010. 4, 25

[21] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, et al. **Gene Ontology: tool for the unification of biology**. *Nature genetics*, **25**(1):25, 2000. 4

[22] P. Jaiswal, S. Avraham, K. Ilic, E.A. Kellogg, S. McCouch, A. Pujar, L. Reiser, S.Y. Rhee, M.M. Sachs, M. Schaeffer, et al. **Plant Ontology (PO): a controlled vocabulary of plant structures and growth stages**. *Comparative and Functional Genomics*, **6**(7-8):388–397, 2005. 4, 33

[23] M. Bodén and T.L. Bailey. **Associating transcription factor-binding site motifs with target GO terms and target genes**. *Nucleic acids research*, **36**(12):4108–4117, 2008. 4

[24] F.A. Buske, M. Bodén, D.C. Bauer, and T.L. Bailey. **Assigning roles to DNA regulatory motifs using comparative genomics**. *Bioinformatics*, **26**(7):860–866, 2010. 4

[25] A. Conesa, S. Götz, J.M. García-Gómez, J. Terol, M. Talón, and M. Robles. **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research**. *Bioinformatics*, **21**(18):3674–3676, 2005. 4

[26] R.L. Poole. **The TAIR database**. *METHODS IN MOLECULAR BIOLOGY-CLIFTON THEN TOTOWA-*, **406**:179, 2007. 4

[27] R.L. Scholl, S.T. May, and D.H. Ware. **Seed and molecular resources for Arabidopsis**. *Plant physiology*, **124**(4):1477–1480, 2000. 4

[28] P. Jaiswal, J. Ni, I. Yap, D. Ware, W. Spooner, K. Youens-Clark, L. Ren, C. Liang, W. Zhao, K. Ratnapu, et al. **Gramene: a bird's eye view of cereal genomes**. *Nucleic acids research*, **34**(suppl 1):D717–D723, 2006. 4

REFERENCES

[29] K. Youens-Clark, E. Buckler, T. Casstevens, C. Chen, G. DeClerck, P. Derwent, P. Dharmawardhana, P. Jaiswal, P. Kersey, AS Karthikeyan, et al. **Gramene database in 2010: updates and extensions**. *Nucleic acids research*, **39**(suppl 1):D1085–D1094, 2011. 4

[30] I.Y. Tecle, N. Menda, R. Buels, and L. Mueller. **SGN Database: From QTLs to Genomes**. 2009. 4

[31] C.J. Lawrence, Q. Dong, M.L. Polacco, T.E. Seigfried, and V. Brendel. **MaizeGDB, the community database for maize genetics and genomics**. *Nucleic acids research*, **32**(suppl 1):D393–D397, 2004. 4

[32] JG Lees, JK Heriche, I. Morilla, JA Ranea, and CA Orengo. **Systematic computational prediction of protein interaction networks**. *Physical Biology*, **8**:035008, 2011. 4

[33] M. Min, H. Cai, W. Zheng, Z. Yang, X. Li, X. Fan, and Q. Feng. **PlaPID: a database of protein-protein interactions in plants**. In *Bioinformatics and Biomedical Engineering (iCBBE), 2010 4th International Conference on*, pages 1–4. IEEE, 2010. 4

[34] P. Pagel, S. Kovac, M. Oesterheld, B. Brauner, I. Dunger-Kaltenbach, G. Frishman, C. Montrone, P. Mark, V. Stümpflen, H.W. Mewes, et al. **The MIPS mammalian protein–protein interaction database**. *Bioinformatics*, **21**(6):832–834, 2005. 4

[35] S. Yellaboina, D.B. Dudekula, and M.S.H. Ko. **Prediction of evolutionarily conserved interologs in Mus musculus**. *BMC genomics*, **9**(1):465, 2008. 4

[36] I. Xenarios, L. Salwinski, X.J. Duan, P. Higney, S.M. Kim, and D. Eisenberg. **DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions**. *Nucleic acids research*, **30**(1):303–305, 2002. 5

[37] K. Eilbeck, A. Brass, N. Paton, and C. Hodgman. **INTERACT: an object oriented protein-protein interaction database**. In *Proc Int Conf Intell Syst Mol Biol*, **87**, page 94, 1999. 5

[38] S. De Bodt, D. Carvajal, J. Hollunder, J. Van den Cruyce, S. Movahedi, and D. Inze. **CORNET: a user-friendly tool for data mining and integration**. *Plant Physiol*, **152**(3):1167–79, 2010. 5, 18, 31, 33, 41, 67, 72, 101

[39] A. Rebane, K. Kisand, M. Laan, U. Haljasorg, R. Andreson, and P. Peterson. **AIRE activated tissue specific genes have histone modifications associated with inactive chromatin**. *Human molecular genetics*, **18**(24):4699–4710, 2009. 5

[40] B. Y. Liao and J. Z. Zhang. **Evolutionary conservation of expression profiles between human and mouse orthologous genes**. *Molecular Biology and Evolution*, **23**(3):530–540, 2006. 5, 52, 70, 98, 103

[41] B. Usadel, T. Obayashi, M. Mutwil, F. M. Giorgi, G. W. Bassel, M. Tanimoto, A. Chow, D. Steinhauser, S. Persson, and N. J. Provart. **Co-expression tools for plant biology: opportunities for hypothesis generation and caveats**. *Plant Cell Environ*, **32**(12):1633–51, 2009. 5, 18, 23, 32

[42] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. **Cluster analysis and display of genome-wide expression patterns**. *Proc Natl Acad Sci U S A*, **95**(25):14863–8, 1998. 5, 8, 38

[43] W. P. Lee and W. S. Tzou. **Computational methods for discovering gene networks from expression data**. *Brief Bioinform*, **10**(4):408–23, 2009. 5, 38

[44] S. Bergmann, J. Ihmels, and N. Barkai. **Similarities and differences in genome-wide expression data of six organisms**. *PLoS Biol*, **2**(1):E9, 2004. 5, 6, 15, 25, 39, 108

[45] Y. Lu, P. Huggins, and Z. Bar-Joseph. **Cross species analysis of microarray expression data**. *Bioinformatics*, **25**(12):1476–83, 2009. 5, 6, 14, 25, 39, 108

[46] A. Mustroph, S. C. Lee, T. Oosumi, M. E. Zanetti, H. Yang, K. Ma, A. Yaghoubi-Masihi, T. Fukao, and J. Bailey-Serres. **Cross-kingdom comparison of transcriptomic adjustments to low-oxygen stress highlights conserved and plant-specific responses**. *Plant Physiol*, **152**(3):1484–500, 2010. 5, 14, 39

[47] L. Huminiecki and K. H. Wolfe. **Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse**. *Genome Res*, **14**(10A):1870–9, 2004. 6, 39

[48] Y. Ogata, H. Suzuki, N. Sakurai, and D. Shibata. **CoP: a database for characterizing co-expressed gene modules with biological information in plants**. *Bioinformatics*, **26**(9):1267–8, 2010. 7, 21, 26

[49] S. Movahedi, Y. Van de Peer, and K. Vandepoele. **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**(3):1316–30, 2011. 7, 18, 21, 26, 27, 29, 33, 107

[50] M. Mutwil, S. Klie, T. Tohge, F. M. Giorgi, O. Wilkins, M. M. Campbell, A. R. Fernie, B. Usadel, Z. Nikoloski, and S. Persson. **PlaNet: Combined Sequence and Expression Comparisons across Plant Networks Derived from Seven Species**. *Plant Cell*, **23**(3):895–910, 2011. 7, 18, 21, 26, 29, 33, 34, 67, 107

[51] D. Jupiter, H. Chen, and V. VanBuren. **STARNET 2: a web-based tool for accelerating discovery of gene regulatory networks using microarray co-expression data**. *BMC Bioinformatics*, **10**:332, 2009. 7, 21, 26

[52] Iiris Hovatta et al. *DNA Microarray Data Analysis*. CSC Scientific Computing Ltd., 2005. 8

[53] Lucinda Kay Southworth. *Methods for integrating and comparing coexpression information over multiple data sets and applications in mice aging*. Doctoral Thesis / Dissertation, 2009. Doctoral Thesis / Dissertation. 8

[54] `http://itol.embl.de/`. 8

[55] `http://www.affymetrix.com/`. 8

[56] Z. Wang, M. Gerstein, and M. Snyder. **RNA-Seq: a revolutionary tool for transcriptomics**. *Nature reviews. Genetics*, **10**(1):57–63, 2009. 10, 110

[57] http://www.ncbi.nlm.nih.gov/geo/. 10

[58] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed. **Exploration, normalization, and summaries of high density oligonucleotide array probe level data**. *Biostatistics*, **4**(2):249–64, 2003. 10, 16, 71, 100

[59] P. Sykacek, D. P. Kreil, L. A. Meadows, R. P. Auburn, B. Fischer, S. Russell, and G. Micklem. **The impact of quantitative optimization of hybridization conditions on gene expression analysis**. *BMC Bioinformatics*, **12**:73, 2011. 11, 35

[60] T. Taji, M. Seki, M. Satou, T. Sakurai, M. Kobayashi, K. Ishiyama, Y. Narusaka, M. Narusaka, J. K. Zhu, and K. Shinozaki. **Comparative genomics in salt tolerance between Arabidopsis and aRabidopsis-related halophyte salt cress using Arabidopsis microarray**. *Plant Physiol*, **135**(3):1697–709, 2004. 14

[61] M. Weber, E. Harada, C. Vess, E. Roepenack-Lahaye, and S. Clemens. **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**(2):269–81, 2004. 14

[62] Q. Gong, P. Li, S. Ma, S. Indu Rupassara, and H. J. Bohnert. **Salinity stress adaptation competence in the extremophile Thellungiella halophila in comparison with its relative Arabidopsis thaliana**. *Plant J*, **44**(5):826–39, 2005. 14

[63] K. Vandenbroucke, S. Robbens, K. Vandepoele, D. Inze, Y. Van de Peer, and F. Van Breusegem. **Hydrogen peroxide-induced gene expression across kingdoms: a comparative analysis**. *Mol Biol Evol*, **25**(3):507–16, 2008. 14

[64] Y. Jiao, L. Ma, E. Strickland, and X. W. Deng. **Conservation and divergence of light-regulated genome expression patterns during seedling development in rice and Arabidopsis**. *Plant Cell*, **17**(12):3239–56, 2005. 15

