Integrative transcriptomic approaches to analyzing plant co-expression networks

Dissertation

zur Erlangung des akademischen Grades

"doctor rerum naturalium" (Dr. rer. nat.)

eingereicht im

Institut für Biochemie und Biologie an der

Mathematisch-Naturwissenschaftlichen Fakultät der

Universität Potsdam

Marek Mutwil

Arbeitsgruppe Persson

Max-Plank-Institut für Molekulare Pflanzenphysiologie

Potsdam, den 25.07.2010

Published online at the
Institutional Repository of the University of Potsdam:
URL http://opus.kobv.de/ubp/volltexte/2011/5075/
URN urn:nbn:de:kobv:517-opus-50752
http://nbn-resolving.org/urn:nbn:de:kobv:517-opus-50752

Acknowledgments

Staffan Persson is certainly one of the best supervisors I've ever had. He gave me the freedom to pursue my own ideas and goals whilst providing all the necessary help and support when needed. To work at such an institute, in the company of great scientific minds, has helped my scientific thinking through a colossal improvement. I still have far to go, but I start my post-doctoral efforts from a very solid basis.

I would also like to thank my official university supervisor Ralph Bock and progress commity member Dirk Walther, for superb guidance during my PHD. Thanks to the whole of AG Persson for the good times and great company.

I thank my family for their support; Mum, Jens and the rest of my family, in Denmark and Poland.

Contents

Abstract

It is well documented that transcriptionally coordinated genes tend to be functionally related, and that such relationships may be conserved across different species, and even kingdoms. (Ihmels et al., 2004). Such relationships was initially utilized to reveal functional gene modules in yeast and mammals (Ihmels et al., 2004), and to explore orthologous gene functions between different species and kingdoms (Stuart et al., 2003; Bergmann et al., 2004).

Model organisms, such as *Arabidopsis*, are readily used in basic research due to resource availability and relative speed of data acquisition. A major goal is to transfer the acquired knowledge from these model organisms to species that are of greater importance to our society. However, due to large gene families in plants, the identification of functional equivalents of well characterized *Arabidopsis* genes in other plants is a non-trivial task, which often returns erroneous or inconclusive results.

In this thesis, concepts of utilizing co-expression networks to help infer (i) gene function, (ii) organization of biological processes and (iii) knowledge transfer between species are introduced. An often overlooked fact by bioinformaticians is that a bioinformatic method is as useful as its accessibility. Therefore, majority of the work presented in this thesis was directed on developing freely available, user-friendly web-tools accessible for any biologist.

Zusammenfassung

Es ist bereits ausgiebig gezeigt worden, dass Gene, deren Expression auf Transkriptionsebene koordiniert ist, häufig auch funktional in verwandten Stoffwechselwegen vorkommen, und dass sich dies wahrscheinlich auch Spezies- und sogar Reichübergreifend sagen lässt (Ihmels et al., 2004). Anfänglich wurden solche Beziehungen verwendet, um sogenannte Genfunktionsmodule in Hefe und Säugern aufzudecken (Ihmels et al., 2004), um dann orthologe Genfunktionen zwischen verschiedene Spezies und Reichen zu entdecken (Stuart et al., 2003; Bergmann et al., 2004).

Modellorganismen wie Arabidopsis werden bevorzugt in der Forschung verwendet, weil man durch die schnelle Generationszeit in kurzer Zeit viele Daten erheben kann und aufgrund dessen die Ressourcen- und Informationsvielfalt um ein Vielfaches größer ist. Ein Hauptziel ist der Wissenstransfer von Modellorganismen auf Spezies, die gesellschaftlich von höherer Bedeutung sind wie z.B. Getreidearten oder andere Feldfrüchte. Pflanzen besitzen oft große Genfamilien und die eindeutige Identifizierung von gut charakterisierten Arabidopsisorthologen in besagten Nutzpflanzen ist kein triviales Vorhaben.

In der vorliegenden Arbeit werden Konzepte zur Nutzung von Co-expressionsnetzwerken beschrieben, die helfen sollen (i) Genfunktionen zu identifizieren, (ii) die Organisation von biologischen Prozessen aufzuklären und (iii) das erworbene Wissen auf andere Spezies übertragbar zu machen. Ein häufig von Bioinformatikern übersehender Umstand ist, dass bioinformatische Methoden nur so sinnvoll sind wie ihre Zugänglichkeit. Deshalb basiert der Großteil dieser Arbeit auf freiverfügbaren und vor allem für Biologen nutzerfreundlichen Webtools.

1. Introduction

Biology is currently one of the most rapidly evolving sciences. Cross-diciplinary developments in molecular biology, chemistry and computer science generate vast amount of biological data from high-throughput -*omics* studies. Genomics, in the form of genome sequences from various organisms, have increased our understanding of gene content, gene function and evolution. As of June 2010, over 1500 genomes from prokaryotic, eukaryotic and archae organisms have been fully sequenced, and over 5500 sequencing projects are in progress (Liolios et al., 2010). Transcriptomic studies, i.e. microarrays and deep sequencing of mRNA, reveal how internal or external perturbations affect the responses of an organism. Over 100.000 microarray experiments from various organisms are now publicly available (Rocca-Serra et al., 2003). While not as exhaustive, high through-put studies of metabolome (Hegeman et al., 2010), proteome (Premsler et al., 2009), interactome (Lievens et al., 2009) are also becoming important factors in understanding biology.

The descriptive biology of last century is slowly transforming into science that can both explain and predict complex biological systems (Kitano et al., 2005). While molecular biology of 20th century focused on the elucidating the function and interaction of single components, many studies today use *systems biology* approaches to investigate how different cellular components interact to make up a system (e.g. cell). This may be achieved by the combination of different -*omic* disciplines (Kitano, 2002). It is safe to say that no gene, nor gene product, is an island, and that each component of a cell is directly or indirectly interacting and affecting at least one other component of the cell. Systems biology attempts to model dynamic interactions between different components by measuring changes of the system in response to perturbances (Kirschner, 2002). The major tool of a systems biologist is arguably transcriptomics, which can measure simultaneous expression level of thousands of mRNAs in response to any internal or external stimuli. Indeed, just for *Arabidopsis thaliana* alone, nearly 400 experiments investigating the response of this plant to various stimuli, stresses or genotypes has been performed (Goda et al., 2008, Kilian et al., 2007).

A major discovery while analyzing the accumulating expression data was that functionally related genes tend to be transcriptionally coordinated, i.e. co-expressed (Stuart et al., 2003, Yu et al., 2003). For example, by identifying genes showing similar expression patterns to cellulose synthase (CESA) genes across 408 microarrays, Persson et al., (2005) characterized two novel genes that displayed deficiencies related to cellulose synthesis in *Arabidopsis*. Consequently, using "guilt by association" approaches, co-expression analyses have proved valuable for rapid inferences of gene functions and of biological pathway discovery (Wei et al., 2006; Yonekura-Sakakibara et al., 2008; Usadel et al., 2009).

Arabidopsis has little economic value, but is used to research dicotyledon plants for a number of reasons including a short generation time, large seed production, convenient size, a relatively small and fully sequenced genome, and the existence of well established transformation protocols. The plant has approximately 50% of the genes functionally annotated by sequence homology, and roughly 11% of the genes are associated with distinct biological functions that have been experimentally verified (Saito et al., 2008). Due to little economical value of Arabidopsis, the knowledge obtained in this plant needs to be applied to other species, which may be of greater importance for the society. However, while the exact function of a gene product in a model organism (i.e. knowledge donor) has been proven experimentally, uncovering the identity of the functional equivalent in, for example a crop plant (i.e. knowledge acceptor) is not trivial. Plants generally hold large gene families and sequence comparisons can return a large list of possible candidate genes. However, several studies have showed that co-expressed relationships are conserved across distantly related organsisms, such as yeast, mouse and human (Stuart et al., 2003; Bergmann et al., 2004). Thus, a functional homolog may be rapidly identified by combined sequence and co-expression approaches.

Today, a cell biologist faces three challenges: (i) to define the function of cellular components (e.g. proteins, metabolites), (ii) to understand how those components cooperate to form a living cell, and ultimately (iii) to transfer and apply this knowledge to any organism important for the society. The work presented in this thesis attempts to demonstrate how comparative co-expression analysis in several plant species, combined with available experimental and bioinformatical knowledge can help answer the abovementioned questions. The following chapters in the introduction provide key concepts which were used in this work.

1.1 Transcriptomics

A transcriptome is a collection of all messenger RNA molecules in a cell. The information gained from transcriptomics can provide a platform for the researchers to gain a better understanding of how genes and pathways are involved in biological processes. Although several important steps such as translation efficiency and post-translational modification are not revealed by measuring mRNA levels, transcriptomics via DNA microarrays provide a mature and affordable method for monitoring regulatory changes.

AtGenExpress consortium generated an exhaustive transcriptomic atlas of all *Arabidopsis* organs during different stages of development and during a wide range of abiotic (e.g. heat, cold, drought), biotic (infection) and hormonal (e.g. auxin, cytokinin) treatments (Schmid et al., 2005, Kilian et al., 2007, Goda et al., 2008). Apart from measuring of the mRNA levels, DNA microarray technology has also been applied to study alternative splicing (splicing arrays reviewed in Hallegger et al., 2010), genotype (SNP arrays reviewed in Gupta et al., 2008), and empirical discovery of unknown transcripts (tiling arrays reviewed in Gregory, 2009).

The traditional DNA microarrays, however, represent powerful high throughput tools to measure the expression of thousands of genes simultaneously, and have been successfully applied to study transcriptional changes during developmental programs, responses to internal (genetic) and external (stress) perturbations. DNA microarrays are usually simple glass slides with microscopic spots of DNA probes attached to the surface with each spot consisting of several identical sequences. The sequence of each probe can be targeted towards a certain gene or targeted towards some other sequence in a genome. Two types of microarrays exist: spotted cDNA microarrays utilize cDNA probes spotted onto a glass slide by a robotic “arrayer”, while oligonucleotide arrays employ *in-situ* synthesized short (~25 nucleotide long, as used by for example Affymetrix) or long (~50-70 used typically by Agilent, Illumina, and Nimblegen) oligonucleotides. Spotted cDNA and long oligonucleotide arrays permit hybridization of two samples to the same array, where treatment and control can be labeled with Cy3 (green) and Cy5 (red) fluorophores, respectively, and the differential gene expression can be directly measured by observing Cy3/Cy5 ratio. For short oligonucleotide arrays, one mRNA population per treatment is hybridized to one microarray, requiring at least two microarrays per any comparative experiment.

A workflow over how microarrays are used is presented in Figure 1.1 (reviewed in Churchill, 2002). To use a microarray for expression measurements, the investigator first

needs to formulate the set up of the experiment. Often, a researcher compares two cell populations; one from treatment/disease and one from control. To avoid effects of technical variance (e.g. imperfections of array manufacturing) and biological variance (e.g. discrepancy in responses of two genetically equal cell cultures/organisms to a given treatment), it is generally accepted to use 3 microarrays for each measurement, with mRNA from one or more independent cell cultures.

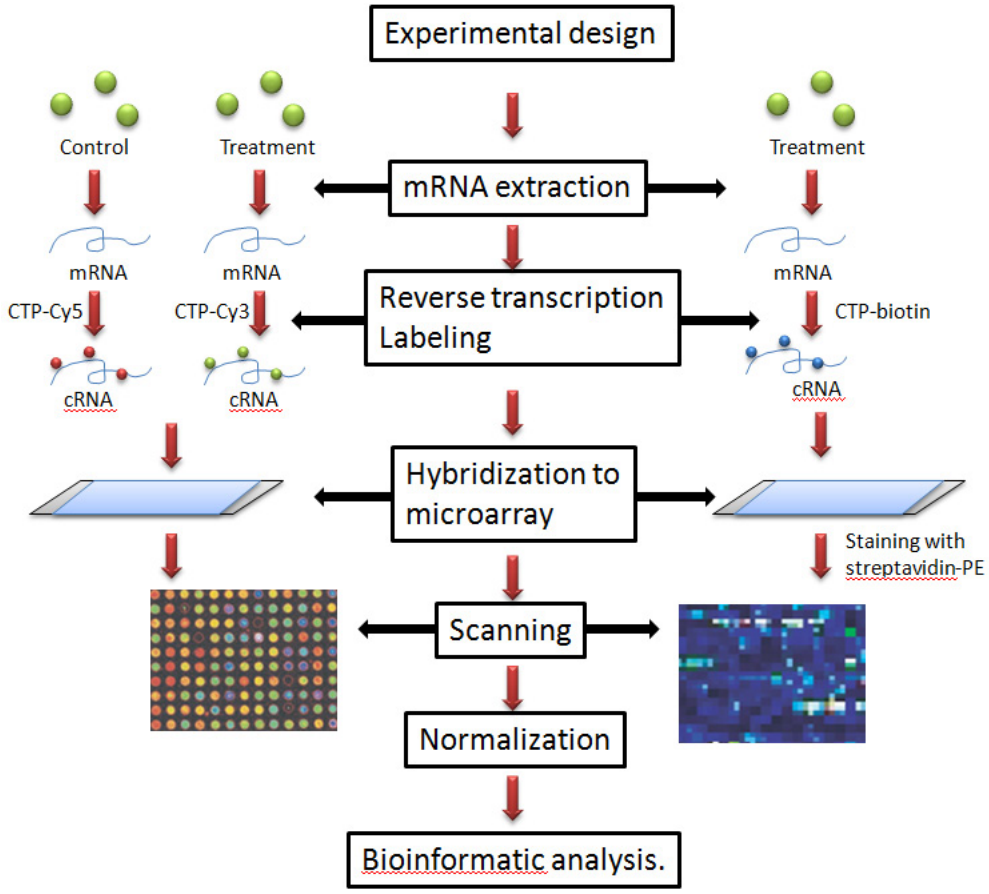

Figure 1.1 Comparison of two color and one color microarray procedures. *For two color arrays, two cell populations, for instance treatment and control, are isolated, RNA is extracted, and cDNA is made, which is used for in vitro transcription (IVT) with Cy3 (green) or Cy5 (red) labeled nucleotides. The two labeled cRNA samples are mixed and hybridized on a glass slide array, which is scanned with a laser, followed by computer analysis of the intensity image. With Affymetrix arrays, one population is used as starting material. Total RNA is extracted and cDNA is prepared. The cDNA is used in an IVT reaction to generate biotinylated cRNA. After fragmentation, this cRNA is hybridized to microarrays, washed and stained with phycoerythrin-conjugated streptavidin, and subsequently scanned on a laser scanner. Figure based on (Staal et al., 2003)*

Extracted mRNA is reverse transcribed into cRNA, and labeled with a fluorescent dye (Cy3/C5 for spotted, or long, oligonucleotide microarrays), or biotin (short oligonucleotide). The array is washed with the labelled cRNA (followed washing with phycoerythrin-conjugated streptavidin for short oligo microarrays), and under a fluorescent light, glow of each spot is quantified using a confocal microscope.

Microarrays are subject to multiple sources of undesired variation, which includes the array manufacturing process (e.g. imperfections during manufacture, dust spots on array), the preparation of the biological sample (e.g. different labeling methods used by different laboratories), the hybridization of the sample to the array (due to the existence of different hybridization protocols), and the quantification of the spot intensities (due to different protocols and types of laser scanners). Normalization is a critical initial step in the analysis of a microarray experiment, where the objective is to balance the individual signal intensity levels across the experimental factors, while maintaining the effect due to the treatment under investigation. The most commonly used normalization techniques assume that the majority of genes are not differentially regulated, or that the number of up-regulated genes roughly equals the number of down-regulated (Do et al., 2006). These assumptions seem to be adequate and do not appear to affect most biological experiments. The normalization strategy of *in situ* synthesized short oligonucletode arrays is different from that of long oligonucleotide or cDNA arrays, due to differences in array structure and labeling scheme. As short oligonucleotide microarrays are one colored, the normalization data is performed at the level of between-array, while normalization of two colored spotted oligonucleotide or cDNA array data is basically conducted at the level of within array. Do et al. (2006) provides a review discussing the different normalization procedures. While different normalization algorithms that can return dramatically different expression estimates, it has been showed that for co-expression analysis, the effect of a normalization procedure is negligible for a larger (>50) number of microarrays (Usadel et al., 2009). Short oligonucleotide microarrays from Affymetrix were used in this study, and the arrays were normalized using the popular RMA (Irizarry et al., 2003) and MAS5 algorithms (www.affymetrix.com).

1.2 Co-expression analysis

Functionally related genes tend to show coordinated spatiotemporal expression, i.e. co-expression (Figure 1.2; Yu et al., 2003). Co-expression analysis therefore employs "guilt-by-association" paradigm, where investigator assumes that the query gene is involved in the

same biological process as genes co-expressed with the query. While such assumption disregards the fact that coordination between mRNAs and corresponding proteins is surprisingly low, probably due to translational regulation (Maier et al., 2009), co-expression analysis has been successfully applied to functionally characterize previously unknown genes from yeast (Yu *et al.* 2003) and human (Lee *et al.* 2004).

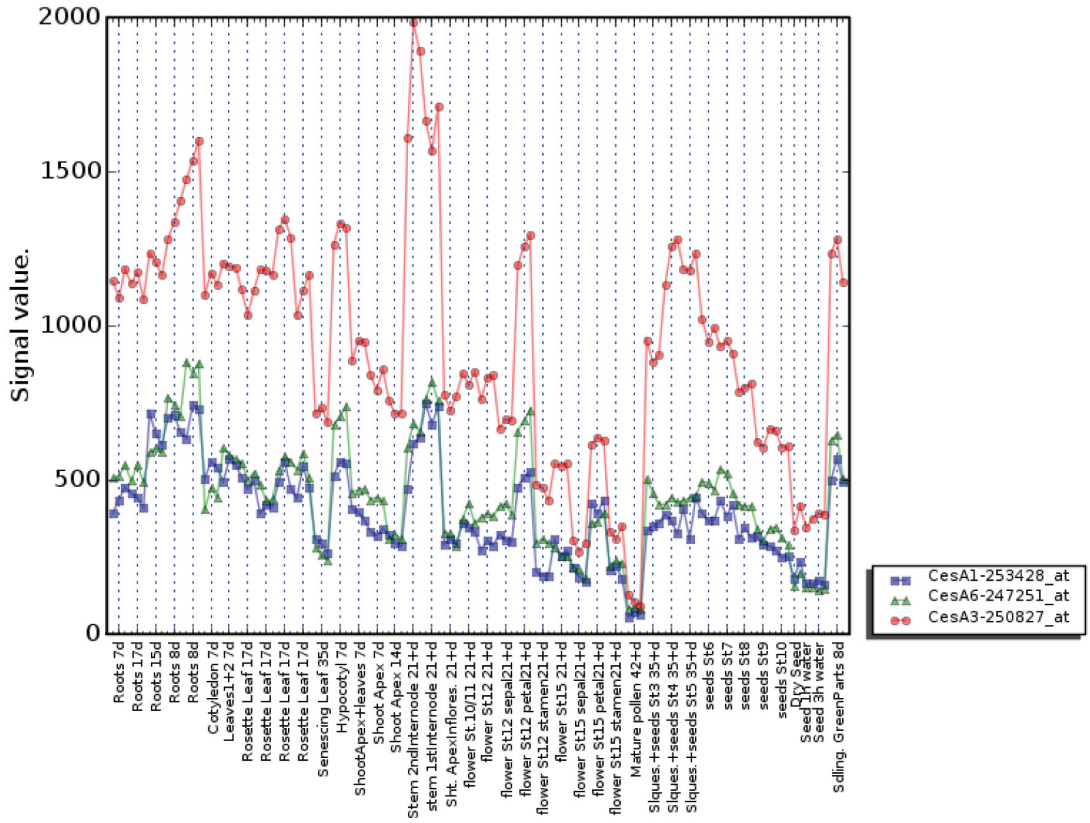

Figure 1.2 Expression values (y-axis) of Arabidopsis CESA1, 3, and 6 genes across microarray data from different tissues (x-axis). *The three CESA genes correlate with average Pearson correlation coefficient (r-value) of 0.87. Output from GeneCAT's ExpressionProfiling tool (Mutwil et al., 2008).*

Co-expression analysis in its simplest form can be used to: (1) find novel genes involved in the biological process of interest, and (2) suggest the biological process a gene is involved in. For example, (1) transcript of *Arabidopsis* gene *At1g31330* (photosystem I subunit F) is strongly co-expressed with other subunits of photosystem I and II complexes, and also with gene *At1g08380*, annotated as "predicted protein" (Table 1.1). While *At1g08380* gene is not characterized, it is likely to be also involved in photosynthesis. (2) Let's assume that the molecular function of *At1g08380* was defined in an experimental study, yet the biological

context of this gene is unknown. However, the gene is strongly associated with other genes involved in photosynthesis, implying the involvement of *At1g08380* in this process.

r-value	Gene ID	Description
1	at1g31330	photosystem I reaction center subunit III family protein
0.98206	at5g66570	Extrinsic subunit of photosystem II
0.9815	at4g12800	photosystem I reaction center subunit XI
0.97977	at1g55670	photosystem I reaction center subunit V
0.97963	at1g08380	expressed protein
0.97789	at4g02770	photosystem I reaction center subunit II
0.97714	at4g28750	photosystem I reaction center subunit IV
0.97547	at3g16140	photosystem I reaction center subunit VI
0.97473	at1g52230	photosystem I reaction center subunit VI, chloroplast, putative
0.97373	at3g61470	chlorophyll A-B binding protein (LHCA2)

Table 1.1 Co-expression analysis of photosystem I subunit At1g31330. *Top ten co-expressed genes are showed.*

In *Arabidopsis*, there are several instances in which co-expression analyses have successfully been used to identify genes not previously associated with a given biological process. Perhaps the most explored co-expressed process is the formation of secondary cell wall in *Arabidopsis* (Brown *et al.,* 2005; Persson *et al.,* 2005). Both studies identified candidate genes from co-expression analysis, and subsequently took a reverse genetics approach and showed through mutant analyses that several of the predicted genes were essential for secondary cell wall integrity.

Several groups have used co-expression approaches in attempts to associate genes with specific pathways or with given functions (Aoki et al., 2007), and with biological processes. For example, Ehlting et al., (2008) used co-expression analysis to ascribe potential function to members of the large cytochrome P450 superfamily in *Arabidopsis*. Horan *et al.* (2008) assigned 104 proteins of unknown function (PUFs) along with 269 proteins of known function (PKFs) as being involved in a wide variety of abiotic stresses, whereas a further 206 PUF genes, along with 608 PKFs, could be associated with specific stresses.

Co-expression relationships can be represented as tables (e.g. Table 1.1), or as graphs (introduced in chapter 1.4), in which nodes represent genes and connecting edges represent significant co-expression. The representation of co-expression relationships as networks

enable biologists to more readily contextualize their genes or proteins of interest, and are discussed and utilized in later chapters of this thesis.

1.2.1 Expression similarity metrics

Pearson Correlation Coefficient (PCC), or r-value, is the most commonly used metric to score expression similarity between any two genes. The PCC measures the tendency of the expression levels of a pair of genes to respond in the same (or opposite) direction across different samples. It ranges from -1, indicating that the two genes respond in completely opposite directions, to +1, in which case the two genes respond in the same manner across all samples (Figure 1.3). Thus, a positive correlation coefficient indicates that an increase in the expression level of one gene is likely reflected by an increase in the expression level of the other. In cases where the coefficient is zero, no association can be detected.