[65] L. Ma, C. Chen, X. Liu, Y. Jiao, N. Su, L. Li, X. Wang, M. Cao, N. Sun, X. Zhang, J. Bao, J. Li, S. Pedersen, L. Bolund, H. Zhao, L. Yuan, G. K. Wong, J. Wang, and X. W. Deng. **A microarray analysis of the rice transcriptome and its comparison to Arabidopsis**. *Genome Res*, **15**(9):1274–83, 2005. 15, 40, 67, 69

[66] N. R. Street, A. Sjodin, M. Bylesjo, P. Gustafsson, J. Trygg, and S. Jansson. **A cross-species transcriptomics approach to identify genes involved in leaf development**. *BMC Genomics*, **9**:589, 2008. 15, 67

[67] G. Hardiman. **Microarray platforms–comparisons and contrasts**. *Pharmacogenomics*, **5**(5):487–502, 2004. 16

[68] P. Zimmermann, B. Schildknecht, D. Craigon, M. Garcia-Hernandez, W. Gruissem, S. May, G. Mukherjee, H. Parkinson, S. Rhee, U. Wagner, et al. **MIAME/Plant–adding value to plant microarray experiments**. *Plant Methods*, **2**(1):1, 2006. 16

[69] T. Barrett and R. Edgar. **Gene expression omnibus: microarray data storage, submission, retrieval, and analysis**. *Methods Enzymol*, **411**:352–69, 2006. 16

[70] H. Parkinson, U. Sarkans, N. Kolesnikov, N. Abeygunawardena, T. Burdett, M. Dylag, I. Emam, A. Farne, E. Hastings, E. Holloway, N. Kurbatova, M. Lukk, J. Malone, R. Mani, E. Pilicheva, G. Rustici, A. Sharma, E. Williams, T. Adamusiak, M. Brandizi, N. Sklyar, and A. Brazma. **ArrayExpress update–an archive of microarray and high-throughput sequencing-based functional genomics experiments**. *Nucleic Acids Res*, **39**(Database issue):D1002–4, 2011. 16

[71] R. P. Wise, R. A. Caldo, L. Hong, L. Shen, E. Cannon, and J. A. Dickerson. **BarleyBase/PLEXdb**. *Methods Mol Biol*, **406**:347–63, 2007. 16

[72] A. C. Fierro, F. Vandenbussche, K. Engelen, Y. Van de Peer, and K. Marchal. **Meta Analysis of Gene Expression Data within and Across Species**. *Curr Genomics*, **9**(8):525–34, 2008. 16, 108

[73] J. Quackenbush. **Microarray data normalization and transformation**. *Nat Genet*, **32 Suppl**:496–501, 2002. 16

[74] B. D. Gregory, J. Yazaki, and J. R. Ecker. **Utilizing tiling microarrays for whole-genome analysis in plants**. *Plant J*, **53**(4):636–44, 2008. 16

[75] M. Schmid, T. S. Davison, S. R. Henz, U. J. Pape, M. Demar, M. Vingron, B. Scholkopf, D. Weigel, and J. U. Lohmann. **A gene expression map of Arabidopsis thaliana development**. *Nat Genet*, **37**(5):501–6, 2005. 18, 59, 71, 87, 99

[76] A. Druka, G. Muehlbauer, I. Druka, R. Caldo, U. Baumann, N. Rostoks, A. Schreiber, R. Wise, T. Close, A. Kleinhofs, A. Graner, A. Schulman, P. Langridge, K. Sato, P. Hayes, J. McNicol, D. Marshall, and R. Waugh. **An atlas of gene expression from seed to seed through barley development**. *Funct Integr Genomics*, **6**(3):202–11, 2006. 18

[77] V. A. Benedito, I. Torres-Jerez, J. D. Murray, A. Andriankaja, S. Allen, K. Kakar, M. Wandrey, J. Verdier, H. Zuber, T. Ott, S. Moreau, A. Niebel, T. Frickey, G. Weiller, J. He, X. Dai, P. X. Zhao, Y. Tang, and M. K. Udvardi. **A gene expression atlas of the model legume Medicago truncatula**. *Plant J*, **55**(3):504–13, 2008. 18

[78] Y. Jiao, S. L. Tausta, N. Gandotra, N. Sun, T. Liu, N. K. Clay, T. Ceserani, M. Chen, L. Ma, M. Holford, H. Y. Zhang, H. Zhao, X. W. Deng, and T. Nelson. **A transcriptome atlas of rice cell types uncovers cellular, functional and developmental hierarchies**. *Nat Genet*, **41**(2):258–63, 2009. 18

[79] L. Wang, W. Xie, Y. Chen, W. Tang, J. Yang, R. Ye, L. Liu, Y. Lin, C. Xu, J. Xiao, and Q. Zhang. **A dynamic gene expression atlas covering the entire life cycle of rice**. *Plant J*, **61**(5):752–66, 2010. 18

[80] K. D. Edwards, A. Bombarely, G. W. Story, F. Allen, L. A. Mueller, S. A. Coates, and L. Jones. **TobEA: an atlas of tobacco gene expression from seed to senescence**. *BMC Genomics*, **11**:142, 2010. 18

REFERENCES

[81] M. Libault, A. Farmer, T. Joshi, K. Takahashi, R. J. Langley, L. D. Franklin, J. He, D. Xu, G. May, and G. Stacey. **An integrated transcriptome atlas of the crop model Glycine max, and its use in comparative analyses in plants**. *Plant J*, **63**(1):86–99, 2010. 18

[82] S. P. Ficklin and F. A. Feltus. **Gene Coexpression Network Alignment and Conservation of Gene Modules between Two Grass Species: Maize and Rice**. *Plant Physiol*, **156**(3):1244–56, 2011. 21, 27, 29, 33, 107

[83] R. Xu and 2nd Wunsch, D. **Survey of clustering algorithms**. *IEEE Trans Neural Netw*, **16**(3):645–78, 2005. 23

[84] Patrik Dhaeseleer. **How does gene expression clustering work?** *Computational Biology*, **23**(12):1499–1501, 2005. 23

[85] K. Vandepoele, M. Quimbaya, T. Casneuf, L. De Veylder, and Y. Van de Peer. **Unraveling transcriptional control in Arabidopsis using cis-regulatory elements and coexpression networks**. *Plant Physiol*, **150**(2):535–46, 2009. 25, 31, 51, 67, 71

[86] F. Luo, Y. Yang, J. Zhong, H. Gao, L. Khan, D. K. Thompson, and J. Zhou. **Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory**. *BMC Bioinformatics*, **8**:299, 2007. 25

[87] E. W. Sayers, T. Barrett, D. A. Benson, E. Bolton, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, S. Federhen, M. Feolo, I. M. Fingerman, L. Y. Geer, W. Helmberg, Y. Kapustin, D. Landsman, D. J. Lipman, Z. Lu, T. L. Madden, T. Madej, D. R. Maglott, A. Marchler-Bauer, V. Miller, I. Mizrachi, J. Ostell, A. Panchenko, L. Phan, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, M. Shumway, K. Sirotkin, D. Slotta, A. Souvorov, G. Starchenko, T. A. Tatusova, L. Wagner, Y. Wang, W. J. Wilbur, E. Yaschenko, and J. Ye. **Database resources of the National Center for Biotechnology Information**. *Nucleic Acids Res*, **39**(Database issue):D38–51, 2011. 25

[88] R. D. Finn, J. Mistry, J. Tate, P. Coggill, A. Heger, J. E. Pollington, O. L. Gavin, P. Gunasekaran, G. Ceric, K. Forslund, L. Holm, E. L. Sonnhammer, S. R. Eddy, and A. Bateman. **The Pfam protein families database**. *Nucleic Acids Res*, **38**(Database issue):D211–22, 2010. 25

[89] S. Proost, M. Van Bel, L. Sterck, K. Billiau, T. Van Parys, Y. Van de Peer, and K. Vandepoele. **PLAZA: A Comparative Genomics Resource to Study Gene and Genome Evolution in Plants**. *Plant Cell*, 2009. 26, 50, 73, 74, 75, 101, 102

[90] M. Martinez. **Plant protein-coding gene families: emerging bioinformatics approaches**. *Trends Plant Sci*, **in press**, 2011. 26

[91] A. Kuzniar, R. C. van Ham, S. Pongor, and J. A. Leunissen. **The quest for orthologs: finding the corresponding gene across genomes**. *Trends Genet*, **24**(11):539–51, 2008. 26

[92] M. Van Bel, S. Proost, E. Wischnitzki, S. Movahedi, C. Scheerlinck, Y. Van de Peer, and K. Vandepoele. **Dissecting plant genomes with the PLAZA comparative genomics platform**. *Plant Physiology*, **158**(2):590–600, 2012. 26

[93] E. Lyons, B. Pedersen, J. Kane, M. Alam, R. Ming, H. Tang, X. Wang, J. Bowers, A. Paterson, D. Lisch, et al. **Finding and comparing syntenic regions among Arabidopsis and the outgroups papaya, poplar, and grape: CoGe with rosids**. *Plant physiology*, **148**(4):1772–1781, 2008. 26

[94] M. Humphry, P. Bednarek, B. Kemmerling, S. Koh, M. Stein, U. Gobel, K. Stuber, M. Pislewska-Bednarek, A. Loraine, P. Schulze-Lefert, S. Somerville, and R. Panstruga. **A regulon conserved in monocot and dicot plants defines a functional module in antifungal plant immunity**. *Proc Natl Acad Sci U S A*, **107**(50):21896–901, 2010. 26, 68

[95] M. D. Chikina and O. G. Troyanskaya. **Accurate Quantification of Functional Analogy among Close Homologs**. *PLoS Comput Biol*, **7**(2):e1001074, 2011. 26, 27, 32, 67

[96] P. Zarrineh, A. C. Fierro, A. Sanchez-Rodriguez, B. De Moor, K. Engelen, and K. Marchal. **COMODO: an adaptive coclustering strategy to identify conserved coexpression modules between organisms**. *Nucleic Acids Res*, **39**(7):e41, 2011. 26, 27

[97] K. Vandepoele and Y. Van de Peer. **Exploring the plant transcriptome through phylogenetic profiling**. *Plant Physiol*, **137**(1):31–42, 2005. 27, 40

[98] M. Kanehisa, S. Goto, M. Furumichi, M. Tanabe, and M. Hirakawa. **KEGG for representation and analysis of molecular networks involving diseases and drugs**. *Nucleic Acids Res*, **38**(Database issue):D355–60, 2010. 29, 108

[99] N. Tsesmetzis, M. Couchman, J. Higgins, A. Smith, J. H. Doonan, G. J. Seifert, E. E. Schmidt, I. Vastrik, E. Birney, G. Wu, P. D'Eustachio, L. D. Stein, R. J. Morris, M. W. Bevan, and S. V. Walsh. **Arabidopsis reactome: a foundation knowledgebase for plant systems biology**. *Plant Cell*, **20**(6):1426–36, 2008. 29, 73, 108

[100] B. Usadel, F. Poree, A. Nagel, M. Lohse, A. Czedik-Eysenberg, and M. Stitt. **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**(9):1211–29, 2009. 29, 108

[101] W. Huang da, B. T. Sherman, and R. A. Lempicki. **Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists**. *Nucleic Acids Res*, **37**(1):1–13, 2009. 29

[102] M.E. Smoot, K. Ono, J. Ruscheinski, P.L. Wang, and T. Ideker. **Cytoscape 2.8: new features for data integration and network visualization**. *Bioinformatics*, **27**(3):431, 2011. 29

[103] K. Vandepoele, K. Vlieghe, K. Florquin, L. Hennig, G.T.S. Beemster, W. Gruissem, Y. Van de Peer, D. Inzé, and L. De Veylder. **Genome-wide identification of potential plant E2F target genes**. *Plant physiology*, **139**(1):316–328, 2005. 31

[104] N. Takahashi, M. Quimbaya, V. Schubert, T. Lammens, K. Vandepoele, I. Schubert, M. Matsui, D. Inzé, G. Berx, and L. De Veylder. **The MCM-binding protein ETG1 aids sister chromatid cohesion required for postreplicative homologous recombination repair**. *PLoS genetics*, **6**(1):e1000817, 2010. 31

[105] Y. Van de Peer, J. A. Fawcett, S. Proost, L. Sterck, and K. Vandepoele. **The flowering world: a tale of duplications**. *Trends Plant Sci*, **14**(12):680–8, 2009. 32

[106] S. Tarazona, F. García-Alcalde, J. Dopazo, A. Ferrer, and A. Conesa. **Differential expression in RNA-seq: A matter of depth**. *Genome Research*, 2011. 34

[107] S. Ouyang, W. Zhu, J. Hamilton, H. Lin, M. Campbell, K. Childs, F. Thibaud-Nissen, R. L. Malek, Y. Lee, L. Zheng, J. Orvis, B. Haas, J. Wortman, and C. R. Buell. **The TIGR Rice Genome Annotation Resource: improvements and new features**. *Nucleic Acids Res*, **35**(Database issue):D883–7, 2007. 35

[108] T. Casneuf, Y. Van de Peer, and W. Huber. **In situ analysis of cross-hybridisation on microarrays and the inference of expression correlation**. *BMC Bioinformatics*, **8**:461, 2007. 35, 71

[109] M. H. Dai, P. L. Wang, A. D. Boyd, G. Kostov, B. Athey, E. G. Jones, W. E. Bunney, R. M. Myers, T. P. Speed, H. Akil, S. J. Watson, and F. Meng. **Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data**. *Nucleic Acids Research*, **33**(20), 2005. 35

[110] R. A. Studer and M. Robinson-Rechavi. **How confident can we be that orthologs are similar, but paralogs differ?** *Trends Genet*, **25**(5):210–6, 2009. 39

[111] L. Sterck, S. Rombauts, K. Vandepoele, P. Rouze, and Y. Van de Peer. **How many genes are there in plants (... and why are they there)?** *Curr Opin Plant Biol*, **10**(2):199–203, 2007. 40

[112] Y. Pilpel, P. Sudarsanam, and G. M. Church. **Identifying regulatory networks by combinatorial analysis of promoter elements**. *Nat Genet*, **29**(2):153–9, 2001. 41

[113] S. Orzechowski. **Starch metabolism in leaves**. *Acta Biochim Pol*, **55**(3):435–45, 2008. 49

[114] F. F. Fu and H. W. Xue. **Co-expression analysis identifies Rice Starch Regulator1 (RSR1), a rice AP2/EREBP family transcription factor, as a novel rice starch biosynthesis regulator**. *Plant Physiol*, 2010. 50

[115] H. Liu, R. Sachidanandam, and L. Stein. **Comparative genomics between rice and Arabidopsis shows scant collinearity in gene order**. *Genome Res*, **11**(12):2020–2026, 2001. 50

[116] K. Vandepoele, Y. Saeys, C. Simillion, J. Raes, and Y. Van De Peer. **The automatic detection of homologous regions (ADHoRe) and its application to microcolinearity between Arabidopsis and rice**. *Genome Res*, **12**(11):1792–801, 2002. 50

[117] X. Y. Ren, M. W. Fiers, W. J. Stiekema, and J. P. Nap. **Local coexpression domains of two to four genes in the genome of Arabidopsis**. *Plant Physiol*, **138**(2):923–34, 2005. 50

[118] S. K. Palaniswamy, S. James, H. Sun, R. S. Lamb, R. V. Davuluri, and E. Grotewold. **AGRIS and AtRegNet. a platform to link cis-regulatory elements and transcription factors into regulatory networks**. *Plant Physiol*, **140**(3):818–29, 2006. 51

[119] K. Higo, Y. Ugawa, M. Iwamoto, and T. Korenaga. **Plant cis-acting regulatory DNA elements (PLACE) database: 1999**. *Nucleic Acids Res*, **27**(1):297–300, 1999. 51

[120] E. T. Chan, G. T. Quon, G. Chua, T. Babak, M. Trochesset, R. A. Zirngibl, J. Aubin, M. J. Ratcliffe, A. Wilde, M. Brudno, Q. D. Morris, and T. R. Hughes. **Conservation of core gene expression in vertebrate tissues**. *J Biol*, **8**(3):33, 2009. 59

[121] K. Aoki, Y. Ogata, and D. Shibata. **Approaches for extracting practical information from gene co-expression networks in plant biology**. *Plant Cell Physiol*, **48**(3):381–90, 2007. 67

[122] T. H. Lee, Y. K. Kim, T. T. Pham, S. I. Song, J. K. Kim, K. Y. Kang, G. An, K. H. Jung, D. W. Galbraith, M. Kim, U. H. Yoon, and B. H. Nahm. **RiceArrayNet: a database for correlating gene expression from transcriptome profiling, and its application to the analysis of coexpressed genes in rice**. *Plant Physiol*, **151**(1):16–33, 2009. 67

[123] T. Obayashi and K. Kinoshita. **Coexpression landscape in ATTED-II: usage of gene list and gene network for various types of pathways**. *J Plant Res*, **123**(3):311–9, 2010. 67

[124] M. Ha, W. H. Li, and Z. J. Chen. **External factors accelerate expression divergence between duplicate genes**. *Trends Genet*, **23**(4):162–6, 2007. 68

[125] D. Meinke, R. Muralla, C. Sweeney, and A. Dickerman. **Identifying essential genes in Arabidopsis thaliana**. *Trends Plant Sci*, **13**(9):483–91, 2008. 68

[126] L. T. Macneil and A. J. Walhout. **Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression**. *Genome Res*, **in press**, 2011. 68

[127] X. Y. Ren, W. J. Stiekema, and J. P. Nap. **Local co-expression domains in the genome of rice show no microsynteny with Arabidopsis domains**. *Plant Mol Biol*, **65**(1-2):205–17, 2007. 69

[128] K. E. Hastings. **Strong evolutionary conservation of broadly expressed protein isoforms in the troponin I gene family and other vertebrate gene families**. *J Mol Evol*, **42**(6):631–40, 1996. 70, 80

[129] L. Duret and D. Mouchiroud. **Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate**. *Mol Biol Evol*, **17**(1):68–74, 2000. 70, 80, 109

REFERENCES

[130] E. P. Rocha and A. Danchin. **An analysis of determinants of amino acids substitution rates in bacterial proteins**. *Mol Biol Evol*, **21**(1):108–16, 2004. 70

[131] I. Tirosh and N. Barkai. **Evolution of gene sequence and gene expression are not correlated in yeast**. *Trends in genetics : TIG*, **24**(3):109–13, 2008. 70

[132] K. D. Makova and W. H. Li. **Divergence in the spatial pattern of gene expression between human duplicate genes**. *Genome Res*, **13**(7):1638–45, 2003. 70

[133] I. K. Jordan, L. Marino-Ramirez, and E. V. Koonin. **Evolutionary significance of gene expression divergence**. *Gene*, **345**(1):119–26, 2005. 70

[134] B. Lemos, B. R. Bettencourt, C. D. Meiklejohn, and D. L. Hartl. **Evolution of proteins and gene expression levels are coupled in Drosophila and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions**. *Mol Biol Evol*, **22**(5):1345–54, 2005. 70

[135] M. A. Sartor, A. M. Zorn, J. A. Schwanekamp, D. Halbleib, S. Karyala, M. L. Howell, G. E. Dean, M. Medvedovic, and C. R. Tomlinson. **A new method to remove hybridization bias for interspecies comparison of global gene expression profiles uncovers an association between mRNA sequence divergence and differential gene expression in Xenopus**. *Nucleic Acids Res*, **34**(1):185–200, 2006. 70

[136] G. Blanc and K. H. Wolfe. **Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution**. *Plant Cell*, **16**(7):1679–91, 2004. 70

[137] A. Tanay, A. Regev, and R. Shamir. **Conservation and evolvability in regulatory networks: the evolution of ribosomal regulation in yeast**. *Proc Natl Acad Sci U S A*, **102**(20):7203–8, 2005. 70

[138] M. Jain, A. Nijhawan, R. Arora, P. Agarwal, S. Ray, P. Sharma, S. Kapoor, A. K. Tyagi, and J. P. Khurana. **F-box proteins in rice. Genome-wide analysis, classification, temporal and spatial gene expression during panicle and seed development, and regulation by light and abiotic stress**. *Plant Physiol*, **143**(4):1467–83, 2007. 71

[139] M. Li, W. Xu, W. Yang, Z. Kong, and Y. Xue. **Genome-wide gene expression profiling reveals conserved and novel molecular functions of the stigma in rice**. *Plant Physiol*, **144**(4):1797–812, 2007. 71

[140] K. Birnbaum, D. E. Shasha, J. Y. Wang, J. W. Jung, G. M. Lambert, D. W. Galbraith, and P. N. Benfey. **A gene expression map of the Arabidopsis root**. *Science*, **302**(5652):1956–60, 2003. 71