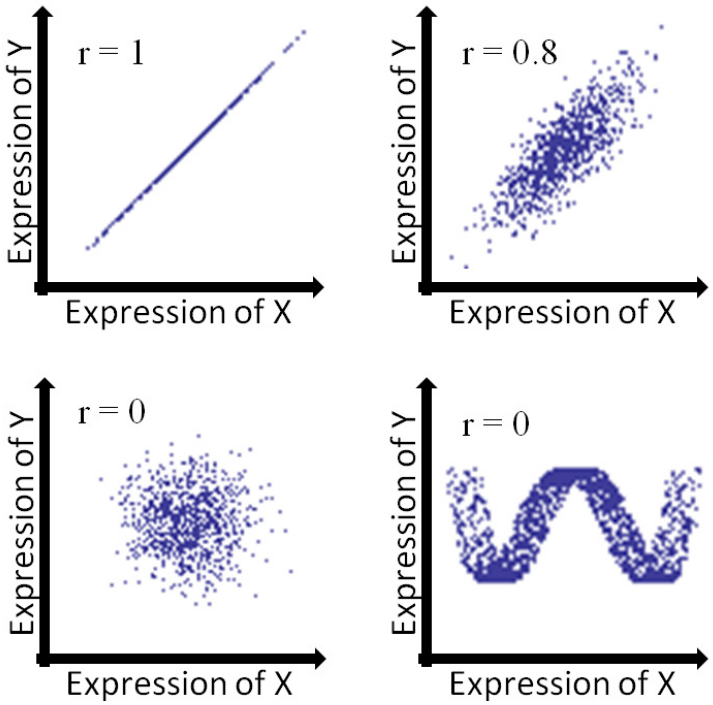

Figure 1.3. Exemplary plots of expression values of gene x vs. gene y. *Corresponding PCC values are given.*

Pearson correlation coefficient is obtained by the formula (1.1):

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \times \sqrt{\sum (y_i - \bar{y})^2}} \quad (1.1)$$

where x_i, y_i represent the expression values of gene x and y in array i, and $\bar{x}$ and $\bar{y}$ are the average expression of genes x and y across all arrays. A feature of Pearson correlation coefficient is its sensitivity for outliers. If both genes show very high expression, i.e. from an outlier in one sample because of a single 'bad' array, the Pearson correlation coefficient can approach +1, falsely indicating a strong correlation when none in fact exists.

Because of the Pearson correlation coefficient's sensitivity to outliers, other metrics more robust to outliers, such as *Spearman Correlation Coefficient* (SCC), can be applied. While SCC utilizes the same formula (1.1) for calculating the correlation coefficient, the expression values are first transformed into ranks, where for each gene the lowest expression level in the expression dataset gets a value of 1, the second lowest 2, and so on. This transformation diminishes the effect of an outlier as for large amount of arrays, the difference between an outlier and highest biologically relevant expression value is 1. However, in some cases such an 'outlier' might actually be driven by biology, and thus be highly meaningful. It may therefore be relevant to manually inspect the plotting of the values. Another interesting characteristic of the Pearson correlation coefficient is that it can only be +1 if the relationship in question is strictly linear, whereas the Spearman correlation can be used to detect non-linear associations if one gene is monotonically rising or falling in its expression levels with respect to the other.

Nevertheless, there are many kinds of relationships that cannot be uncovered with either correlation coefficient. For this reason the *mutual information*, derived from information theory, has been suggested to decipher relationships between genes (Steuer *et al.,* 2002). The mutual information quantifies the reduction in the uncertainty of one gene given knowledge about another gene. In the case that there is no interdependence it assumes the value of zero, but unlike correlation coefficients there is no upper bound for the mutual information score. Because the calculation of the mutual information more samples are needed for an estimate of the mutual information than for the estimation of correlation coefficients. In principle, the mutual information is suited to find any kind of relationship between genes. Complex non-linear types of relationships, however, do not seem to be present in all examples from real biological data (Daub *et al.* 2004), and are much harder to interpret biologically.

1.2.2 Setting threshold for biologically relevant co-expression

While abovementioned similarity metrics return a numerical value that reflects the strength of expression profile similarity, a more difficult task is to define a biologically relevant cut-off.

A commonly used method in statistics is calculation of p-values, which refers to the likelihood of obtaining the same or a better co-expression score than the observed by chance alone. The p-value can be calculated using formula (1.2):

$$ss = \frac{r \times \sqrt{n-2}}{\sqrt{1-r^2}} \quad (1.2)$$

where, r is the co-expression score and n is number of microarrays used in the study. However, due to a large number of genes on any array, p-value metrics often return very large amount of significantly correlated genes. Also, in cases where many arrays are used to compute the co-expression scores, a correlation coefficient as low as 0.2 can become highly significant. Therefore, especially when considering large number of samples, a significant correlation might not be of biological relevance. In addition, for Pearson's correlation analysis, the data have to be normally distributed for each gene and bivariate normally distributed for gene pairs, which might not hold true for all gene pairs. Also, the hypothesis of independence of experimental conditions cannot be essentially satisfied because the experiments were conducted for a particular biological purpose, resulting in deviations from the necessary behavior of the data (e.g. strong enrichment for microarrays representing one tissue type).

Rather than using p-value, existing co-expression approaches use subjective cut-offs. For example, r-value of 0.7 is often used as accepted cut-off (e.g. Srinivasasainagendra et al., 2008), as r^2= 0.49, which amounts to ~50% shared variance between expression profiles of two genes.

1.2.3 Meta-analysis of co-expression relationships.

Several studies have combined co-expression with other types of analyses.

Cis-*regulatory elements* in the promoter are the major contributions to the spatiotemporal regulation of gene expression. Several investigations observed that co-expressed genes share similar *cis*-regulatory sites (e.g. Wang et al., 2003), and studies combining co-expression analysis with promoter motif showed that this approach can effectively predict novel cis-elements. For example, Toufighi et al., (2005) showed that genes co-expressed due to abscinic acid treatment contain ACGT cis-element, which has been shown to mediate response to the hormone (Hobo et al., 1999). Consequently, bioinformatic

tools that can extract enriched regulatory sites from promoters of co-expressed genes have been developed for plants (e.g. Mariño-Ramírez, 2009, Toufighi et al., (2005)).

Comparative co-expression analyses between different species and kingdoms have revealed that co-expression relationships between corresponding gene pairs are robust (van Noort et al., 2003). For example, Stuart *et al.* (2003) combined co-expression analysis with gene sequence, and showed that certain modules of genes that are co-expressed in one species may be similarly co-expressed in other species, indicating that not only gene sequence, but also regulation can be conserved. Using this type of comparative analysis the authors predicted functionalities of different genes between species, and proved the predictions for a gene involved in cell proliferation in *C. elegans* using an RNAi approach (Stuart *et al.* 2003). A study by Geisler-Lee et al., (2007) combined co-expression analysis, *in silico* sub-cellular localization prediction, together with protein-protein interaction data from yeast and other species, to predict the *Arabidopsis* interactome. The study found total of 1,159 high confidence, 5,913 medium confidence, and 12,907 low confidence interactions in Arabidopsis proteins.

Ihmels *et al.* (2004) analysed how different genes that encode metabolic enzymes in yeast are transcriptionally wired. They found that most such connections formed linear arrangements following predicted pathway structures. Comparable analyses were subsequently undertaken in *Arabidopsis*. One such analysis investigated the transcriptional coordination of genes associated with secondary metabolism, and found that the genes encoding the main enzymes in the investigated pathways display a clear linear relationship to each other (Gachon *et al.* 2005). This analysis was followed by a wider analysis conducted on all genes associated with different metabolic pathways (Wei *et al.* 2006). Similar to Ihmels *et al.* (2004) observations, this study revealed that genes associated with distinct metabolic pathways were more tightly co-expressed than those in different pathways.

Co-expression analysis has also been combined with other omics techniques, such as metabolomics, to estimate coordination between gene expression and metabolite content, and to assess metabolite regulated gene circuits. Hirai *et al.* 2004 showed that both genes and metabolites associated with the glucosinolate pathway responded in an organized fashion, suggesting that this approach can reveal genes infuencing metabolic content of cell.

These and other examples demonstrate the importance of combining co-expression analyses with other large scale data, which may provide a system-wide glimpse into a cell. (Sweetlove & Fernie, 2005).

1.3 Functional gene ontology

Functional ontology analysis via Gene Ontology and Mapman Ontology are used in this thesis (chapter 3 & 4) to help infer the biological function of co-expression networks. A great advantage of ontologies is that they are precise and can be used to implement biological knowledge into algorithms.

Gene Ontology is a collection of terms used to describe gene products and was created to facilitate standardization and exchangeability of gene description (Ashburner et al., 2000). The Gene Ontology project provides an ontology of defined terms representing gene product properties. The ontology covers three domains: (i) cellular component (e.g. mitochondrium), (ii) molecular function (e.g. oxidoreductase activity) and (iii) biological process (e.g. oxidative phosphorylation). The GO vocabulary is designed to be species-neutral, and includes terms applicable to prokaryotes and eukaryotes, single and multicellular organisms.

Mapman ontology was designed specifically for plants (Usadel et al., 2006). Genes are assigned based on their annotation into largely non-redundant and hierarchically organised BINs. Each BIN consists of items of similar biological function and can be further split into sub-BINs, corresponding to submodes of the biological function.

1.4 Graphs and graphs theory

Graphs, or networks, are a conceptual construct that show the relationships in a system. A graph is often depicted as a series of nodes that are connected by lines (edges). The structure of a graph can be employed to define and analyse different properties that would not be visible when analyzing associations presented as a list.

Several studies have explored the properties of biological networks (Ihmels et al., 2004; Mentzen & Wurtele, 2008; Ma et al., 2007). The structture of biological networks may generally be described by power-law related relationships, i.e. a small number of nodes appear to have a large number of connections while most nodes have very few connections (Albert, 2005). Another apparent feature is that essentiality correlates with high node degree in both co-expression and protein-protein interaction networks in several species, i.e. essential genes have more edges that non-essential genes (Bergmann et al., 2004; Jeong et al., 2001; Carlson et al., 2006).

Properties of essential genes in co-expression networks are further discussed in chapter 3.3.11.

1.5 Clustering

Clustering is a process of grouping objects that are similar to each other, without considering any prior knowledge of their true membership. Clustering methods are widely used in co-expression analyses, as the process groups genes with similar expression patterns.

Genome-scale co-expression networks, probably due to the power-law nature, are highly heterogenous, with regions consisting of densely connected nodes, interspersed with regions of low density. The densely connected regions represent clusters of highly co-expressed genes, most likely involved in same biological process, and are therefore of great interest to a biologist. Clustering of microarray data is therefore a common procedure in transcriptomics.

The most popular algorithms in clustring of gene expression values are hierarchical clustering and k-means clustering (Eisen et al. 1998, Sherlock G., 2000). Several algorithms clustering networks, such as Markov Clustering and MCODE have been applied to protein-protein interaction networks (van Dongen, 2000, Bader and Hogue, 2003). Thalamuthu et al. (2006) provided a omprehensive review on clustering of expression data.
Clustering of co-expression networks, together with development of novel Heuristic Cluster Chiseling Algorithm (HCCA) are discusses in chapter 3.3.4.

1.6 Outline and contributions

In this thesis, new concepts of utilizing co-expression networks to help infer (i) gene function, (ii) organization of biological processes and (iii) knowledge transfer between species were introduced. An often overlooked fact by bioinformaticians is that a bioinformatic method is as useful as its accessibility. Therefore, majority of the work presented in this thesis was directed on developing freely available, user-friendly web-tools accessible for any biologist.

In Chapter 2, co-expression tool GeneCAT (Gene Co-expression Analysis Toolbox) for *Arabidopsis* and Barley is introduced. This platform provides the user both with standard co-expression tools, such as gene clustering and expression profiling, and also includes tools that use multiple bait-genes and makes functional inferences across different organisms by combining sequence comparison and co-expression analysis.

In Chapter 3, AraNet, web-tool for *Arabidopsis* based on co-expression network, is introduced. To better visualize the genome-wide co-expression network of *Arabidopsis*, a novel graph clustering algorithm was developed. Available phenotypic data for *Arabidopsis*,

together with ontological analysis of the network was combined to assign function to different network regions. To cluster the co-expression network, we have developed HCCA algorithm. We investigated the properties of the co-expression network in terms of gene essentiality, and characterized six novel genes, essential for Arabidopsis development. Importantly, we have introduced an analysis of transcriptional associations of Mapman ontology terms, which might reflect the coordination of biological processes in plants. We further explored the predictive power of this network through mutant analyses, and identified six new genes that are essential to plant growth.

In Chapter 4, we have extended AraNet with six additional model plant species, creating PlaNet. We implemented a comparative network algorithm that estimates similarities between network structures. Thus, the platform can be used to swiftly infer similar co-expressed network vicinities within and across species and can predict the identity of functional homologs.

The context of the results are discussed in the final chapter along with an outlook.

2. GeneCAT - Novel webtools that combine BLAST and co-expression analyses

2.1 Abstract

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

2.2 Introduction

The ability to measure the activity of several thousands of genes simultaneously has revolutionized the way we currently view biological processes. Substantial amounts of such expression data that represent experiments from a variety of tissues, developmental stages and stimuli, are currently publicly available for different organisms. Widely used public microarray data repositories are ArrayExpress (Parkinson et al., 2007) and Gene Expression Omnibus (GEO). As each microarray experiment often generates large amounts of expression data, it is often difficult for researches without background in bioinformatics to extract the information they seek. However, several web-based tools that analyze collections of publicly available microarray data for the plant model organism *Arabidopsis thaliana* have therefore been developed, including Genevestigator (Zimmermann et al., 2004), *Arabidopsis* Co-expression Tool (ACT; Manfield et al., 2006), Botany Array Resource (Toufighi et al.,

2005) , CSB.DB (Steinhauser et al., 2004) and ATTED-II (Obayashi et al., 2007). These tools provide comparative gene analyses including cis-element prediction, expression profiling and co-expression analysis. In addition, a tool that combines co-expression and predicted protein-protein interactions has recently been developed (Gesiler-Lee et al., 2007). It therefore appears that future web-tools will combine different types of data to facilitate a more complex and multi-dimensional view of organisms such as *Arabidopsis*.

Several studies exploit the fact that genes which are functionally related may be transcriptionally coordinated (Stuart et al., 2003; Bergmann et al., 2004). Recent studies have shown that this is also the case in plants (Brown et al., 2005; Persson et al., 2005; Wei et al., 2006; Hirai et al., 2007). Consequently, most of the current web-based tools are mainly focused on retrieving expression and/or co-expression patterns for individual genes. We have extended and refined this process and produced several new tools under the banner Gene Co-expression Analysis Toolbox (GeneCAT). This platform provides the user both with standard co-expression tools, such as gene clustering and expression profiling, and also includes tools that use multiple bait-genes and makes functional inferences across different organisms by combining BLAST and co-expression. GeneCAT is preloaded with data sets for two plant model organisms, *Arabidopsis* and barley, and data set from other species can readily be added. To increase the accessibility to the tools we have made GeneCAT accessible either via the web (www.genecat.mpg.de) or as platform independent source-code, upon request.

2.3 Results and Discussion

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

2.3.1 Expression Profiling and Tree View - Cellulose synthases

The Expression Profiling tool generates line plots of expression profiles for a specified set of genes within *Arabidopsis* and barley. The ExpressionTree tool uses these data to generate dendrograms corresponding to the tightness of co-expression for the same set of genes. To exemplify these tools we analyzed the cellulose synthase (*CESA*) genes from both

Arabidopsis and barley. There are 10 and at least 8 members of the CESA families in *Arabidopsis* and barley, respectively. The current model for cellulose synthesis proposes that at least three different CESA proteins are assembled into a functional complex (reviewed in Mutwil et al., 2008). Mutant analyses have shown that *AtCESA1*, *3* and *6* are necessary for primary cell wall cellulose synthesis in *Arabidopsis* (Arioli et al., 1998; Fagard et al., 2000; Persson et al., 2007). Similarly, *AtCESA4*, *7* and *8* are required for cellulose production during secondary cell wall formation (Turner and Somerville, 1997). A similar divergence of the *CESA* genes associated with primary and secondary cell wall synthesis is also predicted in barley (Burton et al. 2004).

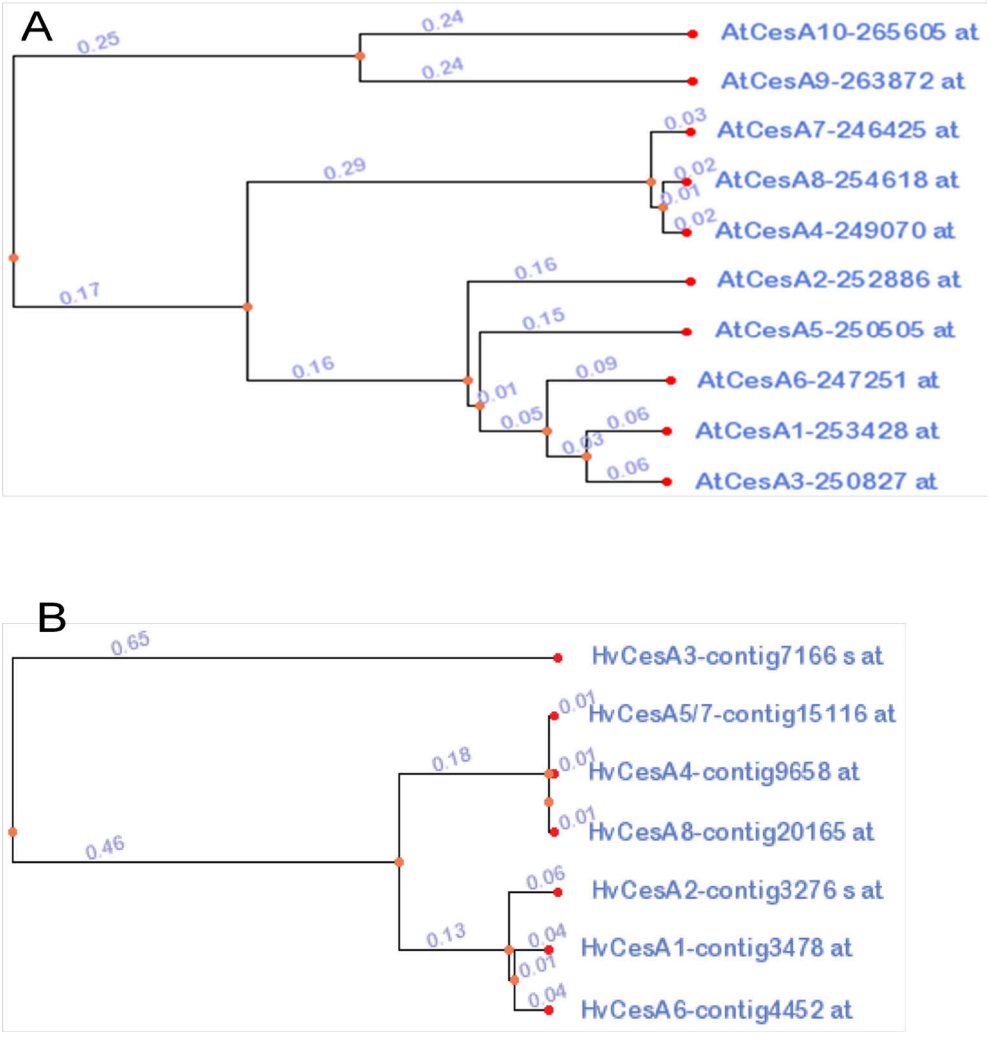

Figure 2.1. ExpressionTree analysis for cellulose synthase genes. *A. ExpressionTree analysis of AtCesA genes. B. ExpressionTree analysis of HvCesA genes. Numbers above edges signify branch lenghts, where lenght = 1 - PCC.*

The 10 *CESA* genes from *Arabidopsis* were analyzed using the ExpressionTree tool (Figure 2.1A). Two tight clusters were evident; one consisting of *AtCESA4*, *7* and *8* and the other including *AtCESA1*, *3* and *6* corresponding to secondary and primary cell wall biosynthesis, respectively. Interestingly, *AtCESA2 and AtCESA5* are tightly associated with the primary cell wall *AtCESAs*, and have recently been implicated to be functionally redundant to *At*CESA6 (Persson et al., 2007; Desprez et al., 2007). Similar to *Arabidopsis*, the expression of the eight Barley *HvCESA*s create two tight clusters consisting of *HvCESA1*, *2* and *6*, and *HvCESA5/7*, *4* and *8* (Figure 2.1B), suggesting that these groups of *HvCESA*s form functional complexes in barley. These data are consistent with results obtained by q-RT-PCR obtained by Burton et al (2004). The high sequence similarity of HvCESA5 and HvCESA7 makes it impossible to distinguish between these homologs (Burton et al 2004).

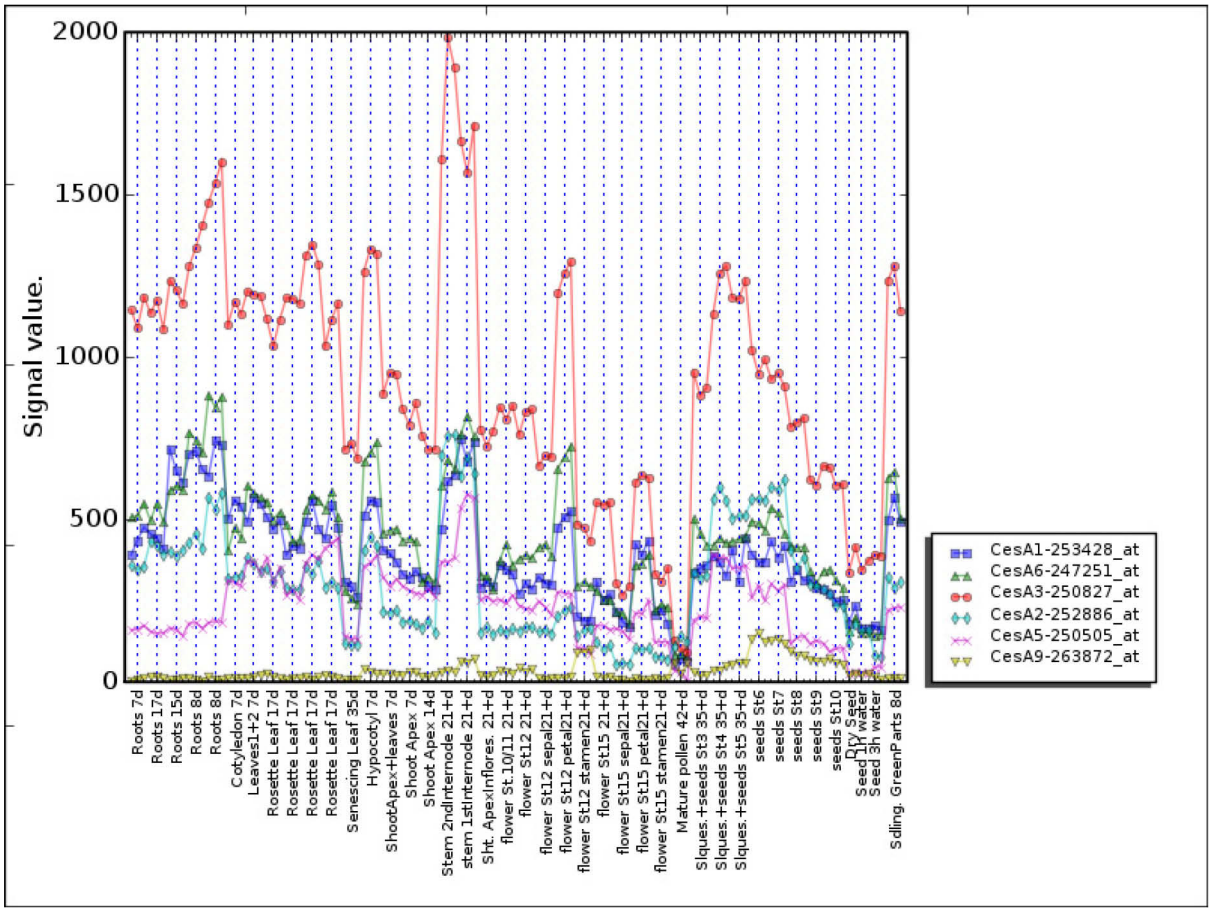

Figure 2.2 ExpressionProfiler output of primary cell wall associated CESAs.