[141] G. J. Norton, M. J. Aitkenhead, F. S. Khowaja, W. R. Whalley, and A. H. Price. **A bioinformatic and transcriptomic approach to identifying positional candidate genes without fine mapping: an example using rice root-growth QTLs**. *Genomics*, **92**(5):344–52, 2008. 71

[142] H. Walia, C. Wilson, P. Condamine, X. Liu, A. M. Ismail, L. Zeng, S. I. Wanamaker, J. Mandal, J. Xu, X. Cui, and T. J. Close. **Comparative transcriptional profiling of two contrasting rice genotypes under salinity stress during the vegetative growth stage**. *Plant Physiol*, **139**(2):822–35, 2005. 71

[143] J. Kilian, D. Whitehead, J. Horak, D. Wanke, S. Weinl, O. Batistic, C. D'Angelo, E. Bornberg-Bauer, J. Kudla, and K. Harter. **The AtGenExpress global stress expression data set: protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses**. *Plant J*, **50**(2):347–63, 2007. 71

[144] P. J. Swarbrick, K. Huang, G. Liu, J. Slate, M. C. Press, and J. D. Scholes. **Global patterns of gene expression in rice cultivars undergoing a susceptible or resistant interaction with the parasitic plant Striga hermonthica**. *New Phytol*, **179**(2):515–29, 2008. 71

[145] F. Chen, A. J. Mackey, J. K. Vermunt, and D. S. Roos. **Assessing performance of orthology detection strategies applied to eukaryotic genomes**. *PLoS One*, **2**(4):e383, 2007. 73

[146] O. Thimm, O. Blasing, Y. Gibon, A. Nagel, S. Meyer, P. Kruger, J. Selbig, L. A. Muller, S. Y. Rhee, and M. Stitt. **MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes**. *Plant J*, **37**(6):914–39, 2004. 74

[147] J. D. Thompson, T. J. Gibson, and D. G. Higgins. **Multiple sequence alignment using ClustalW and ClustalX**. *Curr Protoc Bioinformatics*, **Chapter 2**:Unit 2 3, 2002. 75

[148] Z. Yang. **PAML: a program package for phylogenetic analysis by maximum likelihood**. *Comput Appl Biosci*, **13**(5):555–6, 1997. 75

[149] L. Zhang and W. H. Li. **Mammalian housekeeping genes evolve more slowly than tissue-specific genes**. *Molecular Biology and Evolution*, **21**(2):236–9, 2004. 80, 109

[150] C. Pal, B. Papp, and L. D. Hurst. **Highly expressed genes in yeast evolve slowly**. *Genetics*, **158**(2):927–31, 2001. 80

[151] B. Y. Liao and J. Zhang. **Low rates of expression profile divergence in highly expressed genes and tissue-specific genes during mammalian evolution**. *Molecular Biology and Evolution*, **23**(6):1119–28, 2006. 80, 99, 109

[152] D. T. Odom, R. D. Dowell, E. S. Jacobsen, W. Gordon, T. W. Danford, K. D. MacIsaac, P. A. Rolfe, C. M. Conboy, D. K. Gifford, and E. Fraenkel. **Tissue-specific transcriptional regulation has diverged significantly between human and mouse**. *Nature genetics*, **39**(6):730–2, 2007. 81

[153] H. Nie, R. P. Crooijmans, A. Lammers, E. M. van Schothorst, J. Keijer, P. B. Neerincx, J. A. Leunissen, H. J. Megens, and M. A. Groenen. **Gene expression in chicken reveals correlation with structural genomic features and conserved patterns of transcription in the terrestrial vertebrates**. *PLoS One*, **5**(8):e11990, 2010. 99

[154] R. M. Piro, U. Ala, I. Molineris, E. Grassi, C. Bracco, G. P. Perego, P. Provero, and F. Di Cunto. **An atlas of tissue-specific conserved coexpression for functional annotation and disease gene prediction**. *Eur J Hum Genet*, **19**(11):1173–80, 2011. 100

[155] A. Sturn, J. Quackenbush, and Z. Trajanoski. **Genesis: cluster analysis of microarray data**. *Bioinformatics*, **18**(1):207–8, 2002. 104

[156] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. **Cytoscape: a software environment for integrated models of biomolecular interaction networks**. *Genome Res*, **13**(11):2498–504, 2003. 104

[157] J.C. Marioni, C.E. Mason, S.M. Mane, M. Stephens, and Y. Gilad. **RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays**. *Genome research*, **18**(9):1509–1517, 2008. 110

[158] R.M. Davidson, C.N. Hansey, M. Gowda, K.L. Childs, H. Lin, B. Vaillancourt, R.S. Sekhon, N. de Leon, S.M. Kaeppler, N. Jiang, and C.R. Buell. **Utility of RNA sequencing for analysis of maize reproductive transcriptomes**. *The Plant Genome*, **4**(3):191–203, 2011. 110

[159] V. Costa, C. Angelini, I. De Feis, and A. Ciccodicola. **Uncovering the complexity of transcriptomes with RNA-Seq**. *J Biomed Biotechnol*, **853916**, 2010. 110

[160] R.M. Davidson, M. Gowda, G. Moghe, H. Lin, B. Vaillancourt, S.H. Shiu, N. Jiang, and C. Robin Buell. **Comparative transcriptomics of three Poaceae species reveals patterns of gene expression evolution**. *The Plant Journal*, 2012. 110

[161] M. Jain. **Next-generation sequencing technologies for gene expression profiling in plants**. *Briefings in Functional Genomics*, **11**(1):63–70, 2012. 110

[162] R. Li, Y. Li, K. Kristiansen, and J. Wang. **SOAP: short oligonucleotide alignment program**. *Bioinformatics*, **24**(5):713–714, 2008. 110

[163] C. Trapnell, L. Pachter, and S.L. Salzberg. **TopHat: discovering splice junctions with RNA-Seq**. *Bioinformatics*, **25**(9):1105–1111, 2009. 110

[164] H. Li and R. Durbin. **Fast and accurate short read alignment with Burrows–Wheeler transform**. *Bioinformatics*, **25**(14):1754–1760, 2009. 110

[165] B. Langmead, C. Trapnell, M. Pop, and S.L. Salzberg. **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome**. *Genome Biol*, **10**(3):R25, 2009. 110

[166] T.D. Wu and C.K. Watanabe. **GMAP: a genomic mapping and alignment program for mRNA and EST sequences**. *Bioinformatics*, **21**(9):1859, 2005. 110

[167] T.D. Wu and S. Nacu. **Fast and SNP-tolerant detection of complex variants and splicing in short reads**. *Bioinformatics*, **26**(7):873, 2010. 110

[168] A. Mortazavi, B.A. Williams, K. McCue, L. Schaeffer, and B. Wold. **Mapping and quantifying mammalian transcriptomes by RNA-Seq**. *Nature methods*, **5**(7):621–628, 2008. 110

[169] S. Anders and W. Huber. **Differential expression analysis for sequence count data**. *Genome Biol*, **11**(10):R106, 2010. 110

[170] L. Wang, Z. Feng, X. Wang, X. Wang, and X. Zhang. **DEGseq: an R package for identifying differentially expressed genes from RNA-seq data**. *Bioinformatics*, **26**(1):136–138, 2010. 110

[171] A. Roberts, H. Pimentel, C. Trapnell, and L. Pachter. **Identification of novel transcripts in annotated genomes using RNA-Seq**. *Bioinformatics*, **27**(17):2325–2329, 2011. 110

[172] B. Langmead, K.D. Hansen, and J.T. Leek. **Cloud-scale RNA-sequencing differential expression analysis with Myrna**. *Genome Biology*, **11**(8):R83, 2010. 110

[173] M.D. Robinson, D.J. McCarthy, and G.K. Smyth. **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data**. *Bioinformatics*, **26**(1):139–140, 2010. 110

[174] A. Oshlack, M.D. Robinson, and M.D. Young. **From RNA-seq reads to differential expression results**. *Genome Biology*, **11**(12):220, 2010. 110

[175] `http://www.affymetrix.com/support/technical/byproduct.affx`. 119

[176] `http://www.ebi.ac.uk/~guy/exonerate/`. 119

Appendix A

Step by Step Creating a Custom Made CDF file

The primary data needed to create an Affymetrix custom made CDF file:

- A Probe file in FASTA format[1]
- A cDNA file (including UTRs) in FASTA format

Initially a probe:cDNA alignment is done using Exonerate tool[2] (176) (allowing 0 or 1 mismatch and no gap).

exonerate –showalignment no –showvulgar yes -s 116 probe_file cDNA_file > exonerate-align.txt

In the next step, unique probes are selected in gene/transcript level. Then in case of alternative splicing:

- At gene level probes are filtered only if they are shared between at least 80% of the transcripts.
- At transcript level, check if there are any probes available which uniquely map to each transcript. (see figure 2.6 in chapter 2)

Finally probes which uniquely map to a specific gene/transcript are grouped in to a probeset. Each probeset should at least include 8 probes (see the complete flowchart in figure A.1) .

[1]Probe files can be found in Affymetix website (175).
[2]Exonerate program is a generic tool for sequence alignment.

A. STEP BY STEP CREATING A CUSTOM MADE CDF FILE

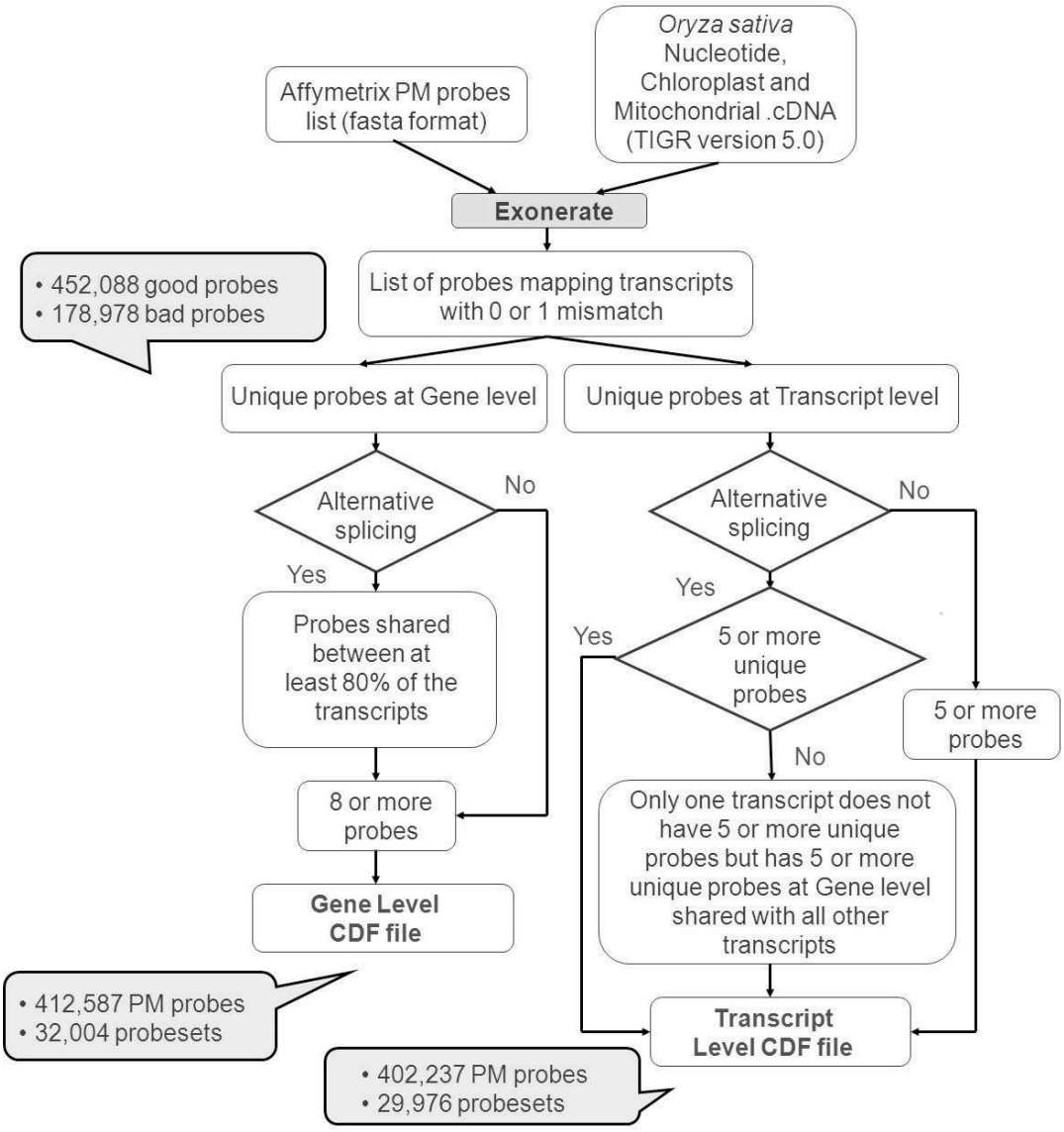


Figure A.1: Making custom made CDF file flowchart - A step by step algorithm creating a gene/transcript level custom made CDF file.

Appendix B

Supplemental Table 1

experiment_name	experiment_description	species
GSM100439.CEL-GSM100440.CEL-GSM149409.CEL	Root_Azucenagrown in hydroponics for one week with 0ppm arsenate_(ctrl)	osa
GSM100441.CEL-GSM100442.CEL-GSM149410.CEL	Root_Azucena grown in hydroponics for one week with 1ppm arsenate	osa
GSM100443.CEL-GSM100444.CEL-GSM149411.CEL	Root_Bala grown in hydroponics for one week with 0ppm arsenate_(ctrl)	osa
GSM100445.CEL-GSM100446.CEL-GSM149412.CEL	Root_Bala grown in hydroponics for one week with 1ppm arsenate	osa
GSM154829.CEL-GSM154831.CEL-GSM154832.CEL	Root_DMSO treatment_30min	osa
GSM154939.CEL-GSM154940.CEL-GSM154941.CEL	Root_DMSO treatment_120min	osa
GSM154942.CEL-GSM154943.CEL-GSM154944.CEL	Root_tZ treatment_120min	osa
GSM154945.CEL-GSM154946.CEL-GSM154947.CEL	Leaf_DMSO treatment_30min	osa
GSM154954.CEL-GSM154955.CEL-GSM154956.CEL	Leaf_tZ treatment_120min	osa
GSM154957.CEL-GSM154958.CEL	Leaf_(ctrl)	osa
GSM154959.CEL-GSM154960.CEL	Leaf_OsRR6 overexpressor	osa
GSM159172.CEL-GSM159173.CEL	Coleoptiles_4 days old_aerobic	osa
GSM159271.CEL-GSM159272.CEL	Coleoptiles_4 days old_anoxic	osa
GSM159177.CEL-GSM159178.CEL-GSM159179.CEL	Root_7 day old_light grown_seedlings	osa
GSM159180.CEL-GSM159181.CEL-GSM159182.CEL	Leaf_Mature	osa
GSM159183.CEL-GSM159184.CEL-GSM159185.CEL	Leaf_Y leaf subtending the shoot apical meristem from mature plants	osa
GSM159186.CEL-GSM159187.CEL-GSM159188.CEL	Shoot apical meristem_SAM upto 0.5 mm at floral transition stage	osa
GSM159189.CEL-GSM159190.CEL-GSM159191.CEL	Inflorescence_young inflorescence upto 3 cm	osa
GSM159192.CEL-GSM159193.CEL-GSM159194.CEL	Inflorescence_3to5 cm	osa
GSM159195.CEL-GSM159196.CEL-GSM159197.CEL	Inflorescence_5to10 cm	osa
GSM159198.CEL-GSM159199.CEL-GSM159200.CEL	Inflorescence_10to15 cm	osa
GSM159201.CEL-GSM159202.CEL-GSM159203.CEL	Inflorescence_15to22 cm	osa
GSM159204.CEL-GSM159205.CEL-GSM159206.CEL	Inflorescence_22to30 cm	osa
GSM159207.CEL-GSM159208.CEL-GSM159209.CEL	Seed_0to2 days after pollination (dap)	osa
GSM159210.CEL-GSM159211.CEL-GSM159212.CEL	Seed_3to4 days after pollination (dap)	osa
GSM159213.CEL-GSM159214.CEL-GSM159215.CEL	Seed_5to10 days after pollination (dap)	osa
GSM159216.CEL-GSM159217.CEL-GSM159218.CEL	Seed_11to20 days after pollination (dap)	osa
GSM159219.CEL-GSM159220.CEL-GSM159221.CEL	Seed_21to29 days after pollination (dap)	osa

GSM159259.CEL-GSM159260.CEL-GSM159261.CEL	Seedling_7 day old	osa
GSM159262.CEL-GSM159263.CEL-GSM159264.CEL	Drought stress_Seedling_7 day old	osa
GSM159265.CEL-GSM159266.CEL-GSM159267.CEL	Salt stress_Seedling_7 day old	osa
GSM159268.CEL-GSM159269.CEL-GSM159270.CEL	Cold stress_Seedling_7 day old	osa
GSM173080.CEL-GSM173086.CEL-GSM173089.CEL	MH63_(ctrl)	osa
GSM174883.CEL-GSM174888.CEL	mock_3 dpi leaf	osa
GSM174884.CEL-GSM174890.CEL	mock_4 dpi leaf	osa
GSM174885.CEL-GSM174887.CEL	FR13_3 dpi leaf	osa
GSM174886.CEL-GSM174889.CEL	FR13_4 dpi leaf	osa
GSM195218.CEL-GSM195219.CEL-GSM195220.CEL	Stigma	osa
GSM195221.CEL-GSM195222.CEL-GSM195223.CEL	Ovary_Stigma	osa
GSM195224.CEL	Suspension cell_Stigma	osa
GSM195225.CEL	Shoot_Stigma	osa
GSM195226.CEL	Root_Stigma	osa
GSM195227.CEL	Anther_Stigma	osa
GSM195228.CEL	Embryo_Stigma	osa
GSM195229.CEL	Endosperm_Stigma	osa
GSM195230.CEL	Seed_5day_Stigma	osa
GSM240994.CEL-GSM240995.CEL-GSM240996.CEL	Zhonghua11_7 DAF_(ctrl)	osa
GSM261998.CEL-GSM261999.CEL	Root_cultivar IAC165_uninfected roots_2days after mock inoculation	osa
GSM262002.CEL-GSM262003.CEL	Root_cultivar IAC165_uninfected roots_4days after mock inoculation	osa
GSM262004.CEL-GSM262005.CEL	Root_cultivar IAC165_roots infected with Striga hermonthica at 4days post inoculation	osa
GSM262006.CEL-GSM262007.CEL	Root_cultivar IAC165_uninfected roots 11days after mock inoculation	osa
GSM262008.CEL-GSM262009.CEL	Root_cultivar IAC165_roots infected with Striga hermonthica at 11days post inoculation	osa
GSM262010.CEL-GSM262011.CEL	Root_cultivar Nipponbare_uninfected roots_2days after mock inoculation	osa
GSM262014.CEL-GSM262015.CEL	Root_cultivar Nipponbare_uninfected roots_4days after mock inoculation	osa

GSM262016.CEL-GSM262017.CEL	Root_cultivar Nipponbare_roots infected with Striga hermonthica at 4days post inoculation	osa
GSM262018.CEL-GSM262019.CEL	Root_cultivar Nipponbare_uninfected roots_11days after mock inoculation	osa
GSM262020.CEL-GSM262021.CEL	Root_cultivar Nipponbare_roots infected with Striga hermonthica at 11days post inoculation	osa
GSM267999.CEL	Leaf_green leaf tip in the vegetative stage	osa
GSM275405.CEL-GSM275406.CEL-GSM275407.CEL	Root_Azucena root tip_in contact with 60% wax layer	osa
GSM275411.CEL-GSM275412.CEL-GSM275413.CEL	Root_Azucena root tip_blucked in 60% wax layer	osa
GSM275414.CEL-GSM275415.CEL-GSM275416.CEL	Root_Azucena root tip_above 60% wax layer	osa
GSM357122.CEL-GSM357133.CEL-GSM357134.CEL	Seedling_14 day old_light grown	osa
GSM357135.CEL-GSM357136.CEL-GSM357137.CEL	Seedling_heat shock_14 day old_light grown	osa
ATGE_1	Cotyledons	ath
ATGE_10	Rosette leaf #4	ath
ATGE_12	Rosette leaf #2	ath
ATGE_16	Rosette leaf #10	ath
ATGE_17	Rosette leaf #12	ath
ATGE_19	Leaf 7,petiol	ath
ATGE_22	Developmental drift, entire rosette after transition to flowering but before bolting	ath
ATGE_25	Leaf, senescing leaves	ath
ATGE_26	Leaf, cauline leaves	ath
ATGE_27	Stem, 2nd internode	ath
ATGE_28	Leaf, 1st node	ath
ATGE_29	Shoot apex, inflorescence (after bolting)	ath
ATGE_3	Roots	ath
ATGE_31	Flowers stage 9	ath
ATGE_32	Flowers stage 10, 11	ath
ATGE_33	Flowers stage 12	ath
ATGE_34	Flowers stage 12 sepals	ath
ATGE_35	Flowers stage 12 petals	ath