We further examined the *AtCESA* expression levels using the Expression Profiling tool. Six of the *AtCESA*s, *AtCESA1*, *2*, *3*, *5*, *6* and *9* are anticipated to be involved in primary cellulose synthesis (Somerville, 2006). Recently, two studies provided evidence for functional redundancies for *AtCESA*2, 5, 6 and 9 (Persson et al., 2007; Desprez et al., 2007). To

evaluate the redundancies based on transcriptional patterns we plotted the expression levels for these *At*CESAs across different tissues (Figure 2.2). *AtCESA*2 and 6 are expressed throughout the plant. Consistent with this, double mutants between *cesa2* and *cesa6* result in additive phenotypic traits compared to the *cesa2* and *cesa6* parent lines (Persson et al., 2007). Interestingly, double mutants between *cesa5* and *cesa6* result in retardation of seedling growth (Desprez et al., 2007. In addition, *AtCESA9* is highly expressed in seeds and the male organs of flowers. In agreement with this triple mutants between *cesa2*, *cesa6* and *cesa9* result in male gametophytic lethality (Persson et al., 2007).

The use of these tools may similarly predict phenotypic outcomes and may therefore also provide a platform for researchers to choose mutant combinations.

2.3.2 Co-expression using multiple bait genes – Suberin biosynthesis

Genes that are involved in related processes are often co-expressed (Wei et al., 2006). Co-expression analyses therefore generally use a bait gene with a known function that is used to target transcriptionally coordinated genes. This approach typically returns a list of genes that appear co-expressed with the bait gene. It is, however, difficult to prioritize what genes that are most relevant to the process that the bait gene is involved in. It therefore appears that an enrichment of such genes would be highly appreciated by biologists. GeneCAT utilizes two approaches to enrich genes for a given function. First, two or more genes that are involved in functionally related processes may be used as bait genes to more accurately identify target genes. Second, target genes that are true positives should in general also exhibit significant transcriptional coordination to each other, thus forming clusters of co-expressed genes (Aoki et al. 2007). Several other tools provide the opportunity to apply such approaches, but GeneCAT is first to relate network information to the list of co-expressed genes. This process is done in three steps. In the first step an average co-expressed gene list is calculated for the bait genes. In the second step, a co-expressed gene network is created by measuring mutual co-expression ranks between the top 50 genes from the list in a pair-wise manner. Any two nodes (genes) that are connected with bold, normal or dashed lines display mutual ranks smaller than 10, 20 or 50, respectively. Blue nodes indicate bait genes and genes connected to these baits are colored green, orange and red if they are linked to any of the bait genes with bold, normal or dashed lines, respectively. The third step implements the color codes from

the network to the co-expressed gene list, thus highlighting genes that exhibit high transcriptional connectivity to the bait genes and other genes in the list.

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

To illustrate how the co-expression tool works we use a multi-gene co-expression approach for the suberin biosynthesis pathway from L-phenylalanine at AraCyc (http://www.*Arabidopsis*.org/biocyc/index.jsp) as an example. Suberin is a waxy, polymeric plant cell wall constituent that regulates water transport and protects against pathogen attacks (Franke and Schreiber, 2007). To enrich for other genes associated with suberin biosynthesis we then used these genes together with the *OMT1* gene as bait genes for the multiple-bait gene co-expression analysis (Table 2.1; Figure 2.3). Several genes that are connected to shikimate, phenylpropanoid, and chorismate biosynthesis are among the most highly ranked genes in the table. For example, two genes annotated as 4-coumarate-CoA ligases (At1g51680 and At3g21240) are among the top ranked genes (Table 2.1).

R value	Affymetrix probe	Locus	Gene Annotation
0.84853	263845_at	at2g37040	phenylalanine ammonia-lyase 1 (PAL1)
0.83857	253276_at	at4g34050	caffeoyl-CoA 3-O-methyltransferase, putative
0.79698	251984_at	at3g53260	phenylalanine ammonia-lyase 2 (PAL2)
0.76834	248200_at	at5g54160	quercetin 3-O-methyltransferase 1 (OMT1)
0.74408	260913_at	at1g02500	S-adenosylmethionine synthetase 1 (SAM1)
0.74355	267470_at	at2g30490	trans-cinnamate 4-monooxygenase / cinnamic acid 4-hydroxylase (C4H)
0.73926	256186_at	at1g51680	4-coumarate--CoA ligase 1 / 4-coumaroyl-CoA synthase 1 (4CL1)
0.72519	248639_at	at5g48930	transferase family protein, similar to anthranilate N-hydroxycinnamoyl/benzoyltransferase
0.68726	261933_at	at1g22410	2-dehydro-3-deoxyphosphoheptonate aldolase, putative
0.67677	258047_at	at3g21240	4-coumarate--CoA ligase 2 / 4-coumaroyl-CoA synthase 2 (4CL2)
0.67345	249910_at	at5g22630	prephenate dehydratase family protein
0.64855	262744_at	at1g28680	transferase family protein
0.64781	254192_at	at4g23850	long-chain-fatty-acid--CoA ligase
0.64220	260153_at	at1g52760	esterase/lipase/thioesterase family protein
0.63753	261749_at	at1g76180	dehydrin (ERD14)
0.63309	253277_at	at4g34230	cinnamyl-alcohol dehydrogenase, putative
0.61164	257771_at	at3g23000	CBL-interacting protein kinase 7 (CIPK7)
0.60580	263426_at	at2g31570	glutathione peroxidase, putative
0.60177	256964_at	at3g13520	arabinogalactan-protein (AGP12)
0.60059	263838_at	at2g36880	S-adenosylmethionine synthetase, putative
0.59751	250339_at	at5g11670	malate oxidoreductase, putative
0.59495	267212_at	at2g44060	late embryogenesis abundant family protein / LEA family protein
0.59441	246627_s_at	at2g45300	3-phosphoshikimate 1-carboxyvinyltransferase
		at1g48860	3-phosphoshikimate 1-carboxyvinyltransferase, putative
0.59430	245780_at	at1g45688	expressed protein
0.59213	262237_at	at1g48320	thioesterase family protein
0.59051	248393_at	at5g52060	BAG domain-containing protein
0.59030	247627_at	at5g60360	cysteine proteinase, putative / AALP protein (AALP)
0.58587	258852_at	at3g06300	Encodes a prolyl-4 hydroxylase
0.58493	255552_at	at4g01850	S-adenosylmethionine synthetase 2 (SAM2)
0.58458	264725_at	at1g22885	expressed protein
0.58370	263711_at	at2g20630	protein phosphatase 2C, putative / PP2C, putative
0.57996	256524_at	at1g66200	glutamine synthetase, putative, similar to glutamine synthetase
0.57864	254224_at	at4g23650	calcium-dependent protein kinase, putative / CDPK
0.57172	259516_at	at1g20450	dehydrin (ERD10)
0.57127	248573_at	at5g49720	endo-1,4-beta-glucanase KORRIGAN (KOR)
0.57113	262619_at	at1g06550	enoyl-CoA hydratase/isomerase family protein
0.56960	245666_at	at1g28280	VQ motif-containing protein
0.56939	248769_at	at5g47730	SEC14 cytosolic factor, putative
0.56763	260453_s_at	at1g72510	expressed protein
		at2g09970	expressed protein
0.56742	261438_at	at1g07590	pentatricopeptide (PPR) repeat-containing protein
		at1g07600	metallothionein-like protein 1A (MT-1A)
0.56634	253606_at	at4g30530	defense-related protein, putative
0.56549	258811_at	at3g03990	esterase/lipase/thioesterase family protein
0.55962	251962_at	at3g53420	plasma membrane intrinsic protein 2A (PIP2A)
0.55902	258037_at	at3g21230	4-coumarate--CoA ligase, putative
0.55664	261412_at	at1g07890	L-ascorbate peroxidase 1, cytosolic (APX1)
0.55555	250827_at	at5g05170	cellulose synthase, catalytic subunit (Ath-B)
0.55199	251739_at	at3g56170	Ca(2+)-dependent nuclease
0.55129	247817_at	at5g58375	expressed protein
0.55120	254262_at	at4g23470	hydroxyproline-rich glycoprotein family protein
0.54987	254103_at	at4g25030	expressed protein

Table 2.1. Co-expression analysis using multiple bait genes involved in suberin synthesis. *Five genes associated with suberin biosynthesis (blue color) were used as bait genes for the co-expression tool at GeneCAT.* Genes that are connected with display highest reciprocal ranks of 10, 20 and 50 to any of the bait genes, are color coded green, orange and red, respectively

These genes convert 4-coumarate into coumaryl-CoA linking the suberin biosynthesis and phenylpropanoid biosynthesis pathways. In addition, a gene annotated as prephenate dehydratase (At5g22630) is associated with the last steps in the phenylalanine biosynthesis pathway and may therefore provide substrate for the suberin biosynthesis.

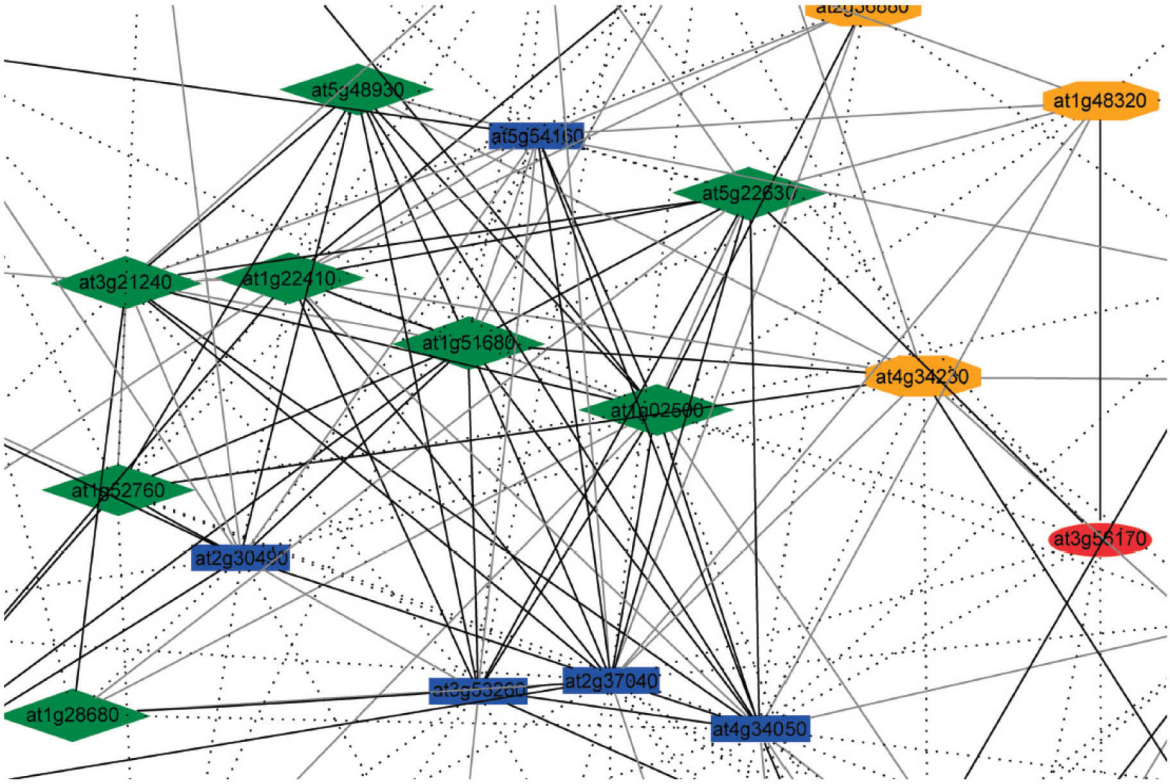

Figure 2.3. Co-expression network for multiple bait genes involved in Suberin biosynthesis*. Cropped co-expression network generated by using At2g37040, At4g34050, At3g53260, At5g54160 and At2g30490 as bait genes. Blue nodes indicate the bait genes for the analysis. Green, orange and red nodes indicate decreasing strength between node and the bait genes, respectively. Similarly, black, grey and dashed lines indicate decreasing strength between any two nodes.*

Furthermore, the cross-wise BLAST analysis of the top 150 genes identified several putative homologs associated with suberin biosynthesis (Table 2.2). These genes may consequently perform similar functions and may be considered as prime candidates for multiple mutant analyses.

By using several connected bait genes for a given process it is therefore apparent that functionally associated genes are enriched.

249910_at 252652_at 266257_at 261758_at	at5g22630 at3g44720 at2g27820 at1g08250	prephenate dehydratase family protein
251962_at 251324_at 265444_s_at 257313_at	at3g53420 at3g61430 at2g37180 at3g26520	tonoplast intrinsic protein, putative, similar to tonoplast intrinsic protein GI:5081419 from (Brassica napus)
247251_at 253428_at 250827_at	at5g64740 at4g32410 at5g05170	cellulose synthase, catalytic subunit, putative
258037_at 256186_at 258047_at	at3g21230 at1g51680 at3g21240	4-coumarate--CoA ligase, putative
245803_at 245483_at 247627_at	at1g47128 at4g16190 at5g60360	cysteine proteinase, putative
255552_at 260913_at 263838_at	at4g01850 at1g02500 at2g36880	S-adenosylmethionine synthetase, putative
259570_at 259516_at 261749_at	at1g20440 at1g20450 at1g76180	dehydrin
263845_at 251984_at	at2g37040 at3g53260	phenylalanine ammonia-lyase
245101_at 267470_at	at2g40890 at2g30490	cytochrome P450, putative
245356_at 257173_at	at4g13940 at3g23810	adenosylhomocysteinase, putative
258023_at 253277_at	at3g19450 at4g34230	cinnamyl-alcohol dehydrogenase, putative
252291_s_at 253099_s_at	at3g49120 at4g37530	peroxidase, putative
267153_at 267154_at	at2g30860 at2g30870	glutathione S-transferase, putative
260180_at 262990_at	at1g70660 at1g23260	ubiquitin-conjugating enzyme family protein
265354_at 247656_at	at2g16700 at5g59890	actin-depolymerizing factor

Table 2.2. Homologs identified by BLAST (BLAST cut-off e-value<10^-07) among genes co-expressed with multiple bait genes associated with suberin biosynthesis. *The table has been formatted from the output of the website. Annotations are abbreviated to fit the table format.*

2.3.4 Forward genetics predictions using Map-O-Matic: photosynthesis

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

To show how the Map-O-Matic tool works (Figure 2.4), we included an example based on photosynthesis.

A)

	Bait: gene X	Bait: gene Y
Genes in current database	R-value	R-value
Gene A	0.9882	0.0147
Gene B	0.9836	-0.0331
Gene C	0.6962	-0.0828
Gene D	0.6470	-0.1132
Gene E	0.0414	-0.2178
Gene F	-0.0128	-0.2731

B)

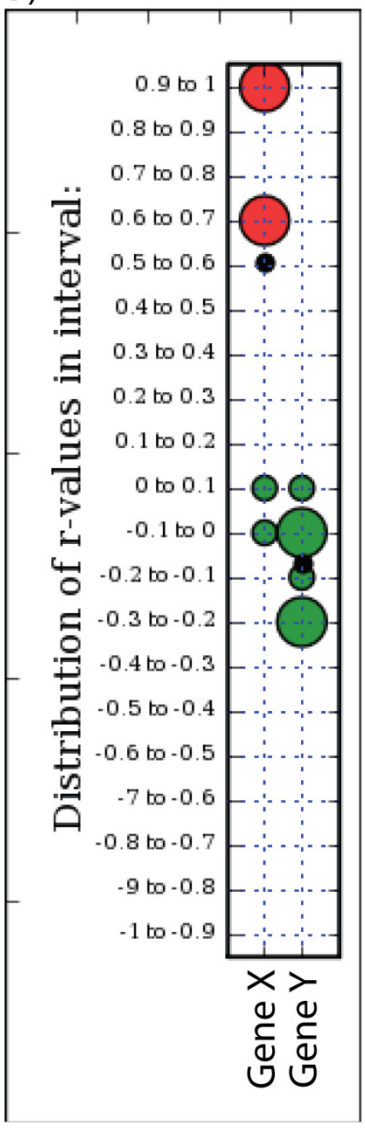

Figure 2.4. Calculation flow of the Map-o-Matic tool.*m The tool shows distribution of r-values of co-expression analysis between query genes and every gene in the current database. A. Map-o-Matic calculates co-expression analysis for two genes X and Y, using database genes A-F and produces two tables. B. The tool visualizes distribution of r-values in each co-expression list using circle sizes to depict the distribution. Genes in the picture are then ordered by the average r-value (black dot) in descending order.*

A mutant that is defective in photosynthesis was identified in *Arabidopsis* and the mutation was mapped to a genomic region of approximately 190 kbp (Muraoka *el al*. 2006). This

region is predicted to hold 57 genes, of which 50 were included on the ATH1 chip. To assess which of these genes that may be likely candidates we compiled a sub-database using the keyword "photosystem" as query, which found 47 genes associated with photosyntheis on the ATH1 chip. We then ran cross-wise co-expression analyses between the 50 candidate genes and the 47 photosystem associated genes. The output from the Map-O-Matic analysis is displayed as a graph with the average co-expression for each of the 50 candidate genes against the 47 photosystem associated genes (Figure 2.5).

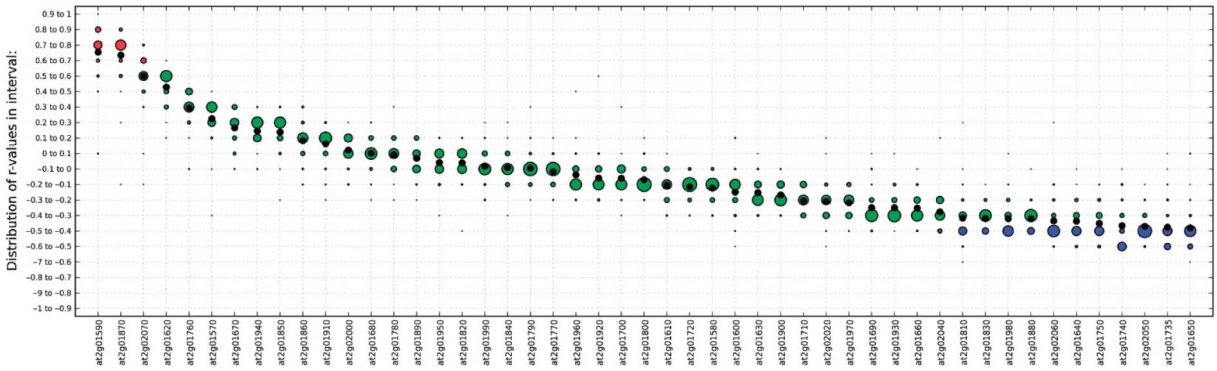

Figure 2.5. Map-O-Matic analysis of a photosynthesis mutant. *Fifty genes corresponding to a genomic region of approximately 190 kbp that was mapped for a photosynthetic defect was cross-wise compared for co-expression with 47 genes associated with the keyword photosystem. Each of the 50 bait genes on the graph is ranked by average coefficient of correlation across the comparison with the 47 photosystem genes. The bait genes are displayed in descending order from left to right, according to average correlation coefficient (depicted as a black dot).*

The top 5 genes of the 50 candidate genes are all highly co-expressed with most of the photosystem-associated genes (Figure 2.5). The gene that corresponded to the phenotypic trait was mapped to At2g01590 (Muraoka et al., 2006), which also was the gene that ranked as the most highly co-expressed gene with the photosystem genes of the 50 genes in the region. The gene ranked second in the analysis, At2g01870, is annotated as 'expressed protein'. Based on its high co-expression with the photosystem genes we suggest that this gene product may also play a direct role in photosynthetic processes. We believe that the Map-O-Matic tool is a powerful way to predict genes that are likely to be involved in specified biological processes.

2.3.5 Combining BLAST and Co-expression using Rosetta - Cellulose synthases

Orthologs in different species can be inferred through BLAST analyses and sequence comparison. These orthologs are then predicted to perform similar molecular functions in the different organisms. If they do perform similar functions we would also expect that other genes involved in the same process would have corresponding orthologs in the different species. Combining BLAST and co-expression analyses may consequently reveal "true" orthologous processes that are conserved in different organisms.

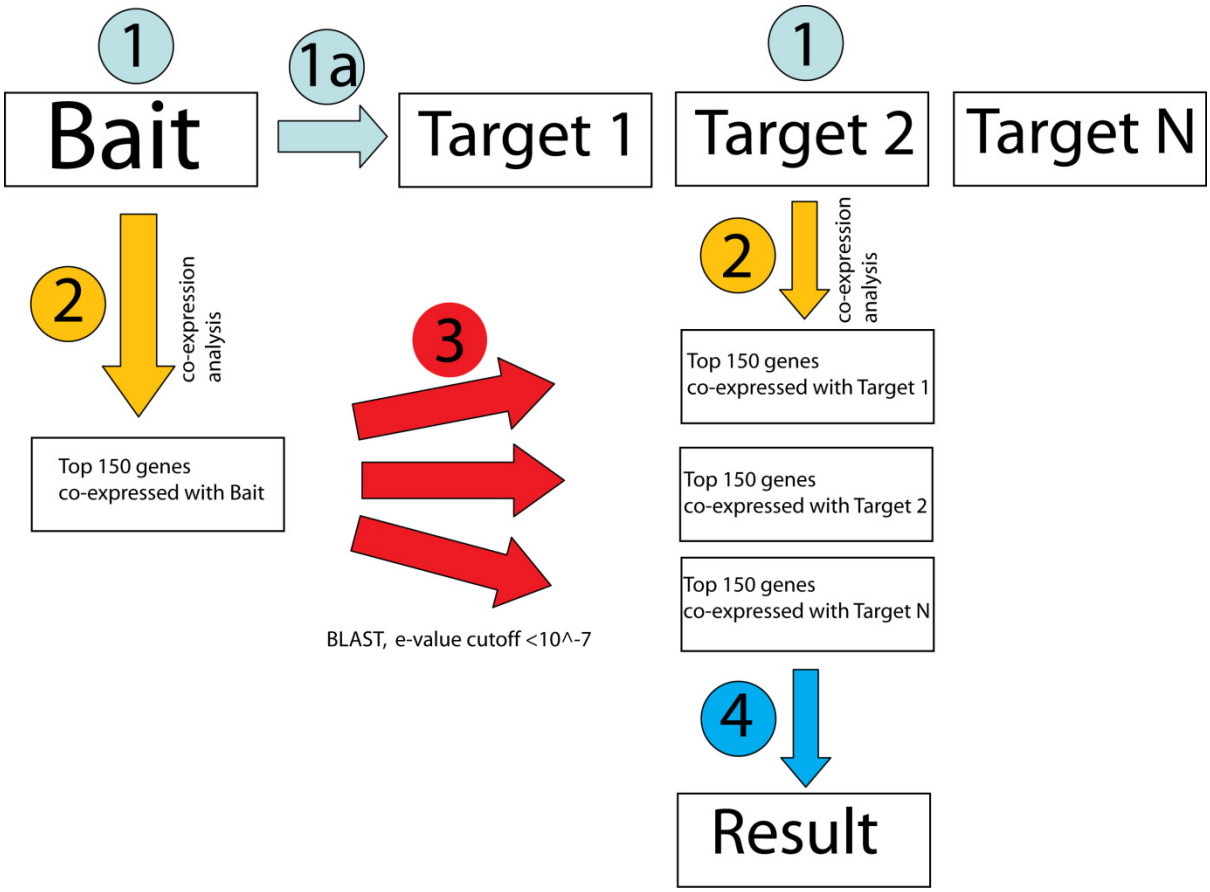

Figure 2.6. Calculation flow of the Rosetta tool. ***1.*** *User specifies the bait and target(s) genes.* ***1a.*** *Targets can also be found by using bait as a query in BLAST search (e-value<10^{-7}).* ***2.*** *Rosetta extracts list of top 150 co-expressed genes for bait gene, and lists the top 150 co-expressed genes for each of the target genes.* ***3.*** *BLAST analysis identifies genes in the target lists that share sequence similarity with genes in the bait list (e-value cut-off < 10^{-7}).* ***4.*** *The results are displayed as a table where genes with sequence similarity above cut-off are grouped in rows.*

Plant species typically contain comparatively large gene families (Cooke et al., 1997). This implies that several gene products may perform similar functions in different organs, tissues and/or developmental stages. It may therefore also be relevant to compare co-expression lists between these homologs to investigate functional conservation within a single species. To demonstrate the application of the Rosetta tool (Figure 2.6), we compared the cellulose synthesis machineries both in *Arabidopsis thaliana* and between *Arabidopsis thaliana* and Barley.

Primary cell wall *CESA1* and secondary cell wall *CESA4* in *Arabidopsis* were used as bait and target (i.e. *Arabidopsis* versus *Arabidopsis*), respectively, for the Rosetta analysis. Using this bait and target, Rosetta identified *AtCESA1*, *AtCESA3*, *AtCESA6* and *AtCESA4*, *AtCESA7*, *AtCESA8* as being associated with primary and secondary cell wall synthesis, respectively, based on the individual genes co-expression profiles (Table 2.3).

Bait: At4g32410 (CESA1)	Target 1: At5g44030 (CESA4)
At4g32410 cellulose synthase *AtCesA1* At5g64740 cellulose synthase *AtCesA6* At5g05170 cellulose synthase *AtCesA3* At5g09870 cellulose synthase *AtCesA5* At4g39350 cellulose synthase *AtCesA2*	At5g17420 cellulose synthase *AtCesA7* At5g44030 cellulose synthase *AtCesA4* At4g18780 cellulose synthase *AtCesA8*
At5g60920 phytochelatin synthetase (COBRA)	At5g15630 COBRA-like 4
At1g05850 chitinase-like protein 1 (CTL1)	At3g16920 CTL2
At5g49720 endo-1, 4-beta-glucanase (KOR)	At1g19940 glycosyl hydrolase family 9 protein
At3g23820 NAD-dependent epimerase/dehydratase	At2g28760 NAD-dependent epimerase/dehydratase At5g59290 UDP-glucuronic acid decarboxylase (UXS3)
At4g12880 plastocyanin-like domain-containing protein	At5g26330 plastocyanin-like domain-containing protein At3g27200 plastocyanin-like domain-containing protein At1g72230 plastocyanin-like domain-containing protein At1g22480 plastocyanin-like domain-containing protein
At5g03040 calmodulin-binding family protein	At2g33990 similar to calmodulin-binding protein At3g59690 calmodulin-binding family protein At3g15050 calmodulin-binding family protein
At3g16850″glycoside hydrolase family 28 protein	At3g42950 glycoside hydrolase family 28 protein At1g80170 polygalacturonase, putative
At1g75500 nodulin MtN21 family protein	At3g45870 integral membrane family protein/nodulin
At3g15480 expressed protein	At4g27435 expressed protein
At1g41830 multicopper oxidase type I family protein	At5g03260 laccase, putative At2g38080 laccase, putative At5g01190 similar to laccase At5g05390 laccase, putative At2g29130 laccase, putative At5g60020 laccase, putative
At3g02350 glycosyl transferase family 8 protein	At5g54690 glycosyl transferase family 8 protein At1g19300 glycosyl transferase family 8 protein
At5g12250 tubulin beta-6 chain (TUB6) At1g20010 tubulin beta-5 chain (TUB5)	At5g12250 tubulin beta-6 chain (TUB6) At5g23860 tubulin beta-8 chain (TUB8)

Table 2.3 Rosetta analysis comparing primary and secondary cellulose biosynthesis in* A. *thaliana. *Corresponding gene families from primary and secondary cell wall are presented in same row.*

Present in the co-expression lists were also genes that are common between the two processes. These include *COBRA* (At5g60920) and *CTL1* (At1g05850) and *COBRA*-like 4

(At5g15630) and*CTL2* (At3g16920) that are associated with primary and secondary cellulose synthesis, respectively (Table 2.3). The COBRA and CTL gene products affect primary and secondary cell wall biosynthesis, although their specific functions are unclear (Mutwil et al., 2008). Several other genes, such as glucanases, family 8 glycosyltransferases and arabinogalactan proteins, also appear to have homologs associated with primary and secondary cellulose production, respectively.

To identify genes associated with secondary cell wall biosynthesis in Barley we used *AtCESA4* from *Arabidopsis* as bait gene and used BLAST to identify targets in Barley (i.e. *Arabidopsis* versus Barley). Rosetta identified 14 probe sets in Barley that have similar sequences compared to *AtCESA4* in *Arabidopsis* (Table 2.4).

Target #	Affymetrix probe	Description	# hits with the bait / # maximum hits:
Target 1:	Contig3478_at	Cellulose synthase-9 related cluster	45/150
Target 2:	HK04P18r_s_at	Cellulose synthase-1 related cluster	39/150
Target 3:	Contig4452_at	Cellulose synthase-1 related cluster	47/150
Target 4:	Contig4451_s_at	Cellulose synthase-1 related cluster	49/150
Target 5:	Contig4451_at	Cellulose synthase-1 related cluster	47/150
Target 6:	**Contig9658_at**	**Cellulose synthase catalytic subunit 10 related cluster**	**57/150**
Target 7:	Contig7166_s_at	Cellulose synthase BoCesA4a related cluster	28/150
Target 8:	Contig5706_at	CSLF6 related cluster	46/150
Target 9:	Contig8067_at	Putative cellulose synthase-like protein OsCslE1 related cluster	25/150
Target 10:	**Contig20165_at**	**Cellulose synthase catalytic subunit 12 related cluster**	**55/150**
Target 11:	Contig20783_at	Cellulose synthase-like protein CslG related cluster	21/150
Target 12:	Contig7068_at	Cellulose synthase-like H1 related cluster	15/150
Target 13:	**Contig15116_at**	**putative cellulose synthase**	**56/150**
Target 14:	rbasd15h01_s_at	CSLD2 related cluster	38/150

Table 2.4. Barley probesets found to have significant (BLAST cut-off e-value<10^-07) sequence similarity to AtCesA4. *The three putative HvCESAs with high similarities to the AtCESA4 co-expression list are high-lighted in bold. The table is formatted from the output from the website.*

Similar to above, Rosetta recognized genes that are common between the co-expressed gene list for *AtCESA4* and the co-expressed gene lists for the 14 Barley probe sets. The comparison of the co-expression profiles revealed that three of the Barley probe sets, corresponding to Contig9658_at, Contig20165_at and Contig_15116_at, had the most similar co-expression profiles to *AtCESA4* in *Arabidopsis (Table 2.4)*. BLAST analysis revealed that these probe sets correspond to secondary cell wall *HvCESA4*, *HvCESA7* and *HvCESA5/7*, respectively. Similar to the analysis in *Arabidopsis*, Rosetta also identified putative *COBRA*-like 4 and *CTL2* orthologs associated with the secondary *HvCESA*s in Barley (Table 2.5).

Bait: At5g44030	Target1: Contig9658_at
At5g44030 cellulose synthase, *AtCesA4* At4g18780 cellulose synthase, *AtCesA8* At5g17420 cellulose synthase, *AtCesA7*	Contig9658_at Cellulose synthase, *HvCesA5//* Contig20165_at Cellulose synthase, *HvCesA*4 Contig15116_at Cellulose synthase, *HvCesA*8
At3g16920 chitinase-like, CTL1	Contig6213_x_at chitinase related Contig6213_s_at chitinase related
At2g38080 putative laccase At5g60020 laccase, putative	Contig18837_at Putative laccase related HVSMEn0005G15f_s_at Putative laccase LAC5-6 related
At5g03170 fasciclin-like arabinogalactan-protein, FLA 12	Contig10000_s_at Putative arabinogalactan protein related Contig10000_at Putative arabinogalactan protein related Contig15105_at Fasciclin-like protein related
At5g15630 Encodes a member of the COBRA family, CTL2	Contig23169_at BRITTLE CULM1 related
At3g62020 germin-like protein, GLP10	Contig10847_at Putative oxalate oxidase related
At2g37090 glycosyl transferase, IRX9	Contig13725_at 3-beta-glucuronosyltransferase related
At4g28500 no apical meristem (NAM) family protein	Contig11856_at No apical meristem (NAM) protein-like
At2g41610 expressed protein	Contig23037_at Expressed protein related
At3g50220 expressed protein At1g09610 expressed protein	Contig25082_at H0212B02.6 protein related
At5g01360 expressed protein	Contig5252_at Hypothetical protein Contig14278_at Expressed protein related Contig21755_at Hypothetical protein Contig13805_at Leaf senescence protein-like related
At5g60490 fasciclin-like arabinogalactan-protein	Contig10000_s_at Putative arabinogalactan protein Contig10000_at Putative arabinogalactan protein Contig15105_at Fasciclin-like protein

Table 2.5. Rosetta analysis comparing secondary cell wall AtCesA4 and HvCesA5/7. *Genes displaying mutual BLAST score e-value<10^-7 are placed in the same row by Rosetta. The table is formatted from the output from the website.*

Thus, Rosetta may rapidly identify homologs that are involved in similar biological processes within and across different organisms and may therefore be used to infer 'true' orthologs.

2.4 Concluding Remarks

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

2.5 Materials and Methods

Implementation and calculation.

GeneCAT is running on Apache server using cgi to link html forms with Python scripts. PhyFi (Fredslund, 2006) and Graphviz (www.graphviz.org) are used for visualization of ExpressionTree and co-expressed gene network, respectively. Calculations are performed on the fly by Python scripts. GeneCAT's source code is freely available upon request.

Microarray data sources and processing.

Databases for *Arabidopsis* and barley use Affymetrix ATH1 and barley1 GeneChips, respectively. *Arabidopsis thaliana* microarray datasets consisting of 1436 RMA normalized ATH1 microarrays data were obtained from TAIR (Rhee *et al*., 2003). Separate *Arabidopsis thaliana* tissue atlas datasets used for ExpressionProfiling were generated by the AtGenExpress project (Schmid *et al.* 2005). For the barley tissue atlas 64 MAS5 normalized microarray datasets were obtained from the BarleyBase (Shen *et al*. 2005) and was created by Druka *et al.* (2006).

3. Assembly of an Interactive Correlation Network for the *Arabidopsis* Genome Using a Novel Heuristic Clustering Algorithm

3.1 Abstract

A vital quest in biology is comprehensible visualization and interpretation of correlation relationships on a genome scale. Such relationships may be represented in the form of networks, which usually require disassembly into smaller manageable units, or clusters, to facilitate interpretation. Several graph clustering algorithms that may be used to visualize biological networks are available. However, only some of these support weighted edges, and none provide good control of cluster sizes, which is crucial for comprehensible visualization of large networks. We constructed an interactive co-expression network for the *Arabidopsis* genome using a novel Heuristic Cluster Chiseling Algorithm (HCCA) that supports weighted edges, and that may control average cluster sizes. Comparative clustering analyses demonstrated that the HCCA performed as well as, or better than, both the commonly used Markov, MCODE, and k-means clustering algorithms. We mapped MapMan ontology terms onto co-expressed node vicinities of the network, which revealed transcriptional organization of previously unrelated cellular processes. We further explored the predictive power of this network through mutant analyses, and identified six new genes that are essential to plant growth. We show that the HCCA partitioned network constitutes an ideal "cartographic" platform for visualization of correlation networks. This approach rapidly provides network partitions with relative uniform cluster sizes on a genome-scale level, and may thus be used for correlation network lay-outs also for other species.

3.2 Introduction

The complete, or partial, genome sequences from a vast number of organisms have increased our understanding of the design principles for biological systems (Kitano**,** 2002). The sequence availability has also provided platforms for various omics technologies, including transcriptomics, interactomics and proteomics (Schena et al.,1995; Li et al., 2004; Baerenfaller et al., 2008). Such techniques have generated an immense amount of data that for the most part is publicly available. One of the central ideas behind the concept of systems biology is to utilize these types of datasets to reveal functional relationships between genes, proteins, and other molecules (Kitano, 2002).

Transcriptional coordination, or co-expression, of genes may uncover groups of functionally related genes (DeRisi et al., 1997; Ihmels et al., 2004; Brown et al., 2005; Persson et al., 2005; Wei et al., 2006). Such relationships was initially utilized to reveal functional gene modules in yeast and mammals (Ihmels et al., 2004), and to explore orthologous gene functions between different species and kingdoms (Stuart et al., 2003; Bergmann et al., 2004). Comparable studies have also been undertaken in plants (Brown et al., 2005; Persson et al., 2005; Hirai et al., 2007). In addition, several web-based tools for plants offer various forms of co-expression analyses. These include CressExpress (Srinivasasainagendra et al., 2008), ATTED-II (Obayashi et al., 2009), *Arabidopsis* Coexpression Data Mining Tools (ACT; Manfield et al., 2006), Genevestigator (Zimmermann et al., 2004), GeneCAT (Mutwil et al., 2008), CSB.DB (Steinhauser et al., 2004), CoreCarb (Mutwil et al., 2009) and Expression Angler of the Bio-Array Resource (BAR; Toufighi et al., 2005). These tools can provide co-expressed gene lists for user specified query genes, and thus represent user-friendly web resources for biologists.

While it appears useful for scientists to examine these types of co-expression lists, more information is generally acquired by visualizing the relationships in the form of networks (Jupiter and VanBuren, 2008). Several studies have explored the properties of such network assemblies (Ihmels et al., 2004; Barabási and Oltvai, 2004; Mentzen and Wurtele, 2008; Ma et al., 2007). The distribution of connections in the networks may generally be described by power-law related relationships, i.e. a small number of nodes appear to have a large number of connections while most nodes have very few connections (Albert, 2005). Another apparent feature is that essentiality correlates with high-connectivity in both co-expression and protein-protein interaction networks in several species (Bergmann et al.,

2004; Jeong et al., 2001; Carlson et al., 2006), though this relationship is less clear in mammalian protein-protein interaction networks (Gandhi et al., 2006; Zotenko et al., 2008).

Although features of co-expression and protein-protein interaction networks have been investigated, the output is generally not very useful for visual inspection, and interpretation. One major task is therefore to make the networks more accessible to biologists, i.e. to produce visualizations of networks that easily may be interpreted (Aoki et al., 2007). For genome-scale networks this requires dividing the network into smaller manageable units, or clusters. Such clustering may, however, artificially assign genes to certain clusters, and therefore skew the output of the biologically "correct" network. It is therefore of importance to maintain as many relevant biological relationships as possible despite division. The ideal number, or sizes, of clusters to maintain these relationships are very rarely known, and is generally very difficult to predict for biological networks. On the other hand, biological networks may also be viewed as clusters within clusters, i.e. as a hierarchical structure that can be viewed on different levels. For example, genes associated with photosynthesis may be viewed as a cluster that belongs to a super-cluster of genes associated with functions in the chloroplast. Thus, the ideal clustering algorithm, and subsequent visualization scheme, should generate partitions of manageable sizes that readily can be re-connected into a whole network to be used for manual inspection.

Several graph clustering algorithms are available, for example Markov Clustering (MCL; van Dongen, 2000), Restricted Neighborhood Search Clustering (RNSC; et al., 2004) and others, but none of these can efficiently control cluster sizes. While these partitioning methods provide useful lay-outs for global biological and clustering interpretations, they are not particularly useful for visual inspection. To overcome this problem we developed a novel Heuristic Cluster Chiselling Algorithm (HCCA), and employed it to construct an interactive correlation network for the *Arabidopsis* genome (AraGenNet: http://aranet.mpimp-golm.mpg.de/aranet). We show that the HCCA-generated cluster solutions were as good, or better, than the commonly used partition algorithms Markov, MCODE and k-means using real world data. We also show that this type of visualization may reveal biological relationships that are not apparent from single gene co-expression approaches. Finally, we explored the network surroundings to identify essential *Arabidopsis* genes, and present six new genes that are essential for plant growth through mutant analyses.

3.3 Results and Discussion

3.3.1 Calculation of Pearson-Based Correlation Networks

To generate a starting network for the HCCA we calculated the degree of transcriptional coordination between all the genes present on the *Arabidopsis* ATH1 array (22,810 probe sets) using 351 RMA normalized microarray datasets from TAIR (Mutwil et al., 2008). Prior to choosing these datasets we removed datasets that displayed poor replication between arrays.

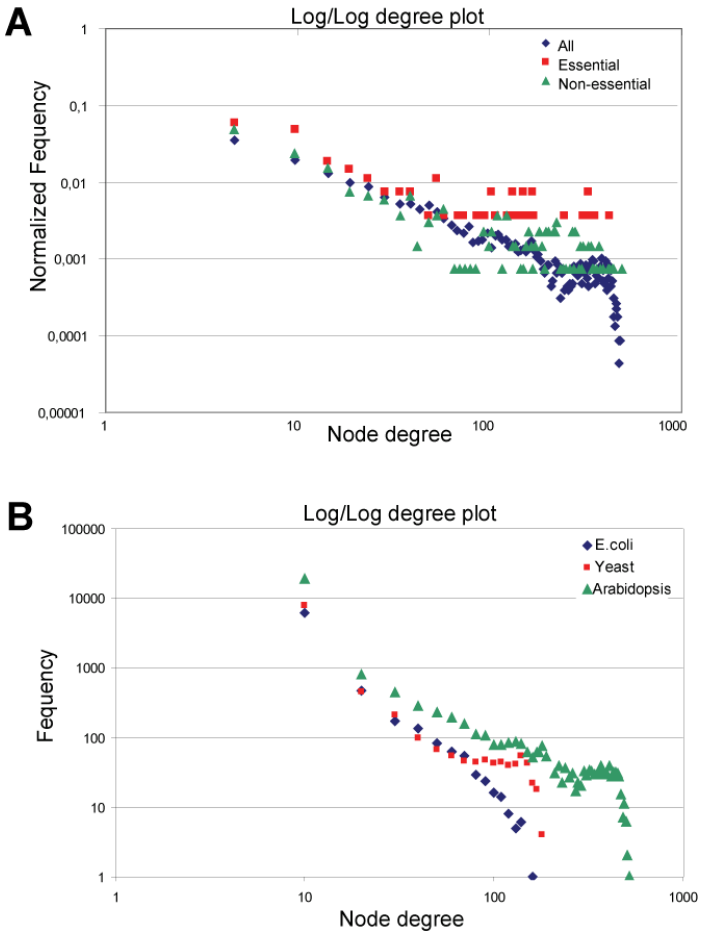

Figure 3.1. Network characteristics and mutant analyses. ***A.*** *Log-log plot of node degree distribution for 261 essential (red points), 1224 non-essential (green points), and all genes (22810 blue points) in the Pearson correlation network (r-value≥0.8) for Arabidopsis.* ***B.*** *Log-log plot of node degree distribution for Pearson correlation networks (r-value≥0.8) from E. coli (blue), yeast (red), and Arabidopsis (green). The x-axis represents the node degree, i.e. the number of connections a node shows, and the y-axis displays the frequency (**B**) or the normalized frequency (**A**), i.e. the number of genes (**B**), or normalized number of genes (**A**), showing this degree.*

Since it is rather difficult to assess whether lowly expressed genes represent noise or real data we chose to include all probe sets in the analysis. We then calculated all-versus-all co-expression network matrix using Pearson correlation coefficient cut-off of 0.8. In contrast to Spearman correlation, Pearson correlations only capture linear relationships between any two given components. However, it is anticipated that most linked expression profiles will adhere to a linear relationship (Daub et al., 2004).

The distribution of connections in Pearson correlation based biological networks may generally be described by power-law related relationships, i.e. a small number of nodes appear to have a large number of connections while most nodes have very few connections (Barabási and Oltvai, 2004). To assess whether the topology of the obtained Pearson correlation network for *Arabidopsis* also followed such a relationship, we calculated the node degree distribution of all individual nodes in the network. Figure 3.1 shows that the node degree distribution is best described by a truncated power-law behaviour. We also observed similar deviations from classical power law behaviour in Pearson correlation networks generated for Yeast, and to a lesser degree for E. coli (Figure 3.1B), in agreement with recent reports (van Noort et al., 2004).

3.3.2 Centrality vs. Essentiality

Another apparent feature in biological networks is that essentiality typically correlates positively with high node degree, i.e. mutations in highly connected nodes tend to result in more severe phenotypes compared to less well connected nodes (Albert, 2005; Jeong et al., 2001; Carlson et al., 2006; Zotenko et al., 2008). To assess if this type of relationship also is evident in our Pearson correlation network, we analyzed gene connectivity vs. embryo lethality. We did this by linking phenotypic data from The *Arabidopsis* Information Resource (TAIR; www.*Arabidopsis*.org) to the genes in our Pearson-based network (R=0.8). Figure 3.1A shows the node degree distribution of embryo lethal genes, genes associated with any type of phenotype, and all genes included on the ATH1 microarray. Whereas the node degree distribution for genes associated with non-lethal phenotypes did not deviate significantly compared to all genes present on the ATH1 gene chips (Figure 3.1A), genes corresponding to embryo lethality were significantly more connected compared to non-essential genes (Wilcoxon test $p<0.05$). Similar observations have also been reported for co-expression, and protein-protein interaction networks in yeast (Albert, 2005; Carlson et al., 2006).

3.3.3 Construction of a Highest Reciprocal Rank - Based Correlation Network in *Arabidopsis*

Several studies have used r-value cut-offs ranging between 0.6 and 0.8 to depict co-expression correlations (for example van Noort et al., 2004). However, different genes have different distributions of r-values, i.e. at a given cut-off some genes may correlate significantly with hundreds of genes while other genes may not correlate with any. Despite this, it is still possible that the latter may hold biologically relevant relationships. For example, the two transcription factors MYB33 (At5g06100) and MYB65 (At3g11440) regulate pollen and anther development, are expressed similarly, and are functionally redundant (Millar and Gubler, 2005). However, an r-value cut-off of 0.8 did not associate these genes transcriptionally (r-value 0.7; data not shown; Mutwil et al., 2008). To minimize this problem we chose to normalize the r-value distributions in the calculated Pearson correlation networks by using highest reciprocal rank (HRR) as they define the mutual co-expression relationship between two genes of interest. Using this approach the MYB33 and MYB65 were readily transcriptionally linked (average rank=2 using GeneCAT; Mutwil et al., 2008). With this approach we were also able to define a connection cut-off, or maximum number of connections, for a given gene. The importance of defining such cut-off is apparent when looking at the distribution of r-values among the data. For example, approximately 1500 genes are only expressed in pollen (estimated from GeneCAT; Mutwil et al., 2008). All of these genes are correlated with each other with an r-value of 0.8 and should therefore be connected to each other in a Pearson-based correlation network (Mentzen and Wurtele, 2008). However, it is virtually impossible to retain any information from such network structure through manual inspection. Instead we argue that displaying these genes in close network vicinities, which is achieved by the HRR-based network, is more useful. In addition, recent results indicate that correlation ranked networks produce sounder results than networks based on correlation co-efficients (Obayashi and Kinoshita, 2009).