ATGE_36	Flowers stage 12 stamens	ath
ATGE_37	Flowers stage 12 carpels	ath
ATGE_39	Flowers stage 15	ath
ATGE_40	Flowers stage 15 pedicels	ath
ATGE_41	Flowers stage 15 sepals	ath
ATGE_42	Flowers stage 15 petals	ath
ATGE_43	Flowers stage 15 stamen	ath
ATGE_45	Flowers stage 15 carpels	ath
ATGE_5	Leaf	ath
ATGE_6	Shoot apex, vegetative	ath
ATGE_7	Seedling, green parts	ath
ATGE_76	Seed stage 3, w/o siliques	ath
ATGE_77	Seed stage 4, w/o siliques	ath
ATGE_78	Seed stage 5, w/o siliques	ath
ATGE_79	Seed stage 6, w/o siliques	ath
ATGE_81	Seed stage 7, w/o siliques	ath
ATGE_82	Seed stage 8, w/o siliques	ath
ATGE_90	Vegetative rosette	ath
ATGE_96	Seedling green parts	ath
ATGE_97	Seedling green parts	ath
ATGE_98	Root	ath
ATGE_99	Root	ath
APL	Phloem	ath
C1	Cortex	ath
GL2	Epidermis	ath
J0571	Ground tissue	ath
J3411	Lateral Root Cap	ath
PET111	Columella	ath
S18	Xylem	ath
scr5	Endodermis	ath
StageI	Root_tip	ath

StageII	Root_tip	ath
StageIII	Root	ath
wol	Stele	ath
RootE002	Root_0h_(ctrl)	ath
RootE152	Root_12h_cold stress	ath
RootE162	Root_24h_cold stress	ath
RootE232	Root_3h_osmotic stress	ath
RootE332	Root_3h_salt stress	ath
RootE342	Root_6h_salt stress	ath
ShootE001	Shoot_0h_(ctrl)	ath
ShootE041	Shoot_6h_(ctrl)	ath
ShootE051	Shoot_12h_(ctrl)	ath
ShootE131	Shoot_3h_cold stress	ath
ShootE141	Shoot_6h_cold stress	ath
ShootE151	Shoot_12h_cold stress	ath
ShootE161	Shoot_24h_cold stress	ath
ShootE221	Shoot_1h_osmotic stress	ath
ShootE231	Shoot_3h_osmotic stress	ath
ShootE241	Shoot_6h_osmotic stress	ath
ShootE251	Shoot_12h_osmotic stress	ath
ShootE261	Shoot_24h_osmotic stress	ath
ShootE331	Shoot_3h_salt stress	ath
ShootE341	Shoot_6h_salt stress	ath
ShootE351	Shoot_12h_salt stress	ath
ShootE361	Shoot_24h_salt stress	ath
ShootE411	Shoot_0.5h_draught stress	ath
ShootE431	Shoot_3h_draught stress	ath

Table B.1: Overview of microarray experiments included in the expression compendia for Arabidopsis and rice.

Appendix C

Supplemental Table 2

ALL			**CONS**				**DIV**				**NON**			
Bin	*Frequency*	*Cumulative*	*Bin*	*Frequency*	*Cumulative*	*Relative*	*Bin*	*Frequency*	*Cumulative*	*Relative*	*Bin*	*Frequency*	*Cumulative*	*Relative*
0.05	870	18.79%	0.05	238	6.69%	27.36%	0.05	230	58.52%	26.44%	0.05	402	59.12%	46.21%
0.1	969	39.72%	0.1	610	23.83%	62.95%	0.1	142	94.66%	14.65%	0.1	218	91.18%	22.50%
0.15	640	53.54%	0.15	559	39.54%	87.34%	0.15	21	100.00%	3.28%	0.15	60	100.00%	9.38%
0.2	443	63.11%	0.2	443	52.00%	100.00%	0.2	0	100.00%	0.00%	0.2	0	100.00%	0.00%
0.25	301	69.61%	0.25	301	60.46%	100.00%	0.25	0	100.00%	0.00%	0.25	0	100.00%	0.00%
0.3	238	74.75%	0.3	238	67.14%	100.00%	0.3	0	100.00%	0.00%	0.3	0	100.00%	0.00%
0.35	259	80.35%	0.35	259	74.42%	100.00%	0.35	0	100.00%	0.00%	0.35	0	100.00%	0.00%
0.4	321	87.28%	0.4	321	83.45%	100.00%	0.4	0	100.00%	0.00%	0.4	0	100.00%	0.00%
0.45	511	98.32%	0.45	511	97.81%	100.00%	0.45	0	100.00%	0.00%	0.45	0	100.00%	0.00%
0.5	78	100.00%	0.5	78	100.00%	100.00%	0.5	0	100.00%	0.00%	0.5	0	100.00%	0.00%
0.55	0	100.00%	0.55	0	100.00%		0.55	0	100.00%		0.55	0	100.00%	
0.6	0	100.00%	0.6	0	100.00%		0.6	0	100.00%		0.6	0	100.00%	
0.65	0	100.00%	0.65	0	100.00%		0.65	0	100.00%		0.65	0	100.00%	
0.7	0	100.00%	0.7	0	100.00%		0.7	0	100.00%		0.7	0	100.00%	
0.75	0	100.00%	0.75	0	100.00%		0.75	0	100.00%		0.75	0	100.00%	
0.8	0	100.00%	0.8	0	100.00%		0.8	0	100.00%		0.8	0	100.00%	
0.85	0	100.00%	0.85	0	100.00%		0.85	0	100.00%		0.85	0	100.00%	
0.9	0	100.00%	0.9	0	100.00%		0.9	0	100.00%		0.9	0	100.00%	
0.95	0	100.00%	0.95	0	100.00%		0.95	0	100.00%		0.95	0	100.00%	
1	0	100.00%	1	0	100.00%		1	0	100.00%		1	0	100.00%	
More	0	100.00%	More	0	100.00%		More	0	100.00%		More	0	100.00%	
	4630			76.85%				8.49%				14.69%		

Table C.1: ECC distributions for all 1:1 orthologous genes.

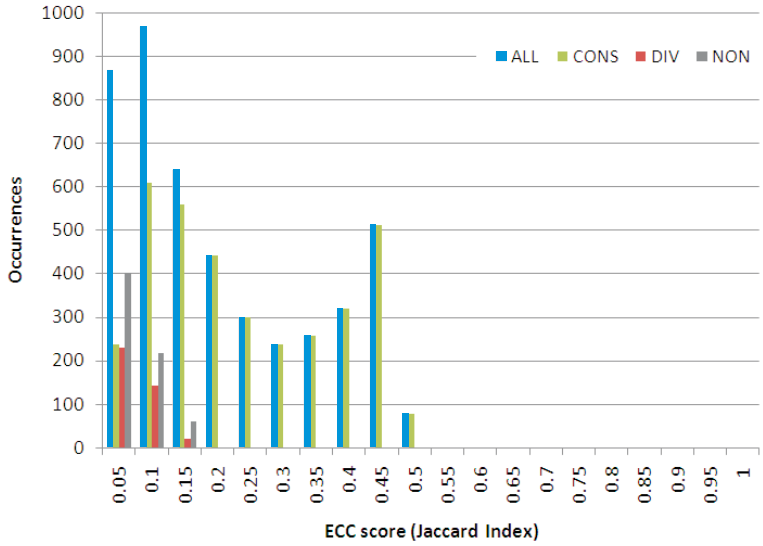


Figure C.1: Supplemental Figure 2.1 - Distribution of ECC scores per ECC category.

C. SUPPLEMENTAL TABLE 2

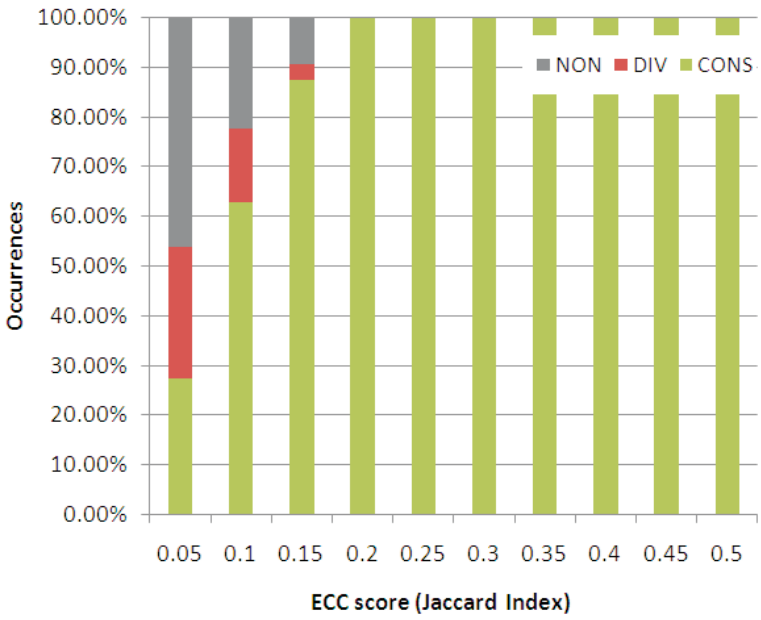


Figure C.2: Supplemental Figure 2.2 - Relative distribution of ECC scores per ECC category.