We set the HRR limit to 30, thus capping the maximum number of edges per node to 30. The resulting HRR network seemed a reasonable compromise between readability and richness of information. In addition, we defined three degrees of co-expression weights using highest reciprocal ranks of 10, 20 and 30 (Mutwil et al., 2008). Similar approaches have also been used by several co-expression web-tools, such as GeneCAT and ATTED-II (Obayashi et al., 2009; Mutwil et al., 2008). The resulting weighted HRR network contained 103,587 edges between 20,785 nodes, and was used as the starting network for the HCCA. As

anticipated, not all the probe sets shared strong correlation with other probe sets, resulting in 2,025 nodes that were not included in the network (data not shown). The HRR based network shared 29,956 edges and 6,942 nodes with Pearson based co-expression network using $r \geq 0.8$ as cut-off (total: 231,882 edges, and 7,178 nodes).

3.3.4 Designing the HCCA

Genome-scale co-expression networks, as other networks, consist of nodes and edges that may form a continuous structure or separate islands of clusters, depending on what cut-off one uses. While the smaller structures in such network may be suitable for visual inspection, other regions may not due to the number of nodes and edges in these regions. To make such regions more accessible it is necessary to partition the network into smaller units, or clusters. Obviously, such partitioning will lead to division of network structures that may, or may not, reflect the ”real” network properties. Most biological networks do not contain sufficient data to assess whether the divisions are justifiable or not. However, the flaws in network divisions may be overcome if the different partitions can be reassembled into the structures they were initiated from. We argue that if we can visualize individual network partitions, or clusters, and put these into context to other clusters then the connectivity between the individual clusters may reflect the larger structures that were partitioned.

Many graph clustering algorithms do not support weighted edges, and do not yield cluster sizes that readily allow visual interpretations. In addition, many graph clustering algorithms do not allow clustering of large networks, i.e. networks consisting of several thousands of nodes. We therefore developed a novel graph clustering algorithm (Figure 3.2), referred to as Heuristic Cluster Chiselling Algorithm (HCCA). The HCCA algorithm takes *n*-value (step size), and desired cluster size range as parameters. The HCCA accepts a network as starting point (Figure 3.2). For each node in the network, the algorithm generates node vicinity networks (NVNs) by collecting all nodes within *n* steps away from the seed node. Nodes with higher connectivity to the outside of the NVN are iteratively removed. The resulting clusters are then ranked by outside-to-inside connectivity ratio, and filtered according to desired cluster size range. Non-overlapping clusters are retained by the algorithm and nodes in these clusters are removed from the network. Nodes associated with rejected clusters are returned to the network, and re-evaluated. The HCCA recursively creates non-overlapping clusters until no nodes are left in the network, or when no more stable

clusters can be obtained (Figure 3.2). In the latter case, remaining nodes are associated with clusters for which they display the highest connectivity to.

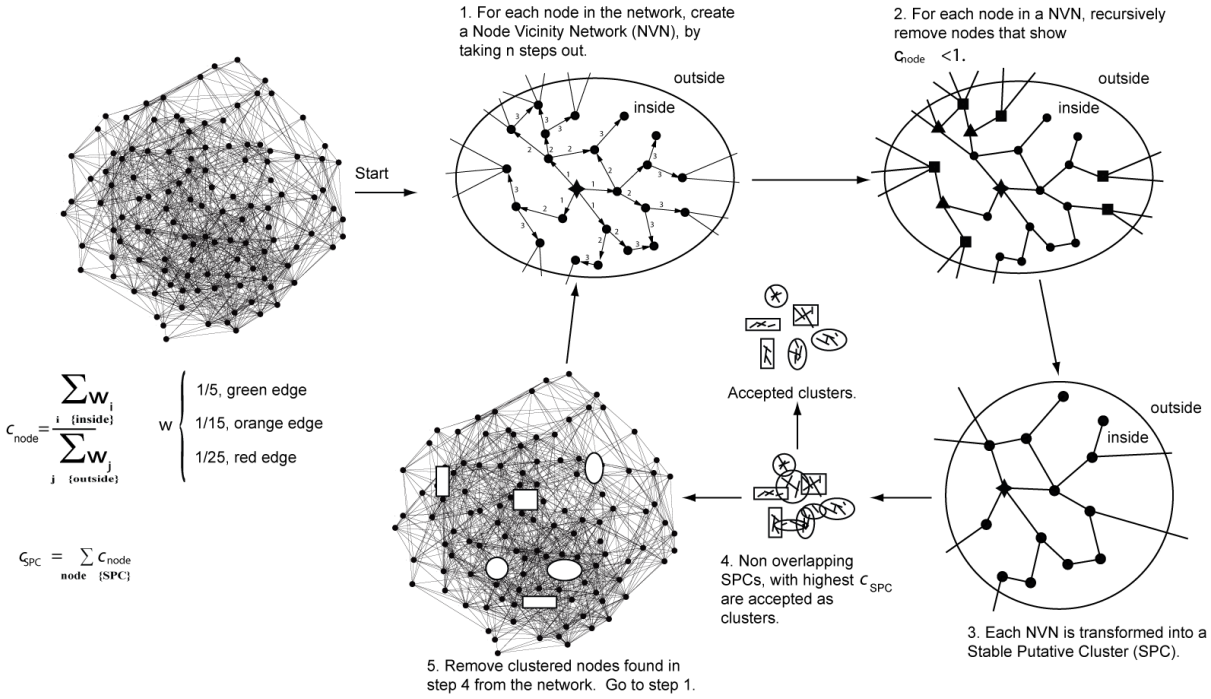

Figure 3.2. Schematic work-flow of the HCCA with n=3. *The HCCA accepts a network as input.* ***1.*** *Each of the m nodes in the network are used to generate Node Vicinity Networks (NVNs) by taking n steps away from a seed node (star).* ***2.*** *Each NVN is then "chiseled" by recursively removing nodes that have higher connectivity to nodes outside of an NVN than to nodes inside the NVN. In this example, squared and triangular nodes are removed in the first and second round of chiseling, respectively.* ***3.*** *The chiseling either completely depletes a NVN of nodes, or produces a Stable Putative Cluster (SPC).* ***4.*** *Non-overlapping SPCs with highest c_{SPC} value are extracted and accepted as clusters.* ***5.*** *Nodes that were accepted as clusters in step (**4**) are removed from the network. The remaining network is then transferred to step **1**, and re-chiseled (**2-5**).*

3.3.5 Visual Inspection of the Network Solutions

To partition the network we used the HCCA, with different steps *n* away from the seed node (Figure 3.2), with desired cluster sizes ranging from 40 to 400. For example, for *n*=3 the HCCA generated 181 clusters that contained approximately 40 to 300 genes per cluster (Figure 3.3A). To assess the biological relevance of the partitioned network we initially compared obtained connections with known biological data through visual inspection. For example, the secondary cell wall cellulose synthase genes *CESA4*, *7* and *8*, have been used

extensively for co-expression analyses (Brown et al., 2005; Persson et al., 2005; Ma et al., 2007). In agreement with these analyses we obtained genes associated with secondary cell wall synthesis, including *IRX6*, *IRX8*, *IRX9*, *IRX12*, and several transcription factors that recently have been implicated in secondary cell wall regulation (Zhong and Ye, 2007), in the network vicinity of the three *CESA* genes (data not shown).

3.3.6 Estimates of Clustering Solutions

A few other graph clustering algorithms also support weighted edge graphs, such as the commonly used MCL (Mentzen and Wurtele, 2008; van Dongen, 2000; Enright et al., 2002). To estimate the quality of the clustering solution obtained by HCCA we therefore clustered the HRR network using the MCL algorithm with a range of different inflation. We have also included MCODE clustering solutions (Bader and Hogue, 2003; Prieto et al., 2008). In addition, we performed clustering using k-means with different settings (Hartigan and Wong, 1979), and then compared the results obtained from the HCCA with the different clustering solutions for the other two algorithms (Figs. 3.3A to 3.3D). We used two different metrics to evaluate the clustering efficiency; the commonly used quantity *modularity* (Newman and Girvan, 2004), which judges partitions by comparing inside-to-outside connectivity ratios, and by the Davies-Bouldin index, which measures the compactness and separation of the obtained clusters (Davies and Bouldin, 1979). Our HCCA approach yielded better cluster partitioning compared to both MCL,k-means and MCODE in terms of *modularity* (Figure 3.3B). In addition, the HCCA solutions were clearly better than all the k-means partitions in terms of the Davies-Bouldin index (Figure 3.3C). However, the MCL and MCODE partitions rendered better Davies-Bouldin scores compared to the HCCA (Figure 3.3C). While the best overall MCL solution was the MCL 1.15 it is important to point out that this partition contain cluster sizes in the range of 2 to 800 genes per cluster (Figure 3.3A), and is therefore not useful for our purposes. These results show that the HCCA performed better than k-means in terms of modularity and Davies-Bouldin index, and also comparably to MCL and MCODE in terms of modularity scores.

When considering modular networks it is generally expected that neighboring nodes fulfill related functions, which also has been recognized in social networks (Wasserman and Faust, 1994). Hence, ideally one co-expressed gene cluster should contain genes associated with similar biological functions. We therefore also tested the overlap of MapMan ontology classes with the clusters generated by the HCCA, MCL, MCODE and k-means. We used an approach similar to ClusterJudge (Gibbons and Roth, 2002) which uses mutual information

between clusters and MapMan ontology terms to score clustering quality (Steuer et al., 2006).

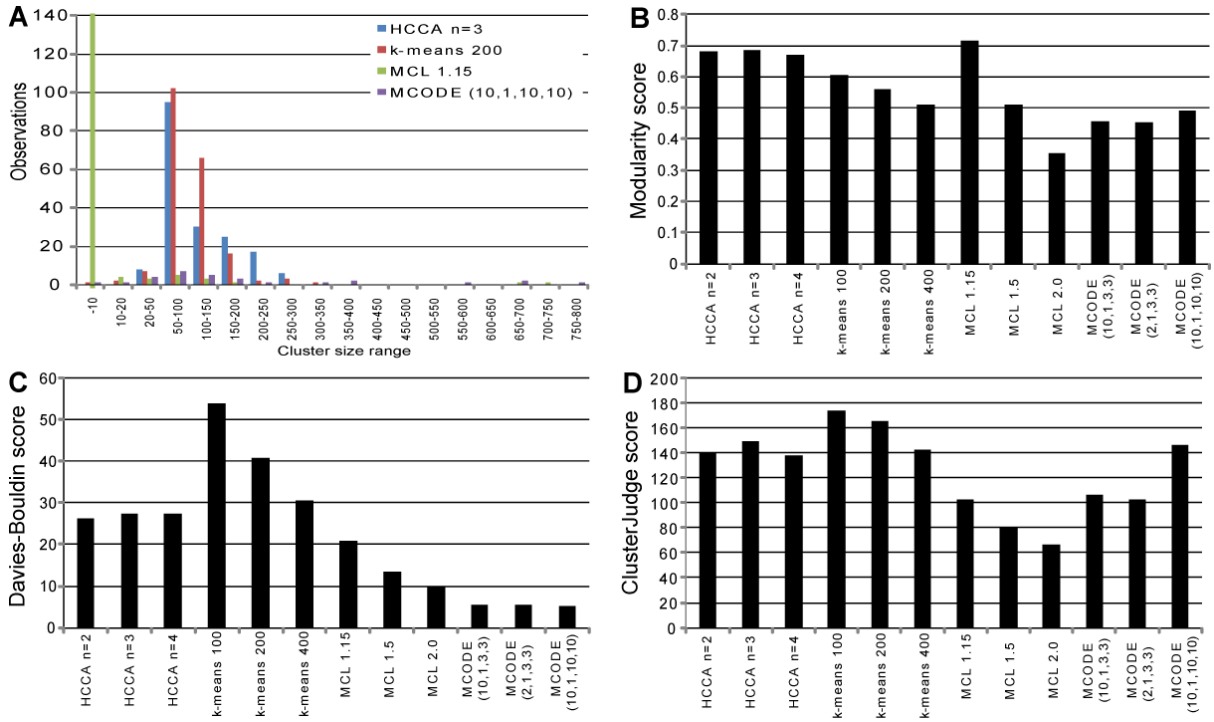

Figure 3.3. Cluster comparison of HCCA, MCL, k-means and MCODE. ***A.*** *Graph displaying the cluster size range (x-axis) vs. number of clusters (y-axis; Frequency) for selected HCCA, MCL k-means and MCODE partitions of the HRR network (HRR cutoff=30).* ***B.*** *Modularity scores for different settings for the HCCA, MCL k-means and MCODE algorithms. k-means 100, 200, and 400 represent desired cluster number parameters for k-means; MCL1.15, 1.5, and 2.0 represent different inflation degrees for the Markov clustering; HCCA n=2, 3, and 4 represent different step size (n) as described in Figure 3.2.; MCODE (A,B,C,D) represent degree cut-off, node score cut-off, k-core and maximum depth, respectively. High modularity values represent better clustering.* ***C.*** *Davies-Bouldin score, or index, for different settings for the HCCA, MCL, k-means and MCODE. The settings are in accordance with* ***B****. Low DB-score represents good clustering* ***D.*** *ClusterJudge scores of the clustering generated by HCCA, MCL, k-means and MCODE. respectively. The settings are in accordance with* ***B****. High ClusterJudge score represents better clustering.*

In brief, this approach scores the overlap between the ontological terms and the clusters, and then subtracts the mean score obtained for randomly assigned clusters and divides this by the standard deviation of the random clustering solutions. Therefore, a score of 0 (or even negative scores) would indicate random biological categories and clusters, whereas higher

scores (which have no upper bound) indicate better concordance between biological categories and clusters. Using this assessment the HCCA partitioned networks scored better than all of the MCL and MCODE partitions and scored nearly as well as the solutions generated by k-means (Figure 3.3D). It is important to note that the latter commonly used algorithm cannot generate clusters based on networks but must use the original expression data, and thus has an inherent advantage compared to the HCCA and MCL.

Taken together, these tests show that the HCCA partitions scored better than k-means in terms of modularity and Davies-Bouldin index, and outperformed the MCL and MCODE solutions in terms of biologically relevant associations.

3.3.7 Robustness of Clustering Towards Node Removal and to Different HRR Cut-offs

The ATH1 microarray chip contains 22,810 probe sets covering roughly 60% of the genes in the *Arabidopsis* genome. This means that approximately 8000 genes are omitted from the chip, and therefore from our analysis. To assess whether omission of such a number of genes may significantly skew the connections in the HRR network we randomly removed approximately 20% of the genes from our datasets and re-clustered the network using HCCA. We repeated this twenty times and then assessed whether the clusters were significantly different by estimating the average adjusted Rand index. Average score for HCCA (n=3) was 0.3818, with only 4 % standard deviation. This value is similar to the value obtained for the comparison of one thousand k-means clustering solution with 100 cluster centers. These data show that the network outline, and HCCA clustering is robust against removal of a significant portion of randomly selected genes, and therefore also should display biologically meaningful correlations despite the absence of some genes on the ATH1 chip. A matrix containing Rand index comparison of the algorithms with different parameteres is available online as supplementary material (Mutwil et al., 2010).

To test how different HRR cut-offs influence the clustering by HCCA, we calculated adjusted Rand indices between networks generated using HRR of 10, 20, 30, 40, and 50. Table 3.1 shows that the adjusted Rand index is relatively high (>0.4) for networks generated by similar HRR cut-offs (HRR20 vs HRR30, HRR30 vs HRR40, and HRR40 vs HRR50), despite that the networks differ dramatically in the number of edges (Table 3.1). Taken together, these results indicate that clusters obtained by HCCA are robust against the parameters used to generate the co-expression networks.

Mutual rank	HRR10	HRR20	HRR30	HRR40	HRR50
HRR10	1	0.3643	0.1595	0.0871	0.053
HRR20		1	0.4763	0.2844	0.1833
HRR30			1	0.4802	0.3257
HRR40				1	0.4407
HRR50					1

Table 3.1 Adjusted Rand index analysis of clustering solutions generated by HCCA using different HRR cutoffs. *Sizes of the networks compared: HRR10=26770 edges, HRR20= 63491 edges, HRR30= 103587 edges, HRR40 = 145644 edges and HRR50 = 189291 edges. The networks contain 22810 nodes each.*

3.3.9 Construction of an Interactive Correlation Network for the *Arabidopsis* Genome

To illustrate the usefulness of the network partition obtained from the HCCA we implemented an interactive co-expression network browser, which we named the *Arabidopsis* Gene Network (AraGenNet; http://aranet.mpimp-golm.mpg.de/aranet). Since the aim of the visualization scheme was to reassemble the partitioned HRR network for manual inspection, the network works on two levels; on assembled cluster level, and on gene level (Figs. 3.4 and 3.5).

The cluster level network (Figure 3.4) represents an overview of the interactions between different partitions, or clusters, and therefore depicts the co-expressed context for individual clusters. We therefore refer to this network as a meta-network. Any two clusters in the meta-network are connected if the combined weight of edges between them was larger than a certain threshold. We set this linkage threshold, or connectivity-score, to 0.02, as this value produced a connection-rich, but readable meta-network (Figs. 3.4A and 3.4B). A node in the meta-network consists of a cluster of co-expressed genes generated from the HCCA (n=3; Figure 3.5). This gene level network becomes visible by clicking on a cluster node in the meta-network. All connections in the gene level network are based on HRR, and are weighted accordingly, i.e. HRR below ten, twenty, and thirty are color coded green, orange, and red, respectively (Figure 3.5). These visualization schemes prove the capability and functionality of the HCCA clustering approach.

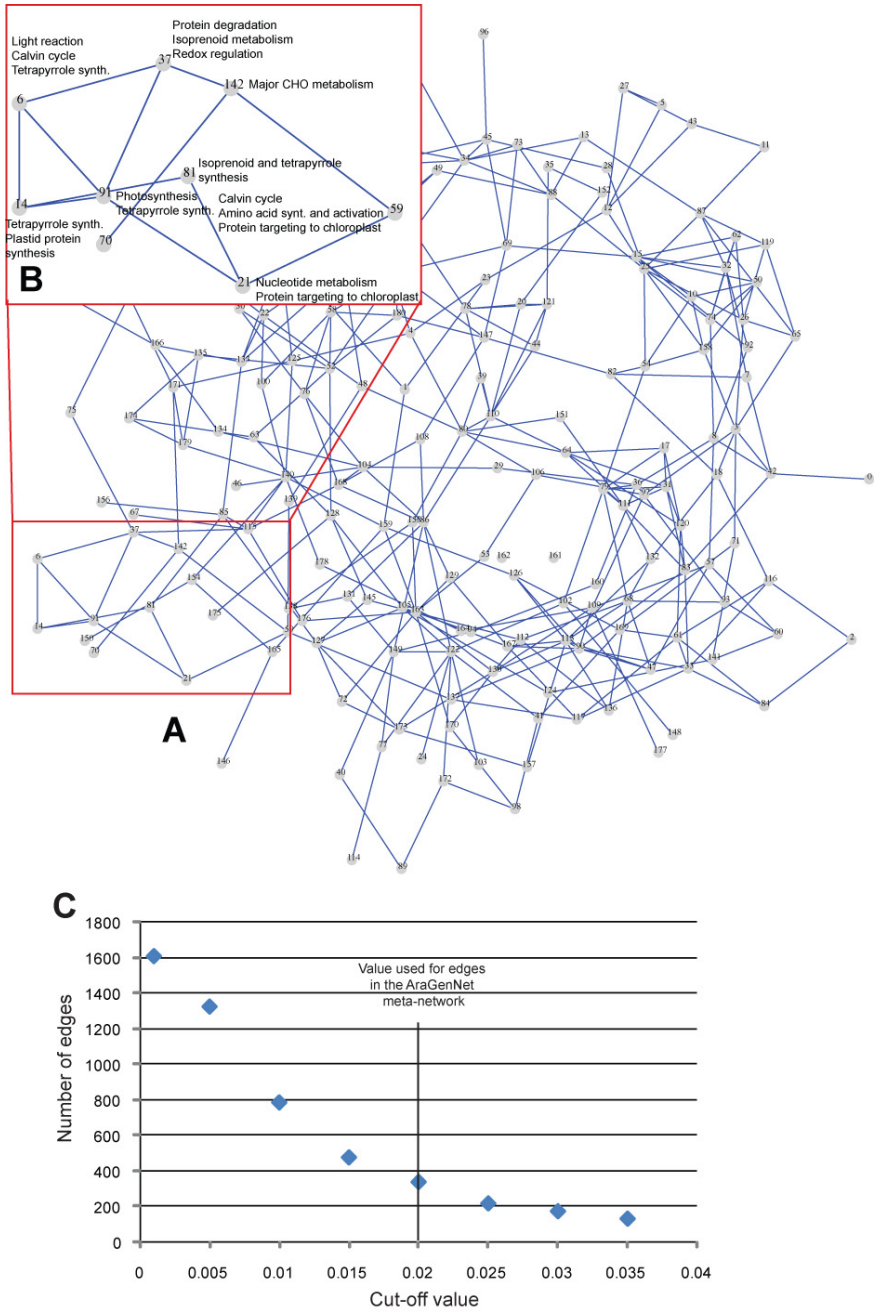

Figure 3.4 Meta-network of coexpressed gene clusters generated by HCCA (n = 3). *A, Nodes in the meta-network, or assembled cluster-level network, represent clusters generated by HCCA. Edges between any two nodes represent interconnectivity between the nodes above threshold 0.02 (according to C). B, Enlarged region depicts part of the meta-network presumably associated with photosynthesis. Cluster annotations were inferred by MapMan terms, phenotypic, and expression data (http://aranet.mpimpgolm. mpg.de/aranet). C, Connectivity cutoff values [c(A,B)] for edges in the meta-network. We used a cutoff of 0.02 for visualization purposes.*

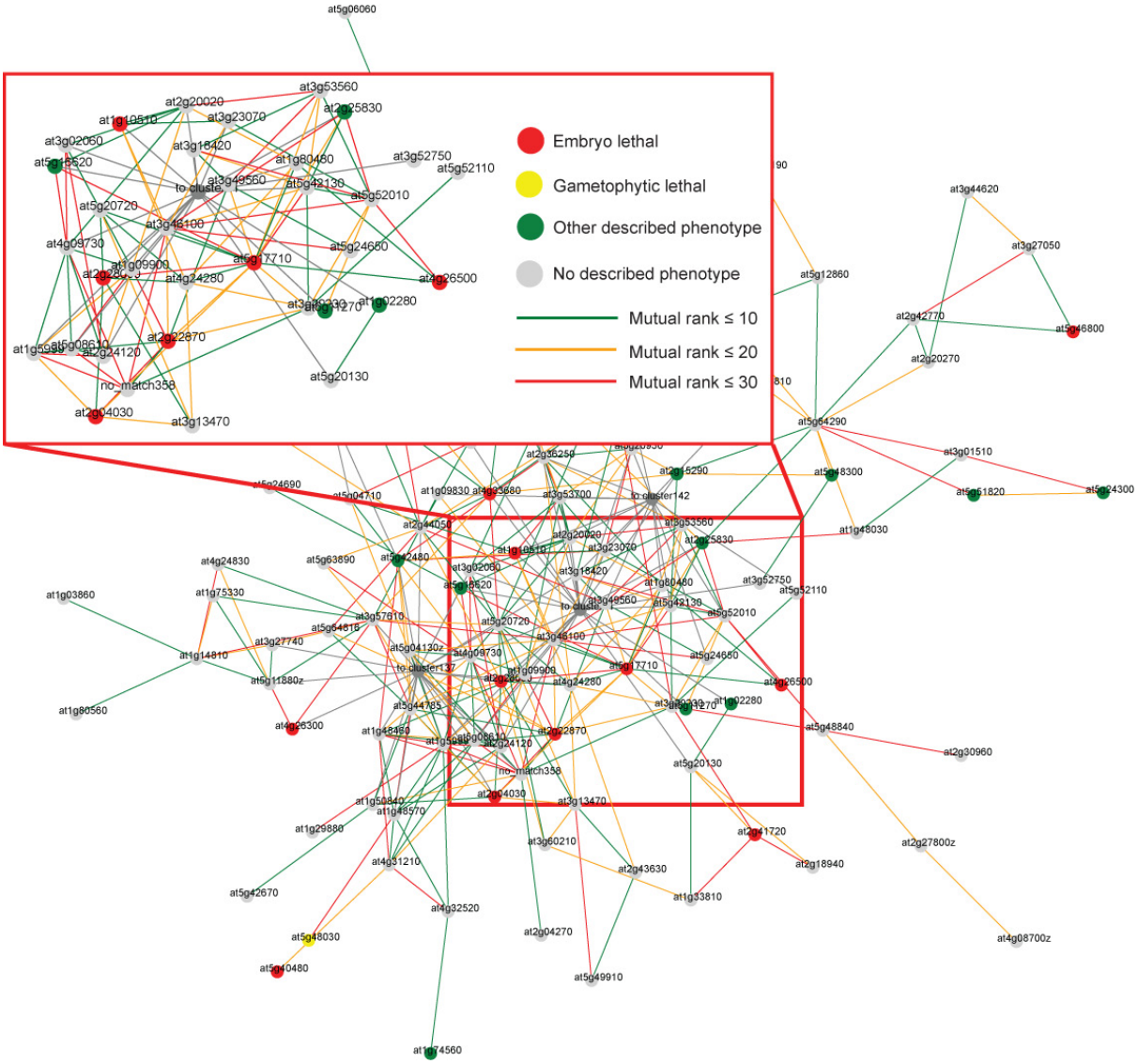

Figure 3.5 Features of HCCA (n = 3) gene cluster 59*. Nodes in this cluster, or gene-level network, represent genes, while edges and edge coloration depict the HRR values between any two nodes. Red, yellow, and green node colors depict gene mutants displaying embryo-lethal, gametophyte-lethal, and other described phenotypes, respectively. Gray nodes represent genes with no described phenotype.*

3.3.10 Phenotype and Ontology Mapping onto Network

Since co-expressed genes often tend to be functionally related (DeRisi et al., 1997; Ihmels et al., 2004; Brown et al., 2005; Persson et al., 2005; Wei et al., 2006) we anticipated that connected clusters in the meta-network would share a certain degree of functional commonalities (Freeman et al., 2007). To assess this we analyzed the genes in each cluster for MapMan ontology term enrichments. We also mapped phenotypic data (http://www.*Arabidopsis*.org/), and tissue-dependent expression profiling for the individual

genes. By combining these analyses we then attempted to describe what biological functions are associated with the individual clusters. For example, mutations in genes associated with cluster 59 (Figure 3.5) often result in embryo lethality, or pale green plants.

The dominant expression profile of genes in this cluster shows high expression in aerial tissues, and low expression in roots, pollen and seeds. MapMan ontology analysis revealed that the most significantly enriched term is amino acid metabolism ($p \leq 10^{-9}$). Taken together these data suggest that cluster 59 is over-represented for genes involved in amino acid metabolism in the chloroplast, and that this function is important for chloroplast development, photosynthesis and embryo development. This conclusion is supported by the fact that cluster 59 was highly enriched for genes with plastidic localization ($p<0.001$; data not shown).

3.3.11 Prediction and Verification of Essential Genes in the Network

To expand the visual features of the network we color coded the severity of the phenotypic traits using red (embryo lethality), yellow (gametophytic lethality), and green (other phenotypes) nodes in the network (Figure 3.5). Interestingly, we observed an uneven distribution of embryo lethal genes per cluster, compared with genes associated with non-lethal phenotypes (Figure 3.6A). For example, the chloroplast associated clusters 21, 59, 137 showed strong enrichment for essential genes ($p<10^{-5}$, Fisher's exact test). This suggests that nodes in clusters associated with certain biological processes are more essential. For example, of the 111 genes associated with cluster 59, twelve are known to be essential for embryo development (Figure 3.6A). As described above this cluster may be associated with amino acid activation in the chloroplast.

We also investigated how the essentiality of a gene is determined by the number and the distances of its homologs in the network. Figure 3.7A shows that embryo lethal genes are clearly over-represented by single-copy genes ($p<0.001$). Furthermore, essential genes tend to be under-represented for genes with family members in the network vicinity, i.e. in the node vicinity network ($p<0.05$; Figure 3.7 B-C). Conversely, non-essential genes tend to be neighbors to their family members ($p<0.05$; Figure 3.7 E-F). Taken together, the probability of essentiality for a given gene therefore appears to depend not only on the connectivity of the gene (Figure 3.1A), but also on its functional uniqueness in the network vicinity, and on its biological role. Interestingly, similar results have recently also been observed in protein-protein interaction studies in yeast (Zotenko et al., 2008). This study convincingly showed

that essentiality corresponded to gene products that are well connected, and that are associated with certain biological functions.

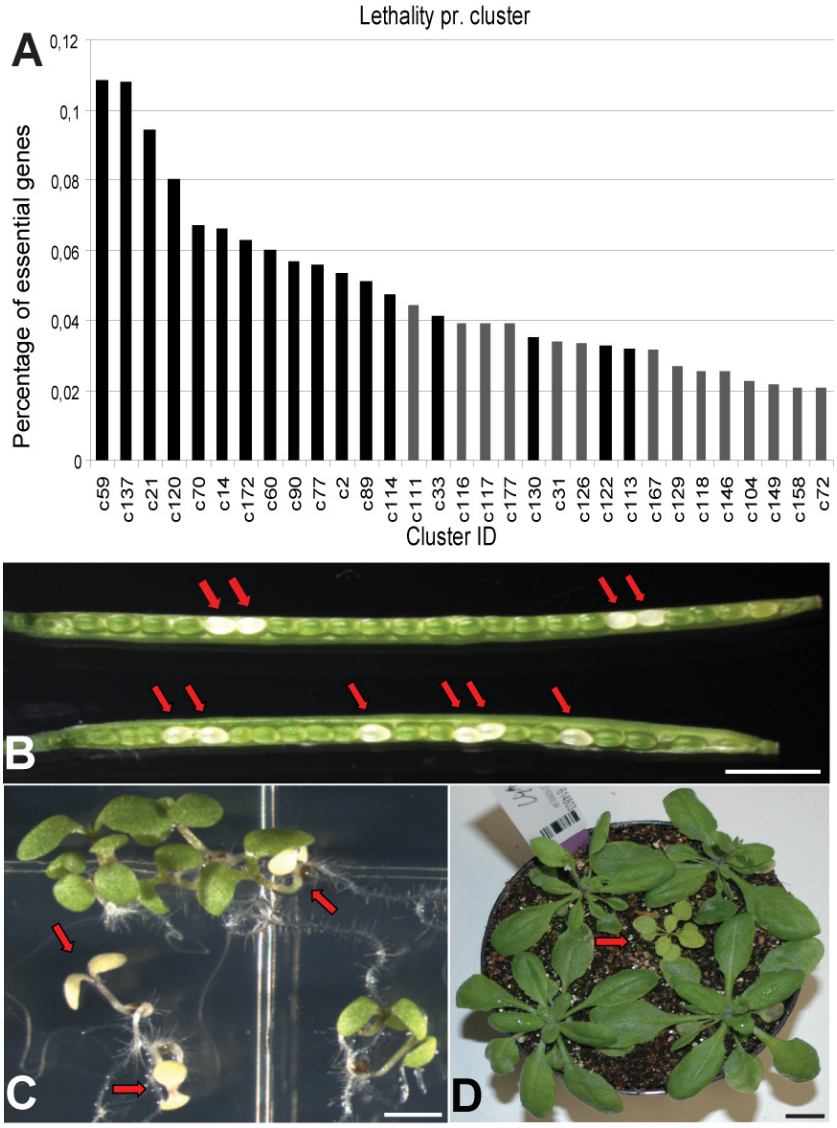

Figure 3.6. Essentiality distribution and mutant phenotypes in the HCCA (n = 3) partitioned network. *A, The graph displays the relative distribution of essential genes per any given cluster in the network (HRR cutoff = 30). Black bars depict clusters significantly enriched ($P < 0.05$) for essential genes. B, Siliques from a plant heterozygous for mutation in At3g14900 (cluster 137). Red arrows indicate chlorotic embryos. Bar = 3 mm. C, Mutant seedlings (At1g15510) from cluster 137 exhibiting pale cotyledons (indicated by arrows). Bar = 3 mm. D, Chlorotic dwarfed mutant (At3g57180; indicated by the arrow) from cluster 21. Bar = 1 cm.*

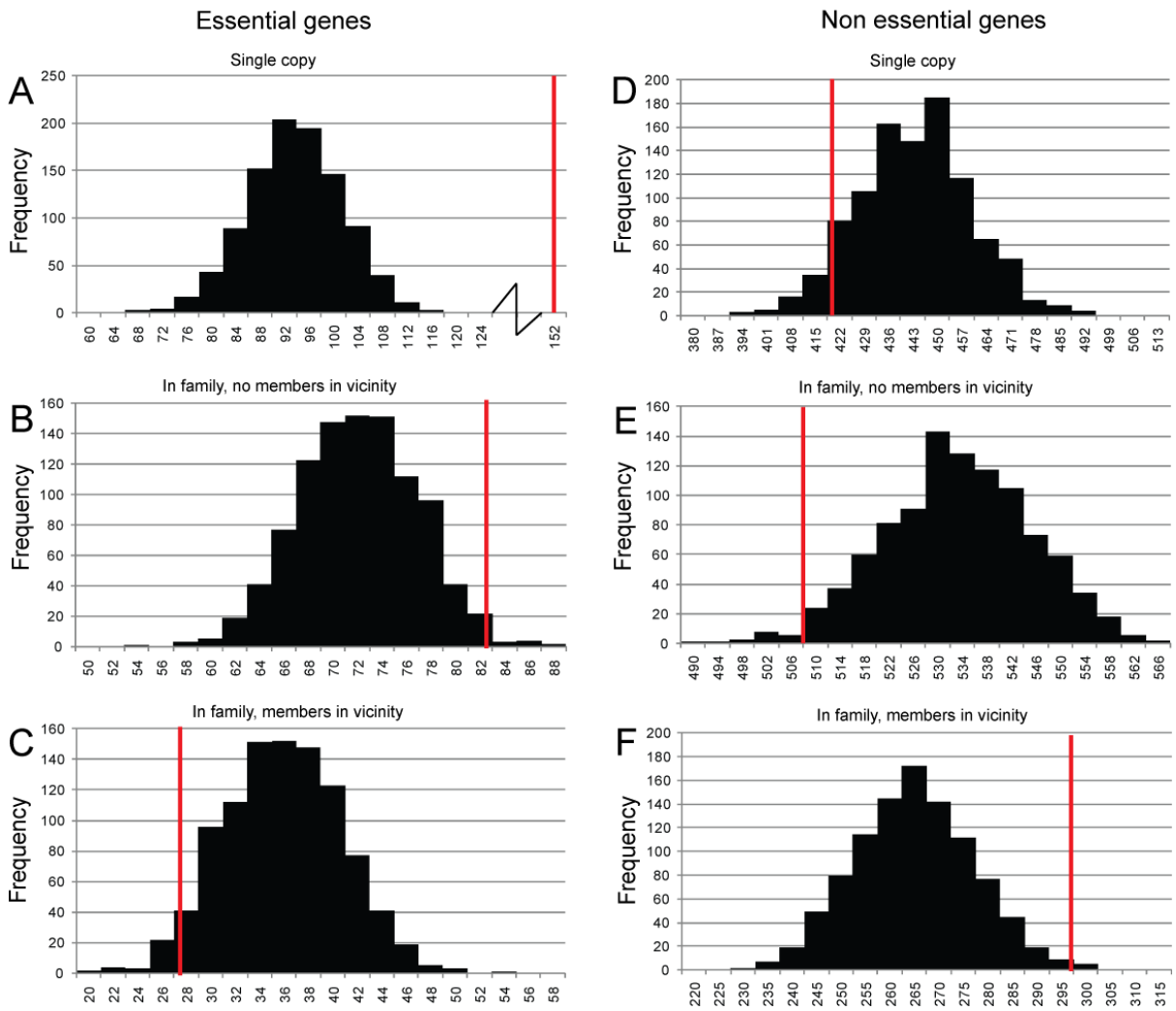

Figure 3.7. Distribution of 1000 random samplings of essential and non-essential genes from the mutual rank network. *A. Distribution of single copy genes from sampling of 261 random genes 1000 times. The number (152) of essential, single copy genes observed in our network is denoted by a red bar. B. Distribution of genes shown to be in a family but unique in the node vicinity network (n=2) from sampling 109 random nodes 1000 times. The observed number (82) of essential genes in family, but unique in the node vicinity network is denoted by red bar. C. Distribution of genes shown to be in a family with family members in node vicinity network (n=2) from sampling of 109 random nodes 1000 times. The observed number (27) of essential genes in family with family members in the node vicinity network is denoted by red bar. D, E, and F correspond to A (1224 nodes sampled), B (802 nodes sampled), and C (802 nodes sampled), respectively, but show distribution for non-essential genes. The observed numbers of non-essential, single copy (422), non-essential, in gene family, but unique in vicinity network (507), and non-essential with family members in vicinity network (295), are denoted by red bars in the figure.*

To explore the prediction of essentiality we chose twenty genes associated with clusters that harbor numerous essential genes, i.e. the connected clusters 21, 59, and 137 (Figure 3.6A), and that are well connected to other essential genes in the network. We ordered T-DNA mutant lines corresponding to these genes and analyzed them for mutant phenotypes (Table 3.2). Out of the twenty mutant lines two resulted in embryo lethality, one in seedling lethality, two in male gametophytic lethality, and one in dwarfed pale green plants (Figure 3.6C to E; Table 3.2). Furthermore, chlorotic cotyledon phenotypes are typically associated with chloroplastic functions (for example Flores-Pérez et al., 2008), supporting our prediction that genes belonging to these clusters, i.e. 21, 59 and 137, are functionally associated with the chloroplast. These results illustrate how a coherent and easy-to-navigate data visualization scheme, such as the AraGenNet, can predict biologically meaningful relationships. Recently, the pollen deficient mutant corresponding to the gene At1g74260 was confirmed by another study (Berthomé et al., 2008).

Gene	T-DNA line	Phenotype	Family size	Family members in vicinity
At3g23940	SALK_069706	Gametophytic lethal	0	0
At1g74260	SALK_050980	Gametophytic lethal	0	0
At5g64580	SAIL_74_G12	Embryo lethal	0	0
At3g14900	SALK_123989	Embryo lethal	0	0
At1g15510	SALK_112251	Seedling lethal	182	38
At3g57180	SALK_068713	Pale green, dwarf	0	0

Table 3.2. Characteristics of mutants. *The family size, and members in vicinity indicate size of a gene family as defined by COG, and number of family members in the gene network vicinity (n=2), respectively.*

3.3.12 Associations of Functional Annotations Using MapMan Ontology

Although the visualization of co-expressed genes may give insight into functional gene patterns and arrangements, an equally relevant quest is to understand how these patterns and arrangements are organized to fulfill cellular functions. To investigate this we explored the notion that co-expressed genes, and therefore network vicinities, often are functionally related (Brown et al., 2005; Persson et al., 2005; Wei et al., 2006; Ihmels et al., 2004). To assess how different ontological terms are transcriptionally connected we used the non-

clustered HRR network (HRR cut-off30), and calculated whether certain MapMan ontology terms were overrepresented in non-overlapping node vicinities (NVNs in Figure 3.2). We then identified terms that co-occurred more often than expected by chance ($p \leq 0.05$).

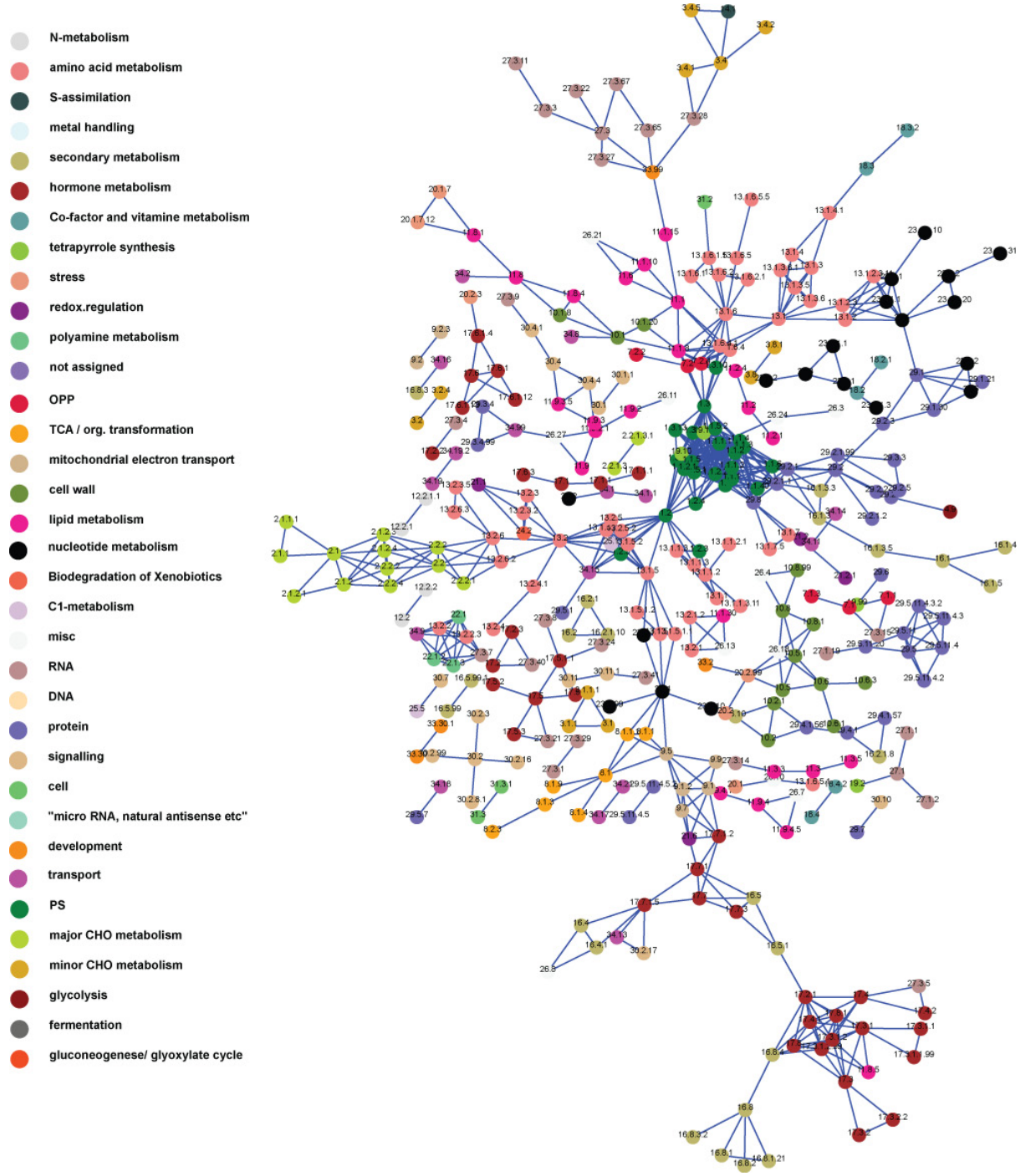

Figure 3.8. Network of coexpressed MapMan ontology terms. *Nodes in this network represent biological processes as defined by MapMan ontology terms. Node colors and numbers depict the different MapMan terms (legend at left), while edges represent significant ($P < 0.001$) associations between the terms based on coexpression. OPP, Oxidative pentose pathway; PS, photosynthesis; CHO, carbohydrate.*

These significantly associated terms were connected, and the resulting ontological network was visualized as an interactive network browser (Figure 3.8; http://aranet.mpimp-golm.mpg.de/aranet/Mapman_network). In order to get a more complete network we also retained connections representing parent-child relationships, which are trivial due to their mutual overlap.

From this visualization it became evident that terms that represent related processes tend to be connected, for example photosynthesis-related processes (dark green) were connected to plastidial protein synthesis (light blue) and to "protein assembly and co-factor ligation" which comprises many proteins involved in the assembly of the plastidial apparatus (light blue). Furthermore, the chloroplast cluster (dark green) is closely associated with genes related to tetra-pyrrole biosynthesis (light green; Figure 3.8). These processes most likely reflect parts of the basal plastidial photosynthetic activity program. Other examples were mitochondrial processes linked to the TCA cycle as well as polyamine synthesis being coupled to arginine degradation more than would be expected by the trivial link of arginine decarboxylase which is present in both processes. Also arabinogalactn proteins were linked to abiotic stress which is in-line with their upregulation upon salt stress (Lamport et al., 2006).

Since biologically relevant associations were confirmed in the MapMan ontology network, we also investigated associations between other biological processes, which were previously unrelated MapMan terms and which might help to generate new functional insights. Interestingly, plant defensins were connected to sphingolipid biosynthesis *in planta*. As often the mode of action of plant defensins seems to be mediated by sphingolipids of the attacking pathogen (Thevissen et al., 2000; 2005; Ramamoorthy et al., 2009), it could be speculated, that plant sphingolipids might play a role in this mechanism as well. Furthermore it might be interesting to investigate what caused the link introduced between aromatic amino acid degradation and starch breakdown (Fig 3.8, lower left corner). Thus, the combination of co-expressed gene vicinities and ontology terms may similarly reveal new associations between different processes in the cell.

3.4 Conclusions

We have constructed an interactive correlation network for *Arabidopsis* using a novel heuristic clustering algorithm (HCCA). The cluster solutions obtained from this clustering algorithm performed as well, or better, than the commonly used clustering algorithms MCL

and k-means. More importantly, by visualizing the portioned clusters we could reassemble the network, and we were therefore able to place the obtained partitions into larger biological contexts. We predicted that unique, well connected genes with certain biological functions tend to be more essential than other genes, and confirmed this by mutant analyses. The presented data therefore show that comprehensible visualization of genome-scale correlation networks may render new insights into the wiring of biological systems. We propose that this type of network visualization constitutes an easy-to-navigate framework for biologists to prioritize genes for functional analyses.

3.5 Materials and Methods

Microarray Data

All calculations for this work were done using python and java scripts. Databases for *Arabidopsis* yeast and E.coli use Affymetrix ATH1 (22 810 probe sets), Affymetrix Yeast Genome S98 (9 335 probe sets) and Affymetrix Ecoli_ASv2 (7312 probesets) GeneChips, respectively. *Arabidopsis* microarray datasets consisting of 1428 ATH1 microarrays were obtained from TAIR (http://www.*Arabidopsis*.org/). Separate *Arabidopsis* tissue atlas datasets containing 121 microarrays, which were used for plotting the gene expression across *Arabidopsis* tissues, were generated by the AtGenExpress project (Schmid et al., 2005), and obtained from TAIR. The data was quality controlled by visual inspection of boxplots of raw PM data and RMA residuals of RMA normalized data, using RMA express program. Cel files showing artifacts on RMA residual plots or visibly deviating from the majority on the PM-boxplots were removed from further analysis. In addition, we removed experiments representing very similar transcriptomic snapshots by iteratively discarding microarrays that displayed Pearson correlation ($r(A,B) \geq 0.95$) to more than three other microarrays. From these analyses we retained 351 microarrays, which subsequently were normalized using R package simpleAffy. The 244 E. coli and 789 yeast microarray datasets used to generate Figure 3.1 were downloaded from GEO (http://www.ncbi.nlm.nih.gov/geo/), RMA normalized and quality controlled as for the arrays for *Arabidopsis*. Names of the cel files used to construct the *Arabidopsis* HRR networks are downloadable from AraGenNet's homepage.