Appendix D

Supplemental Table 3

Species	GO type	GO term	Enrichment	p-value	Description
A. thaliana	BP	GO:0006508	1.88	5.66E-6	proteolysis
A. thaliana	CC	GO:0005737	0.85	7.72E-8	cytoplasm
A. thaliana	CC	GO:0043231	0.68	1.05E-5	intracellular membrane-bounded organelle
A. thaliana	CC	GO:0009536	0.98	6.92E-5	plastid
A. thaliana	CC	GO:0009507	0.99	8.09E-5	chloroplast
Z. mays	CC	GO:0009536	1.98	6.52E-6	plastid
M. truncatula	MF	GO:0005488	1.31	7.44E-80	binding
M. truncatula	MF	GO:0003824	1.37	2.44E-63	catalytic activity
M. truncatula	MF	GO:0005515	1.39	8.58E-38	protein binding
M. truncatula	MF	GO:0016787	1.49	3.67E-22	hydrolase activity
M. truncatula	MF	GO:0030528	1.61	7.89E-18	transcription regulator activity
M. truncatula	MF	GO:0016491	1.68	8.62E-15	oxidoreductase activity
M. truncatula	MF	GO:0003700	1.76	4.69E-12	transcription factor activity
M. truncatula	MF	GO:0042802	1.76	1.40E-9	identical protein binding
M. truncatula	MF	GO:0016788	1.82	2.01E-9	hydrolase activity, acting on ester bonds
M. truncatula	MF	GO:0005215	1.36	6.23E-7	transporter activity
M. truncatula	MF	GO:0022857	1.51	8.38E-7	transmembrane transporter activity
M. truncatula	MF	GO:0022804	1.81	1.00E-6	active transmembrane transporter activity
M. truncatula	MF	GO:0015291	2.21	1.69E-6	secondary active transmembrane transporter activity
M. truncatula	MF	GO:0015297	2.70	2.14E-6	antiporter activity
M. truncatula	MF	GO:0000166	1.46	3.18E-6	nucleotide binding
M. truncatula	MF	GO:0022892	1.43	2.82E-5	substrate-specific transporter activity
M. truncatula	BP	GO:0008152	1.39	1.38E-76	metabolic process
M. truncatula	BP	GO:0009987	1.27	9.25E-53	cellular process
M. truncatula	BP	GO:0050896	1.65	4.70E-25	response to stimulus
M. truncatula	BP	GO:0006807	1.47	9.83E-24	nitrogen compound metabolic process
M. truncatula	BP	GO:0065007	1.56	1.27E-22	biological regulation
M. truncatula	BP	GO:0019222	1.72	9.91E-18	regulation of metabolic process

M. truncatula	BP	GO:0060255	1.72	2.72E-17	regulation of macromolecule metabolic process
M. truncatula	BP	GO:0019219	1.72	9.12E-16	regulation of cellular nucleobase, nucleoside, nucleotide and nucleic acid metabolic process
M. truncatula	BP	GO:0044281	1.71	6.18E-12	small molecule metabolic process
M. truncatula	BP	GO:0055114	1.64	7.48E-11	oxidation reduction
M. truncatula	BP	GO:0042221	1.65	6.58E-10	response to chemical stimulus
M. truncatula	BP	GO:0034641	2.02	1.76E-6	cellular nitrogen compound metabolic process
M. truncatula	BP	GO:0006519	2.04	2.60E-6	cellular amino acid and derivative metabolic process
M. truncatula	BP	GO:0006810	1.20	3.24E-6	transport
M. truncatula	BP	GO:0009628	1.73	3.54E-6	response to abiotic stimulus
M. truncatula	BP	GO:0055085	1.58	8.82E-5	transmembrane transport
M. truncatula	CC	GO:0005623	1.53	1.33E-76	cell
M. truncatula	CC	GO:0044464	1.53	1.33E-76	cell part
M. truncatula	CC	GO:0005622	1.72	4.66E-61	intracellular
M. truncatula	CC	GO:0043227	1.91	8.63E-56	membrane-bounded organelle
M. truncatula	CC	GO:0005737	1.93	1.29E-47	cytoplasm
M. truncatula	CC	GO:0009507	2.72	3.53E-46	chloroplast
M. truncatula	CC	GO:0044434	2.99	1.10E-25	chloroplast part
M. truncatula	CC	GO:0009579	3.18	1.00E-17	thylakoid
M. truncatula	CC	GO:0044436	3.40	1.27E-15	thylakoid part
M. truncatula	CC	GO:0009534	3.34	7.41E-14	chloroplast thylakoid
M. truncatula	CC	GO:0031976	3.34	7.41E-14	plastid thylakoid
M. truncatula	CC	GO:0031984	3.34	7.41E-14	organelle subcompartment
M. truncatula	CC	GO:0031090	2.56	4.44E-13	organelle membrane
M. truncatula	CC	GO:0031967	2.16	7.39E-7	organelle envelope
M. truncatula	CC	GO:0009532	2.43	6.99E-6	plastid stroma
P. trichocarpa	MF	GO:0003676	1.3	2.67E-27	nucleic acid binding

P. trichocarpa	MF	GO:0003677	1.39	7.05E-20	DNA binding
P. trichocarpa	MF	GO:0030528	1.5	2.04E-15	transcription regulator activity
P. trichocarpa	MF	GO:0003700	1.53	1.24E-13	transcription factor activity
P. trichocarpa	MF	GO:0003824	0.58	2.84E-10	catalytic activity
P. trichocarpa	MF	GO:0005515	0.97	4.43E-09	protein binding
P. trichocarpa	MF	GO:0043565	1.76	3.28E-06	sequence-specific DNA binding
P. trichocarpa	MF	GO:0016787	0.86	2.61E-05	hydrolase activity
P. trichocarpa	BP	GO:0009987	0.77	6.49E-27	cellular process
P. trichocarpa	BP	GO:0050789	1.38	9.95E-24	regulation of biological process
P. trichocarpa	BP	GO:0044237	0.87	1.24E-23	cellular metabolic process
P. trichocarpa	BP	GO:0044238	0.8	7.33E-23	primary metabolic process
P. trichocarpa	BP	GO:0043170	0.87	4.88E-20	macromolecule metabolic process
P. trichocarpa	BP	GO:0044260	0.92	1.38E-19	cellular macromolecule metabolic process
P. trichocarpa	BP	GO:0080090	1.48	3.08E-19	regulation of primary metabolic process
P. trichocarpa	BP	GO:0031323	1.48	7.09E-19	regulation of cellular metabolic process
P. trichocarpa	BP	GO:0060255	1.45	1.03E-18	regulation of macromolecule metabolic process
P. trichocarpa	BP	GO:0051171	1.47	4.03E-18	regulation of nitrogen compound metabolic process
P. trichocarpa	BP	GO:0006350	1.44	7.34E-18	cellular transcription
P. trichocarpa	BP	GO:0061018	1.44	7.34E-18	transcription
P. trichocarpa	BP	GO:0009889	1.46	1.49E-17	regulation of biosynthetic process
P. trichocarpa	BP	GO:0050896	1.19	7.53E-16	response to stimulus
P. trichocarpa	BP	GO:0042221	1.44	3.12E-11	response to chemical stimulus
P. trichocarpa	BP	GO:0009719	1.94	5.23E-11	response to endogenous stimulus
P. trichocarpa	BP	GO:0010033	1.7	8.23E-11	response to organic substance
P. trichocarpa	BP	GO:0009725	1.98	1.79E-10	response to hormone stimulus
P. trichocarpa	BP	GO:0051716	2.04	4.17E-10	cellular response to stimulus
P. trichocarpa	BP	GO:0006351	1.47	9.08E-09	cellular transcription, DNA-dependent
P. trichocarpa	BP	GO:0032774	1.47	9.08E-09	RNA biosynthetic process
P. trichocarpa	BP	GO:0061022	1.47	9.08E-09	transcription, DNA-dependent

P. trichocarpa	BP	GO:0070887	2.51	1.91E-08	cellular response to chemical stimulus
P. trichocarpa	BP	GO:0009414	2.75	1.92E-08	response to water deprivation
P. trichocarpa	BP	GO:0071495	2.7	3.31E-08	cellular response to endogenous stimulus
P. trichocarpa	BP	GO:0009755	2.74	1.96E-07	hormone-mediated signaling pathway
P. trichocarpa	BP	GO:0032870	2.74	1.96E-07	cellular response to hormone stimulus
P. trichocarpa	CC	GO:0005623	0.74	1.80E-21	cell
P. trichocarpa	CC	GO:0044464	0.74	1.80E-21	cell part
P. trichocarpa	CC	GO:0005634	1.37	2.82E-15	nucleus
O. sativa	MF	GO:0003824	0.49	1.11E-20	catalytic activity
O. sativa	MF	GO:0004672	1.05	1.78E-20	protein kinase activity
O. sativa	MF	GO:0005524	1.18	2.07E-06	ATP binding
O. sativa	MF	GO:0004713	1.47	9.67E-06	protein tyrosine kinase activity
O. sativa	BP	GO:0009987	0.61	1.69E-38	cellular process
O. sativa	BP	GO:0008219	1.88	1.04E-34	cell death
O. sativa	BP	GO:0016265	1.88	1.04E-34	death
O. sativa	BP	GO:0006915	1.89	1.75E-34	apoptosis
O. sativa	BP	GO:0043170	0.63	2.14E-26	macromolecule metabolic process
O. sativa	BP	GO:0044260	0.63	8.26E-23	cellular macromolecule metabolic process
O. sativa	BP	GO:0044238	0.51	1.19E-20	primary metabolic process
O. sativa	BP	GO:0006468	1.05	2.31E-20	protein amino acid phosphorylation
O. sativa	BP	GO:0006952	1.63	7.83E-19	defense response
O. sativa	BP	GO:0006807	0.6	9.13E-11	nitrogen compound metabolic process
O. sativa	BP	GO:0006139	0.64	2.00E-10	cellular nucleobase, nucleoside, nucleotide and nucleic acid metabolic process
O. sativa	BP	GO:0034660	1.83	4.49E-07	ncRNA metabolic process
O. sativa	BP	GO:0006399	1.92	1.29E-06	tRNA metabolic process
O. sativa	BP	GO:0008037	1.8	3.13E-05	cell recognition
O. sativa	BP	GO:0009856	1.8	3.13E-05	pollination
O. sativa	BP	GO:0009875	1.8	3.13E-05	pollen-pistil interaction
O. sativa	BP	GO:0048544	1.8	3.13E-05	recognition of pollen
O. sativa	CC	GO:0044464	0.77	4.51E-217	cell part