Phenotypic data for *Arabidopsis thaliana*

Phenotypic data for *Arabidopsis* was requested and obtained from TAIR curators, and was divided into essential, gametophytic lethal and non-lethal sets. All the expression data, co-expression network and phenotypic data presented in this work is downloadable from AraGenNets homepage (http://aranet.mpimp-golm.mpg.de/aranet).

Construction of Co-expression Networks

Pearson-based co-expression networks were used for the centrality vs. essentiality study, and for generating log-log plots. These networks were created using the 351 ATH1 microarrays described above. An edge in the network represents two genes with Pearson correlation ($r(A,B) \geq 0.8$). All subsequent analyses were done on highest reciprocal rank (HRR) based networks, including the visualized interactive co-expression network used on the AraGenNet homepage. HRR score between genes A,B is calculated according to:

$$HRR(A,B) = max(r(A,B), r(B,A)) \qquad (3.1)$$

where r(A,B) is correlation rank of gene B in gene A's co-expression list. Any two genes that were present in each others top 10, 20 or 30 correlation lists were connected by green, orange or red connections, respectively. Edges representing HRR values 10, 20 and 30 were assigned weights 1/5, 1/15 and 1/25, respectively.

Any two clusters in the meta-network were connected if the connectivity score exceeded 0.02 according to formula 3.2:

$$c(A,B) = \frac{\frac{\sum_{i \in \{clusterA's\ connections\ to\ clusterB\}} w_i}{\sum_{j \in \{clusterA's\ total\ outgoing\ connections\}} w_j} + \frac{\sum_{k \in \{clusterB's\ connections\ to\ clusterA\}} w_k}{\sum_{l \in \{clusterB's\ total\ outgoing\ connections\}} w_l}}{2} \qquad (3.2)$$

where

$$w \begin{cases} \frac{1}{5}, \text{green edge} \\ \frac{1}{15}, \text{orange edge} \\ \frac{1}{25}, \text{red edge} \end{cases}.$$

We used $c(A,B) \geq 0.02$, which connects clusters A a nd B, if the average mutual weights of edges between the two clusters exceed 0.02. The connectivity score can range from 0 (no

edges between the clusters) to 1 (all outgoing connections from cluster A are targeted to cluster B, and vice versa).

Comparison of a Pearson Correlation Network and a GGM Network

Our Pearson correlation network (R=0.8) was compared to a datasets from a recently published Graphical Gaussian (GGM) network (Ma et al., 2007), and common edges were identified by set comparisons. Approximately one third of the edges in the GGM network were also present in our network, consistent with a previous comparison made between the GGM and a Pearson correlation network (Ma et al., 2007).

Centrality vs. Essentiality

To assess the association of node degree (number of nodes a node is connected to) with phenotype characteristics (essential or non-essential), a node degree of genes showing a phenotype vs. those not showing any phenotype was compared. This was done across 20 co-expression networks generated by using Pearson r-values ranging from 0.9 to -0.9 (steps of 0.1). The median node degree of genes showing a phenotype was compared to median node degree of genes not showing any phenotype at a given r-value cutoff. Significant differences (Wilcoxon test $p<0.05$) in the median node degree between these two classes was used to indicate significant differences between the two classes.

HCCA clustering algorithm

The HCCA can be implemented by a pseudo-code available from the AraGenNet's homepage, and the full source code is available upon request from the authors. A simplified description of the algorithm is depicted in Figure 3.2, and in the Results and Discussion section. Python implementation of HCCA, together with sample networks, is available from AraGenNet's homepage.

Markov Clustering (MCL)

We used the available C code (http://micans.org/mcl/; van Dongen, 2000) for MCL calculations. The method simulates random walks on the graph, with the walking probability respecting the weight, i.e. HRR values, of the edges (HRR value of 10 received weight 1/5,

20 received 1/15, and 30 received 1/25). We used different inflation values, which are the Hadamard power of a stochastic matrix that gives the probabilities for the random walk. Low inflations result in slower random walks, and vice versa. The inflation parameter may range from ≥1 to 5, where small values generate fewer but larger clusters.

K-means Clustering (k-means)

To partition probe sets based on the original data, the expression values for each probe set were centered, scaled, and then subjected to the k-means clustering procedure provided by R using the default Hartigan and Wong algorithm (Hartigan and Wong, 1979).

Comparison of Clustering Solutions

The clustering solutions were judged by *modularity* (Newman and Girvan, 2004) that evaluates the graph partitioning by comparing the sum of edge-weights within clusters with edge-weights linking different clusters. This value is subsequently subtracted by the value that one expects for random partitions. The obtained modularity score ranges between -1 and 1, where 1 represents perfect modularity, 0 represents value expected by chance, and -1 represents value worse than expected by chance.

The partitions were also evaluated by the Davies-Bouldin index (Davies and Bouldin, 1979) using the clusterSim R-package. It is defined by formula 3.3:

$$DB = \frac{1}{n}\sum_{i=1}^{n} max_{i\neq j}\left\{\frac{S_n(Q_i)+S_n(Q_j)}{S(Q_i,Q_j)}\right\} \quad (3.3)$$

with *n* number of clusters, S_n average distance of all objects from the cluster to their cluster center and $S(Q_iQ_j)$distance between two cluster centers. DB score can range from 0 to infinity. Values close to 0 are achieved by good (distant) clustering. However, the value of zero is gained by just one big cluster.

We used adjusted Rand indices to compare two clustering solutions by pairwise affiliation of nodes (Hubert and Arabie, 1985). The scores for biological significance of clusters were calculated using the approximate mutual information between the clustering and MapMan categories (Usadel et al., 2006) having at least 10 members. In the case when the clustering solution did not assign all genes to clusters, only those that could be assigned

where considered. To make the HCCA clustering comparable to k-means, genes not assigned to any cluster by HCCA were not subjected to k-means, as these genes are most likely difficult to cluster. From this mutual information value, the mean mutual information from 1000 random assignments (denoted by $\overline{MI}$) with preserved cluster sizes was subtracted, and the result was divided by the standard deviation (denoted by σ) of these random mutual information values, according to formula (3.4).

$$S = \frac{MI_{cluster} - \overline{MI_{random}}}{\sigma_{random}} \quad (3.4)$$

Overrepresentation Analysis

In order to identify terms which might be associated we randomly sampled approximately 700 non-overlapping NVNs from the whole network and tested for a significant overrepresentation of MapMan terms within these clusters using a Fisher exact test ($p<0.05$ after Benjamini Hochberg correction). This was repeated several times to exclude random effects. Subsequently, we tested for significant co-occurrence of overrepresented terms using again a Fisher exact test.

Uniqueness vs. Essentiality Estimates

In order to group *Arabidopsis thaliana* genes into gene families, a BLASTCLUST analysis on *Arabidopsis* protein sequences obtained from TAIR was performed. Length coverage threshold of 70% and score coverage threshold were used as parameters.

We used random sampling to investigate whether there is correspondence between a gene having essential or non essential characteristics, and its uniqueness in the genome or node vicinity network. So far, 261 genes are characterized as being essential (phenotypic data from TAIR), and 152 of these are single copy genes based on the settings above. To investigate whether essential genes tend to be single copy, we sampled 261 random nodes 1000 times and counted the number of single copy genes acquired in each sampling. To investigate whether essential genes that do belong to gene family tend to be unique in the network vicinity, we sampled 109 (261 total - 152 single copy) random nodes 1000 times. The number of genes unique or non-unique in network vicinity was then counted, and

represented as histogram. The same was done for non-essential genes with characterized non-lethal phenotype (1224 total, 422 single copy).

Plant Cultivation and Mutant Analysis

T-DNA knockout lines (Table 3.2) were obtained from the Nottingham *Arabidopsis* Stock Centre (Alonso et al., 2003). The seeds were surface sterilized, sown on plates containing MS media (1x Murashige and Skoog (MS) salts, 8 g L-1 Agar, 1X B5 vitamins, 10.8 g L-1 Sucrose) and incubated for 48 h at 4°C in the dark. The plates were then incubated for 7 days at 21°C with 16 h photoperiod. T-DNA insertions were confirmed using PCR (data not shown). Pictures of seedlings and siliques were done using Leica MZ 16 FA stereo microscope.

4. PlaNet: Combined sequence and expression comparisons across seven plant species

4.1 Abstract

Model organisms, such as *Arabidopsis*, are readily used in basic research due to resource availability and relative speed of data acquisition. A major goal is to transfer the acquired knowledge from these model organisms to species that are of greater importance to our society. However, due to large gene families in plants, the identification of functional equivalents of well characterized *Arabidopsis* genes in other plants is a non-trivial task, which often returns erroneous or inconclusive results. It is well documented that transcriptionally coordinated genes tend to be functionally related, and that such relationships may be conserved across different species, and even kingdoms.

To exploit such relationships we constructed whole genome co-expression networks for *Arabidopsis* and six important plant crop species. We clustered the networks using the HCCA algorithm, and provide interactive versions of the networks under the banner PlaNet (http://aranet.mpimp-golm.mpg.de). We attempted to assign biological functions to each cluster by assessing enriched ontology terms, and mutant phenotype associations. Importantly, we implemented a comparative network algorithm that estimates similarities between network structures. Thus, the platform can be used to swiftly infer similar co-expressed network vicinities within and across species and can predict the identity of functional homologs. We exemplify this using the co-expressed gene vicinities for the *PSA-D* and Chalcone Synthase genes in *Arabidopsis* as case studies. Finally, we assessed how ontology terms are transcriptionally connected in the seven species, and provide both individual MapMan co-expression networks, as well as a network containing the MapMan co-expressed terms across all seven species. We propose that this platform will considerably improve the transfer of knowledge generated in *Arabidopsis* to valuable crop species.

4.2 Introduction

Various rapidly evolving genomic and post-genomic technologies, including genome sequences and gene expression data, have greatly enhanced our understanding for how biological systems function. As of June 2010, over 1500 genomes from prokaryotic, eukaryotic and archae organisms have been fully sequenced, and over 5500 sequencing projects are in progress (Liolios et al., 2010). In parallel, transcriptional studies via DNA microarrays and deep sequencing methods have generated vast amounts of publicly available expression data for various organisms, with over 6000 microarray datasets available for *Arabidopsis* alone (GEO database, as of June 2010). In essence, the expression data has been generated, and subsequently mined, for hypothesis-driven gene discovery, for example to reveal transcriptional responses to certain genotypes or external stimuli, and for mining coordinate expression of different genes (Usadel et al., 2009). These types of analyses have facilitated the conclusion that functionally related genes tend to be transcriptionally coordinated, i.e. co-expressed (Stuart et al., 2003; Persson et al., 2005). Consequently, using "guilt-by-association" approaches, co-expression analyses have proved valuable for rapid inference of gene function, sub-cellular localization of gene products, and biological pathway discovery (Wei et al., 2006; Yonekura-Sakakibara et al., 2008; Usadel et al., 2009).

Organism	Affymetrix GeneChip*	Probesets	No. of chips*	No. of HCCA obtained clusters	Source database of coding sequences	Percentage of represented genes
Arabidopsis thaliana	ATH1	22,810	351	181	TAIR8 http://www.arabidopsis.org/	~63%
Barley	Barley1	22,840	116	195	harvEST Hv35 http://www.harvest-web.org/	N/A**
Medicago truncala, sativa	Medicago	61,263	105	360	IMGAG 27-02-2008 http://www.medicago.org/genome/IMGAG/	N/A**
Poplar	Poplar	61,413	69	400	Poptr 1.1 Jamboree http://genome.jgi-psf.org/poplar/poplar.home.html	~65%
Rice	Rice	57,380	83	530	Rice Genome annotation v 6.0 http://rice.plantbiology.msu.edu/	~60%
Wheat	Wheat	61,290	150	384	*Triticum aestivum* http://www.harvest-web.org/	N/A**
Soybean	Soybean	61,170	215	549	harvEST Gm 10-12-2009 http://www.harvest-web.org/	N/A**

Table 4.1. Detailed microarray information. **Microarray datasets used in this study and clustering algorithm are available at: http://aranet.mpimp-golm.mpg.de/aranet/downloads. **Due to lack of complete genome sequence, the estimation is not possible.*

However, while co-expression relationships in many cases can provide insight into biological processes and predict genes for functional testing, the representation of genomic content on the microarrays is not complete and the results are therefore also incomplete (Table 4.1). For

example, the widely used *Arabidopsis* ATH1 chip and the Affymetrix rice array cover approximately 63% and 60% of the genes in the *Arabidopsis* and rice genomes, respectively. It is therefore clear that certain transcriptional relationships are not revealed using microarrays. In addition, low spatio-temporal resolution of gene expression contributes to both false negatives, e.g. expression of genes may be rendered as noise due to activity in only specific cell types or stimuli, and false positives, e.g. difficulties in distinguishing pollen and ovule specific genes if only flowers were measured. These caveats should prompt caution by biologists in over-reliance, or at least over-interpretation, of "whole-genome" expression analyses.

Arabidopsis, as the most studied plant species, has approximately 50% of its genes functionally annotated by sequence homology, and approximately 11% of the genes are associated with distinct biological functions that have been experimentally verified (Saito et al., 2008). Still, a major goal is to transfer the knowledge obtained in a model organism (donor), such as *Arabidopsis*, to other species (acceptors), which may be of greater importance for society. After the exact function of a gene product in the knowledge donor has been proven experimentally, uncovering the identity of the functional equivalent in an acceptor species is, however, not trivial. As plants generally hold large gene families, sequence comparison of a gene from the knowledge donor to the genome of the acceptor can return a large list of possible candidate genes. While several of those candidates may perform the same molecular function, they are not necessarily part of the biological process of interest. Intuitively, a functional homolog should be present when the relevant biological process occurs. Thus, functional homologs from different species should be reflected in conserved co-expression patterns. Indeed, several studies have showed that co-expressed relationships are conserved across species and even kingdoms (Stuart et al., 2003; Bergmann et al., 2004). Thus, a functional homolog may be identified by combined sequence and co-expression approaches.

Several web-tools that combine co-expression analyses with sequence, protein-protein interaction, cis-element, and sub-cellular localization prediction have been created for individual plant species (Steinhauser et al., 2004; Manfield et al., 2006; Mutwil et al., 2008; Srinivasasainagendra et al., 2008; Obayashi et al., 2009; Mutwil et al., 2009). The representation of co-expressed relationships as networks has transcended standard single gene analyses, since this enables the biologist to more readily contextualize their genes or proteins of interest (Mao et al., 2009; Mutwil et al., 2010).

Here, we present PlaNet (Plant Network), a platform that integrates genomics, transcriptomics, phenomics and ontology analyses across seven plant species important both for research and human circumstances (http://aranet.mpimp-golm.mpg.de). For comparative analyses we implemented NetworkComparer, a novel pipeline that compares and displays commonalities and differences between the co-expressed vicinity networks (VNs) across selected species. Importantly, considering the incomplete gene coverage of the microarray probes, comparative analysis between species provided insight into the association of a gene with certain processes despite the absence of corresponding microarray probe. We demonstrate the features of the platform by two examples, the photosynthesis related *PSA-D1* gene and several Chalcone Synthase (*CHS*) genes in *Arabidopsis*. However, we are convinced that its utility extends well beyond both these examples, and also our own research interests, which is why we are making it available as a community resource.

4.3 Data sources, construction and structure of PlaNet

Affymetrix microarray datasets (summarized in Table 4.1) for seven plant species (*Arabidopsis*, barley, rice, Medicago, poplar, wheat and soybean) were obtained from GEO (Edgar et al., 2002) and ArrayExpress (Parkinson et al., 2009), and were subjected to deleted residual quality control to remove possible array errors (Persson et al. 2005). The resulting arrays are summarized in Table 4.1, and can be downloaded from http://aranet.mpimp-golm.mpg.de. The starting networks were generated by calculating the degree of co-expression using Pearson correlation cut-off of 0.8 between genes for the respective species, similar to what has been described previously (Mutwil et al., 2010). Social and biological networks generally follow power law distribution (Barabási and Oltvai, 2004). This implies that most of the network nodes are connected to only few other nodes, while a small number of nodes are connected to many other nodes. Indeed, we also observed such relationships in the co-expression networks of each of the seven plant species (Figure 4.1). For visualization of the expression relationships we used the highest reciprocal rank (HRR) between any two genes as a measure (Mutwil et al., 2010), given that it has been demonstrated that rank-based associations produce robust co-expression analyses (Obayashi and Kinoshita, 2009).

As whole genome-scale networks are too large and complex for comprehensive visualization, we first partitioned the networks into manageable clusters, using Heuristic Cluster Chiselling Algorithm (HCCA) with step=3 (Mutwil et al., 2010). In essence, HCCA finds clusters by generating putative clusters for every node in the graph and then recursively

remove nodes that show higher connectivity to nodes outside of a cluster compared to nodes within the cluster. During each recursion new clusters are generated until all nodes are assigned to clusters (Mutwil et al., 2010). HCCA was chosen since this algorithm supports weighted edge graphs and permits the user to specify their own desired cluster size. The latter is crucial for visualization of large networks since large clusters (>400) often are too dense for visual inspection, and conversely small networks (<10) are often biologically meaningless. When the desired cluster size interval is set to 40-200 nodes the algorithm yields between 181 and 549 clusters for the seven plant species (Table 4.1). Following this example further, we used Graphviz (www.graphviz.org) to calculate the layout of the networks. The resulting interactive clusters (available at http://aranet.mpimp-golm.mpg.de) represent co-expressed genes, presumably involved in related biological processes.

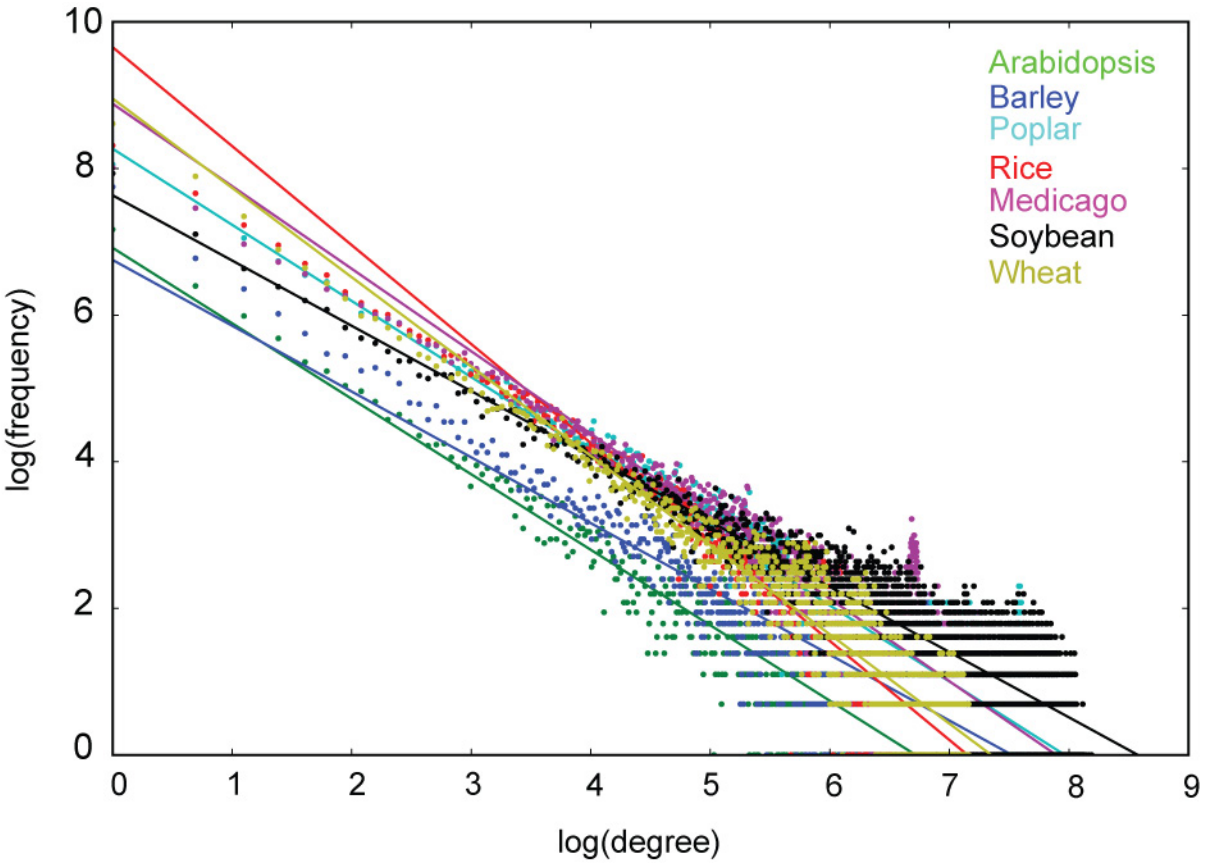

Figure 4.1. Network characteristics. *Log-log plot of node degree distribution for Pearson correlation networks ($r \geq 0.8$) from Arabidopsis (green), Barley (dark-blue), Poplar (light-blue), Rice (red), Medicago (purple), Soybean (black), and Wheat (Mangenta). The x axis represents the node degree (i.e. the number of connections a node holds), and the y axis displays the frequency (i.e. the number of genes) showing this degree.*

To interrogate the resulting networks, the user can specify their gene of interest by probeset ID, gene ID, nucleotide/protein sequence or keyword, which redirects the user to the corresponding gene cluster or to an individual gene specific page (Figure 4.2). The latter page contains the expression profile of the gene across different tissues, a step=2 vicinity network surrounding the gene (i.e. genes co-expressed within two steps from the selected gene), phenotypes found in the VN, and Mapman and Gene Ontology (GO) analyses (Figure 4.2). The phenotype associations are displayed as color-coded nodes, where red, yellow and green represent embryo lethal, gametophytic lethal, and non-lethal phenotypes, respectively. Moving the mouse pointer over a node opens a pop-up window displaying annotation and phenotypic information of a gene, while clicking a node redirects the user to a page dedicated to the corresponding gene.

Partitioning of any object into smaller units indisputably removes information about how the units are arranged to make up the object. To avoid loss of such valuable information we connected the clusters based on mutual co-expression relationships to form a network of clusters, which should reflect the organisation of the genome-wide co-expression network. The resulting "meta-networks" rendered from this analysis thus depict relations between co-expressed gene clusters (Figure 4.2), i.e. a node in this type of network is a HCCA obtained cluster of co-expressed genes (Mutwil et al., 2010). Any two clusters in the meta-network are connected if the number of edges between them exceeds a linkage threshold $\beta \geq 0.02$ (Mutwil et al. 2010). We chose this linkage threshold value to $\beta \geq 0.02$, as it produced informative yet readable network structures. Given that any cluster in the "meta-network" contains genes which are co-expressed to one another, we anticipated that the majority of these genes should be involved in related biological processes. Inferring such relations is, however, not trivial as many genes are not associated with useful annotations. We attempted to get around this problem by combining MapMan and Gene Ontology (GO), available phenotypic data, and tissue dependent expression profiling. For example, the majority of genes in the *Arabidopsis* cluster 77 show ubiquitous expression profiles, and mutations in the genes often show pale green phenotypes or are embryo lethal (Figure 4.2).