O. sativa	CC	GO:0043227	0.87	2.27E-203	membrane-bounded organelle
O. sativa	CC	GO:0043231	0.87	3.15E-203	intracellular membrane-bounded organelle
O. sativa	CC	GO:0044444	0.86	1.67E-186	cytoplasmic part
O. sativa	CC	GO:0009536	1.25	8.18E-108	plastid
O. sativa	CC	GO:0005739	0.91	4.13E-50	mitochondrion
O. sativa	CC	GO:0031988	0.54	9.20E-16	membrane-bounded vesicle
G. max	MF	GO:0019209	9.15	4.18E-4	kinase activator activity
G. max	MF	GO:0030295	9.15	4.18E-4	protein kinase activator activity
V. vinifera	MF	GO:0005488	0.37	2.14E-13	binding
V. vinifera	BP	GO:0008152	0.43	4.25E-26	metabolic process
V. vinifera	BP	GO:0009987	0.45	1.30E-22	cellular process
V. vinifera	BP	GO:0044238	0.44	5.45E-17	primary metabolic process
V. vinifera	BP	GO:0009058	0.56	3.37E-10	biosynthetic process
V. vinifera	BP	GO:0006810	0.76	5.24E-10	transport
V. vinifera	BP	GO:0044281	0.86	1.01E-08	small molecule metabolic process
V. vinifera	BP	GO:0046686	1.95	2.71E-06	response to cadmium ion
V. vinifera	BP	GO:0015031	1.31	3.86E-06	protein transport
V. vinifera	BP	GO:0045184	1.31	3.86E-06	establishment of protein localization
V. vinifera	BP	GO:0010467	0.53	2.83E-05	gene expression
V. vinifera	BP	GO:0006807	0.47	3.40E-05	nitrogen compound metabolic process
V. vinifera	BP	GO:0065007	0.54	3.46E-05	biological regulation
V. vinifera	BP	GO:0050789	0.58	5.73E-05	regulation of biological process
V. vinifera	BP	GO:0006082	0.92	9.12E-05	organic acid metabolic process
V. vinifera	BP	GO:0009056	1.03	9.56E-05	catabolic process
V. vinifera	CC	GO:0005623	0.54	4.59E-29	cell
V. vinifera	CC	GO:0044464	0.54	4.59E-29	cell part
V. vinifera	CC	GO:0005622	0.65	9.78E-25	intracellular
V. vinifera	CC	GO:0005737	0.9	3.35E-24	cytoplasm
V. vinifera	CC	GO:0044444	0.9	2.56E-20	cytoplasmic part
V. vinifera	CC	GO:0043226	0.65	1.27E-17	organelle
V. vinifera	CC	GO:0043227	0.7	2.13E-17	membrane-bounded organelle

V. vinifera	CC	GO:0016020	0.59	3.27E-12	membrane
V. vinifera	CC	GO:0005886	1.06	2.02E-09	plasma membrane
V. vinifera	CC	GO:0005739	1.41	2.74E-09	mitochondrion
V. vinifera	CC	GO:0005773	1.56	4.26E-07	vacuole
V. vinifera	CC	GO:0044429	1.7	3.43E-05	mitochondrial part

Table D.1: GO enrichment results for leaf tissue specific genes ($P < 1.0$E-4, except for G.max).

D. SUPPLEMENTAL TABLE 3

Appendix E

Supplemental Figure 1

E. SUPPLEMENTAL FIGURE 1

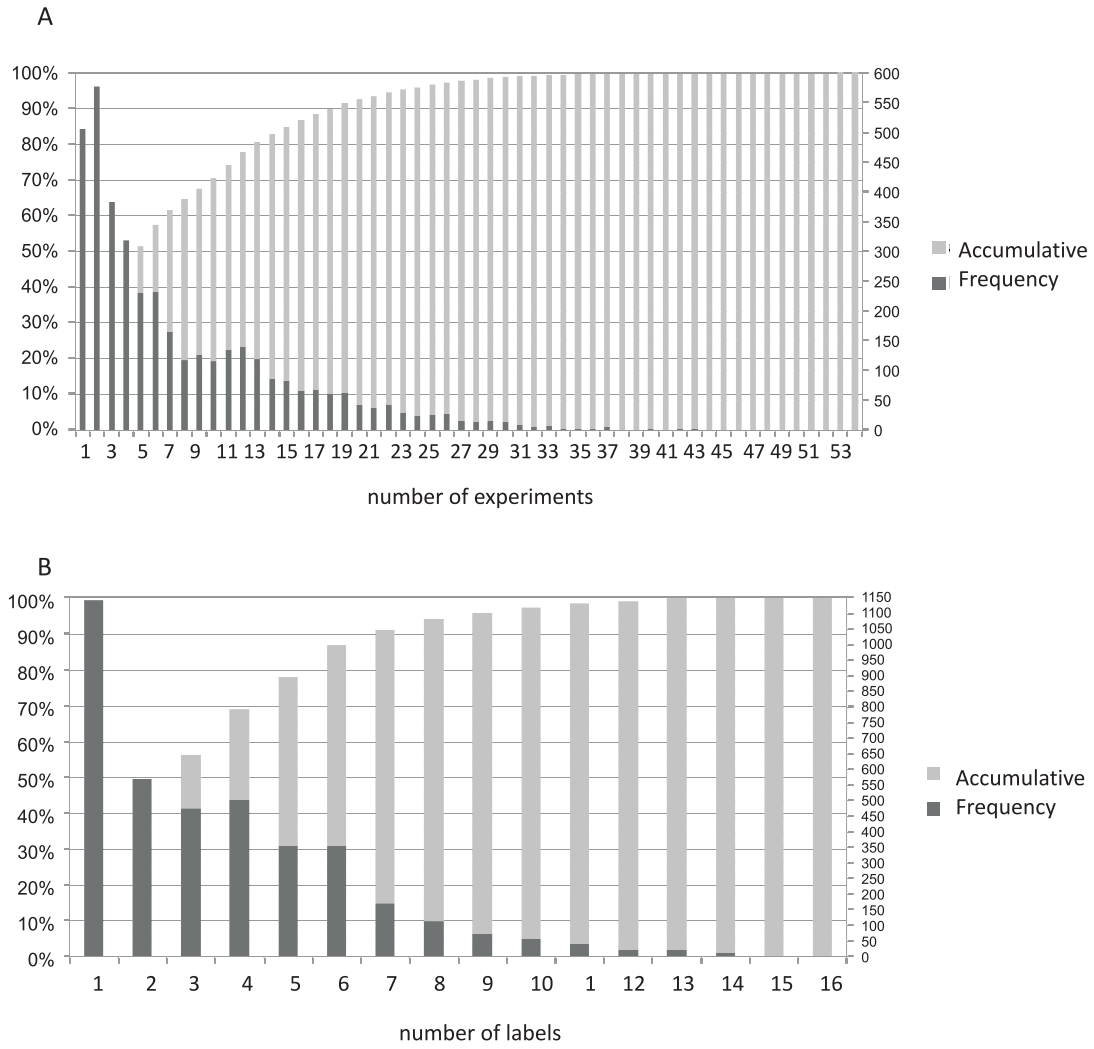


Figure E.1: Frequency of labels and experiments in tissue specific genes - Frequency of tissue specific experiments and labels after applying step.2. A) Green bars show the frequency of experiments defined for each tissue specific gene after applying step.2 and orange bars present the accumulative data, B) Green bars show the frequency of labels defined for each tissue specific gene after applying step.2 and orange bars present the accumulative data.

Appendix F

List of Publications

1. **Comparative co-expression analysis in plant biology** Movahedi, S. and Van Bel, M. and Heyndrickx, K.S. and Vandepoele, K. (2012) Plant, Cell and Environment. doi: 10.1111/j.1365-3040.2012.02517.x

2. **Dissecting plant genomes with the PLAZA comparative genomics platform.** Van Bel, M., Proost, S., Wischnitzki, E., Mohavedi, S., Scheerlinck, S., Van de Peer, Y., and Vandepoele, K. (2011) Plant Physiology. 158(2):590-600 doi:10.1104/pp.111.189514

3. **Comparative network analysis reveals that tissue specificity and gene function are important factors influencing the mode of expression evolution in Arabidopsis and rice.** Movahedi, S., Van de Peer, Y., Vandepoele, K. (2011) Plant Physiology. 156(3), 1316-1330. doi:10.1104/pp.111.177865

4. **Genome-scale computational analysis of DNA curvature and repeats in Arabidopsis and rice uncovers plant-specific genomic properties.** Masoudi-Nejad, A.*, Movahedi, S.*, Jauregui, R. (2011) BMC Genomics. 12,214. doi:10.1186/1471-2164-12-214 (* Contributed equally)

5. **CORNET: a user-friendly tool for data mining and integration.** De Bodt, S., Carvajal, D., Hollunder, J., Van den Cruyce, J., Movahedi, S., Inz, D. (2010) Plant Physiology. 152(3), 1167-79. doi:10.1104/pp.109.147215

F. LIST OF PUBLICATIONS

Appendix G

Curriculum Vitae

EDUCATION

- *Ph.D in Biotechnology (Bioinformatics):* 2008 to 2011 - VIB Department of Biotechnology and Bioinformatics, Ghent University, Belgium
- *M.S in Computer Science:* 2005 to 2007 - Sharif University of Technology, Iran
- *B.S in Computer Science:* 2001 to 2005 - Tehran University of Technology, Iran

PROFESSIONAL EXPERIENCE

- *Researcher in Bioinformatics*
  Next Generation Sequencing data analysis.
  Rijk Zwaan Plant Breeding Company, Fijnaart, the Netherlands (Since 2012)
- *Comparative transcriptomics in plants (PhD)*
  Developing high quality CDF packages for Affymetrix microarrays, building coexpression clusters, defining conserved coexpression networks, performing functional enrichment analysis, applying different statistical analysis on the collected data and studying different biological aspects of conserved coexpression networks which involved developing various analyzing tools and methods using R/Bioconductor and Perl programming language.
  VIB, Department of Biotechnology and Bioinformatics, Ghent University (2008-2011)
  Supervisors: Prof. Dr. Yves Van de Peer and Prof. Dr. Klaas Vandepoele

G. CURRICULUM VITAE

- *CurvaCpG: A genome-scale web based Bioinformatics service (Master)*
  Genome-scale computational analysis of DNA curvature and repeats in Arabidopsis and rice
  Institute of Biophysics and Biochemistry (IBB), Tehran, Iran (2005-2007)
  Supervisor: Prof. Dr. Ali Masoudi-Nejad

- *Using PERL programming language for Genetic Algorithms (Bachelor)*
  Tehran University - Faculty of Science, Department of Mathematics and Computer Science (2001-2005)
  Supervisor: Prof. Dr. Hayedeh Ahrabian