MapMan and GO analysis revealed that this cluster is strongly enriched for genes associated with fatty acid elongation and to lesser degree with amino acid synthesis (Table 4.2; http://aranet.mpimp-golm.mpg.de/aranet/ac77). Based on the combined information from these analyses we predict that cluster 77 holds genes involved in lipid metabolism and chloroplast development.

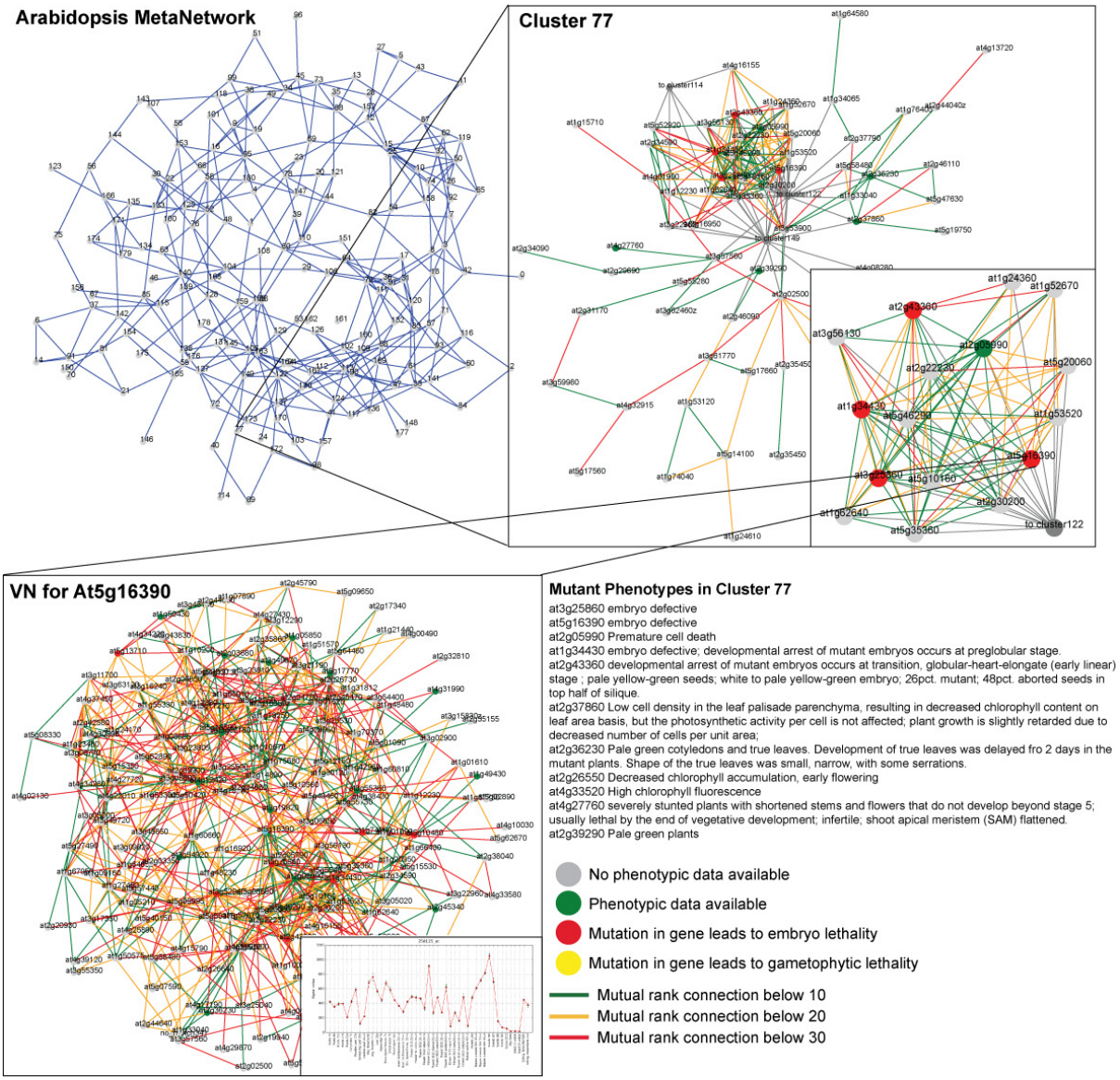

Figure 4.2. Outline of the PlaNet platform. *MetaNetwork (upper left). Each node in this network represents a cluster of co-expressed genes (upper right). We mapped available phenotypic data onto the genes (blow-up insert upper right; red indicates embryo lethality; yellow indicates gametophytic lethality; and green indicates other reported phenotypes, that results when the gene is mutated). The coloured edges indicate strength of the co-expression based on mutual ranks between the individual gene pairs (green indicates a mutual rank below 10; orange indicates a mutual rank between 10 and 20; and red indicates a mutual rank between 20 and 30). Each of the genes in the cluster can also be displayed with its node vicinity network (VN) in which the gene of interest is centered, and the surrounding co-expressed genes are displayed (lower left). This lay-out also includes the expression profile*

of the gene across different tissues. All pages also give information about phenotypic data for the genes in the cluster (Lower right), and enriched ontology terms.

Bin number	Bin annotation	p-value
11	lipid metabolism	2.87E-19
11.1	lipid metabolism,FA synthesis and FA elongation	7.75E-24
11,1,1	lipid metabolism,FA synthesis and FA elongation,Acetyl CoA Carboxylation	6.04E-07
11,1,2	lipid metabolism,FA synthesis and FA elongation,Acetyl CoA Transacylase	0.003165392
11,1,3	lipid metabolism,FA synthesis and FA elongation,ketoacyl ACP synthase	5.90E-05
11,1,4	lipid metabolism,FA synthesis and FA elongation,ACP oxoacyl reductase	0.003165392
11,1,5	lipid metabolism,FA synthesis and FA elongation,beta hydroxyacyl ACP dehydratase	9.88E-06
11,1,6	lipid metabolism,FA synthesis and FA elongation,enoyl ACP reductase	0.003165392
11,1,12	lipid metabolism,FA synthesis and FA elongation,ACP protein	0.021951303
11,1,30	lipid metabolism,FA synthesis and FA elongation,pyruvate kinase	2.96E-05
11,1,31	lipid metabolism,FA synthesis and FA elongation,pyruvate DH	3.03E-07
11.3	lipid metabolism,Phospholipid synthesis	0.011398276
11.9	lipid metabolism,lipid degradation	0.057192339
11,9,2	lipid metabolism,lipid degradation,lipases	0.110769082
11,9,3	lipid metabolism,lipid degradation,lysophospholipases	0.133078577
11,9,3,2	lipid metabolism,lipid degradation,lysophospholipases,carboxylesterase	0.012602403
13	amino acid metabolism	0.000214568
13.1	amino acid metabolism,synthesis	1.71E-05

Table 4.2. Mapman terms associated with Cluster 77 in Arabidopsis.

Analogous to the enrichment of certain biological processes within a cluster, connected clusters also share co-expressed gene pairs and may therefore also be involved in related processes. One such example is evident for genes grouped in the connected clusters 6, 14, 21, 59, 81, 91, 121, 137 and 142 in *Arabidopsis*. Mutations in many of these genes display pale green phenotypes, or result in embryo lethality (data not shown). Most of the genes associated with these clusters are also ubiquitously expressed with the exception of roots, and the clusters are enriched for MapMan ontology terms such as protein targeting to chloroplast, plastid protein synthesis, and photosystem light reaction. We, therefore, find it likely that many of these clusters are associated with chloroplast development and photosynthesis. Many other groups of clusters are also enriched for certain biological functions, such as cell division, protein synthesis, defense and tissue specific development. It may, therefore, be useful to not only explore the direct VN of the gene of interest for functional context, but also to evaluate neighboring clusters for a higher contextual order.

The interactive gene-related networks in PlaNet may thus be browsed on three different levels for each of the individual species: as meta-networks displaying inter-

connectivity between gene clusters, as individual gene clusters, and as single gene pages with surrounding gene VNs (Figure 4.2).

4.4 Comparative co-expression relationships across seven plant species.

The co-expressed network arrangements of the individual species may reveal genes and processes which are associated with the function of a gene of interest. However, due to incomplete coverage of the arrays (Table 4.1), and also of the analyzed datasets, the individual networks most likely lack some candidate genes as well as containing false positive candidates. Here we argue that by comparing network structures across species we may enrich for genes related to the particular biological function of interest. In addition, considering the comparably high knowledge about gene functions in *Arabidopsis* we anticipated that by comparing VNs to other species we would be able to readily infer functional homologs in other species that are likely to have higher societal value.

To compare different network vicinities we included a NetworkComparer pipeline, which can score and display conserved co-expression network structures across species by combining gene sequences with co-expression network structure (Figure 4.3), in the PlaNet platform. For the sequence comparison we binned genes present on the arrays into Protein families (binning downloadable from http://aranet.mpimp-golm.mpg.de/aranet/Downloads). We next obtained probeset target sequence information from Affymetrix homepage (www.affymetrix.com). The probesets were mapped using BLAST to the corresponding best-hit coding sequence as defined by the most current data from genome assembly databases (Table 4.1). Probesets with no gene hit or expected values higher than 0.01 were excluded from further analysis. Then, using reversed position specific BLAST (Marchler-Bauer et al., 2007) with a cut-off expected value of 0.01 we assigned each gene to the best-hit protein (PFAM v23.0) family (Finn et al., 2008). The NetworkComparer pipeline can compare either user defined genes to each other, or a specific gene to all members of the associated protein family (Figure 4.3). Here we illustrate the principle and application of the pipeline by two examples, photosynthesis and the chalcone synthase pathway, however, we also stress that such an approach can be adopted for any gene/pathway/process of interest.

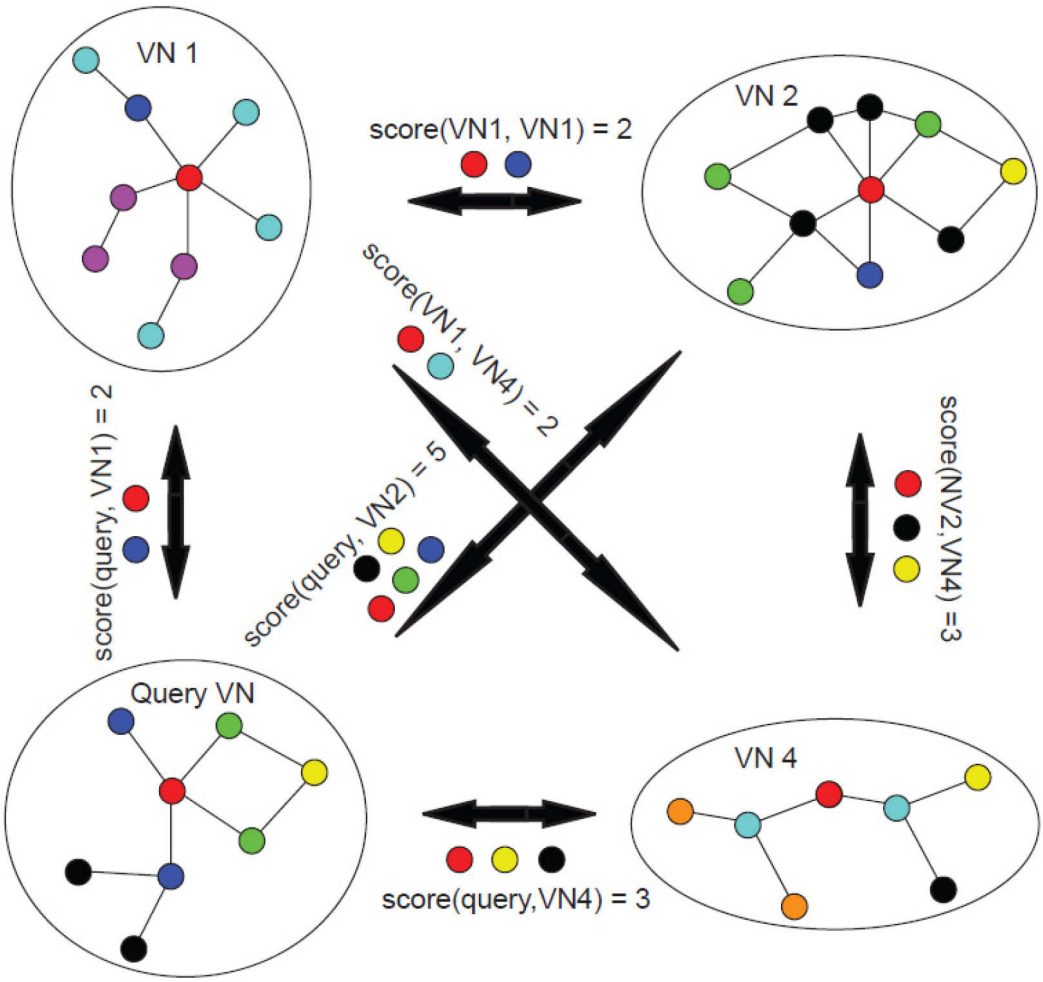

Figure 4.3. Outline for the NetworkComparer pipeline. *Similarities of vicinity networks (VN) are scored by counting the amount of gene families (depicted by colored nodes) they have in common.*

4.4.1 Photosynthesis – AtPSA-D1 and AtPSA-D2

The two *Arabidopsis* genes *AtPSA-D1* and *AtPSA-D2* are co-expressed and belong to the photosystem I reaction center (PSA-D) family, necessary for assembly of the PSI complex (Ihnatowicz et al., 2004). BLAST and phylogenetic analysis of the PSA-D family revealed two, two, three and one members from *Arabidopsis*, Medicago, poplar and rice, respectively (Figure 4.4A). As only a single *PSA-D* related copy is present in rice it appears easy to infer that this gene should represent a functional homolog to the *Arabidopsis* PSA-Ds. Consistent with such idea, the VN for the rice *PSA-D* gene contains many genes for which homologs are found also in the *AtPSA-D1* VN (data not shown). However, the sequence divergence between the AtPSA-Ds and the two Medicago PSA-D proteins is minute. Similarly, the two Poplar PSA-Ds are also at approximately equal sequence distance to the AtPSA-D proteins

(Figure 4.4A). It therefore seems rather difficult to predict which of the PSA-D proteins in poplar and Medicago that have the most closely related biological function to the AtPSA-Ds.

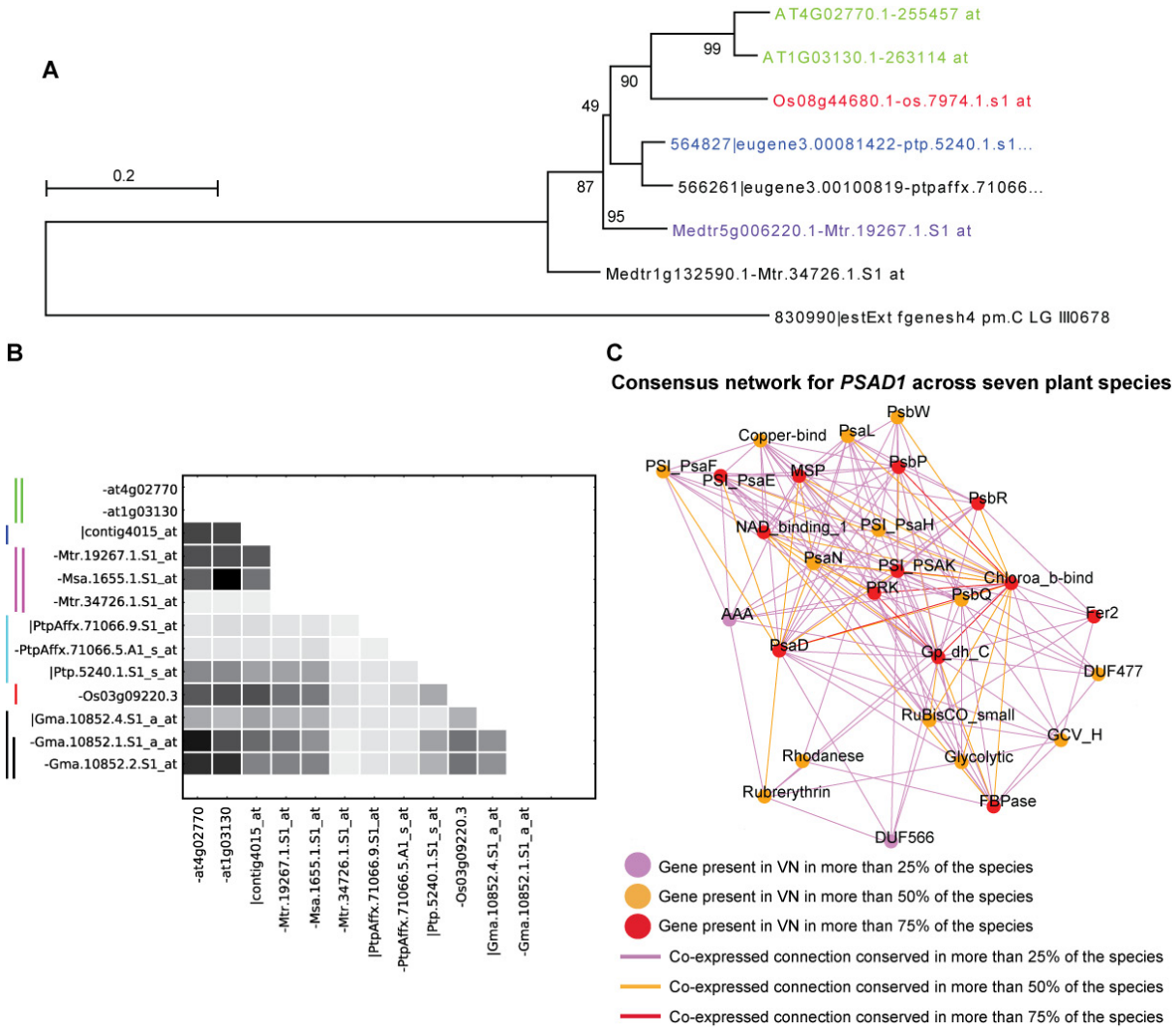

Figure 4.4. Co-expression vicinity networks for PSA-D-related genes across seven plant species. A. *Phylogenetic tree of PSA-D-related proteins in Arabidopsis, poplar, rice, and Medicago. Colour coded proteins/probe sets corresponds to; Green; Arabidopsis PSA-D1and D2, Red; rice PSA-D with a similar VN to AtPSA-D1 and AtPSA-D2, Light-blue; poplar PSA-D with a similar VN to AtPSA-D1 and AtPSA-D2, Purple; Medicago PSA-D with a similar VN to AtPSA-D1 and AtPSA-D2, and Black; Proteins for which the gene VNs have low similarity to the AtPSA-D1 VN. Protein sequences were aligned in MEGA4 (Tamura et al., 2007) using ClustalW and phylogenetic trees were constructed using bootstrapped Neighbour-Joining method (1000 runs).* ***B.*** *Heat-map of co-expressed VNs of PSA-D-related genes across the seven species. X- and Y-axes represent PSA-D-related genes for which VNs were compared. Darker colours indicate more similar VNs. White areas indicate that the two VNs in a species are overlapping, and the values are therefore excluded (inclusion would*

result in artificially high values since the VN areas are overlapping and therefore include the same genes). Coloured vertical lines indicate the different species; Green; Arabidopsis, dark-blue; barley, purple; Medicago, light-blue, poplar, red; rice, and black; soybean. Two lines indicate that the genes are in the same gene VNs. ***C.*** *Combined co-expressed gene VNs for the seven different plant species for PSA-D genes. Colour coded nodes and edges represent presence of certain genes and connections across species. Red, orange and pink nodes indicate that a gene homolog is found in the VN of more than 75%, 50% and 25% of the species, respectively. Similarly, red, orange and pink edges indicate that an edge is found between the two connected gene homologs in more than 75%, 50% and 25% of the species, respectively.*

We attempted to predict which of these, and other *PSA-D*, genes most closely resemble the *AtPSA-D* gene function and therefore used one of the *PSA-D* genes from *Arabidopsis*, *AtPSA-D1* (At4g02770), as query for the NetworkComparer. The pipeline found 13 *PSA-D* associated probesets in the six plant species (Fig 4.4), which were passed on to the comparative analysis. The algorithm implemented in the NetworkComparer first generates VNs for each of the 13 probesets by taking in all co-expressed nodes that are within two steps (step=2) from each probeset (Fig 4.3).

Co-expressed group: 1 **at4g02770** at1g03130	Average score of group:37.0 40 34
Co-expressed group: 2 **contig4015_at**	Average score of group:22.0 22
Co-expressed group: 3 **Mtr.19267.1.S1_at** Msa.1655.1.S1_at Mtr.34726.1.S1_at	Average score of group:14.0 21 19 2
Co-expressed group: 4 PtpAffx.71066.9.S1_at	Average score of group:3.0 3
Co-expressed group: 5 PtpAffx.71066.5.A1_s_at	Average score of group:3.0 3
Co-expressed group: 6 **Ptp.5240.1.S1_s_at**	Average score of group:14.0 14
Co-expressed group: 7 **Os03g09220.3**	Average score of group:20.0 20
Co-expressed group: 8 Gma.10852.4.S1_a_at	Average score of group:10.0 10
Co-expressed group: 9 **Gma.10852.1.S1_a_at** Gma.10852.2.S1_at	Average score of group:23.5 24 23

Table 4.3. NetworkComparer table showing similarity scores of PSAD-1 related probesets to PSAD-1 from Arabidopsis. *Bold probesets were selected for further analysis.*

The VNs are then compared to one another in a pair-wise fashion, where the score value between any two VNs equals the number of protein families they have in common (Fig 4.3).

Thus, VNs with highly similar protein family content should show high mutual score. Results of this comparison are shown as a heatmap and a table. The heat-map graphically reveals the similarity scores of VNs of all probe-sets in the protein family of the query gene. Table 4.3 shows the similarity scores of the query gene to all genes from the analyzed protein family.

It is important to keep in mind that some members from the same gene family may be co-expressed, i.e. present in one another's VNs, and comparison of such genes will, therefore, return an artificially large comparison score. To avoid such artificial enrichments, the pipeline bins the overlapping VNs into co-expression groups. For the *PSA-D* family, nine such co-expression groups were found across the six plant species. The heat-map shows that three probesets representing Medicago *PSA-D* genes are present in one co-expression group (Figure 4.4, Table 4.3). Of the three corresponding genes, Medtr5g006220.1 (represented by probeset Mtr19267.1.S1_at) showed the highest score to co-expression network of *AtPSA-D1* (Msa.1655.1.S1_at could not be readily associated with any current gene models). In addition, of the three *PSA-D* genes from poplar, gene model 564827|eugene3.00081422 (represented by probeset ptp.5240.1.s1_s_at) showed the highest score to *AtPSA-D1* (Fig 4.4, Table 4.3). Taken together, these exemplary results suggest that the second step of the NetworkComparer pipeline can be used to identify potential functional homologs across the different species.

To analyze the commonalities between the *PSA-D* associated networks we chose the highest scoring *PSA-D* gene, i.e. the *PSA-D* genes with the most similar gene VNs to *AtPSA-D1*, from each species (Table 4.3), and sent these to the final step of the analysis. In this step the pipeline extracts and displays the common features of the selected gene VNs in form of a combined network and a table (Figure 4.4C). The combined co-expression VN depicts the frequency for which a given protein family is found in the selected co-expression networks, where red, orange and pink nodes and connections correspond to protein families and family associations found in >75%, >50% and >25% of the networks, respectively.

The co-expression networks of the selected *PSA-D* genes showed strong enrichment of *PSA* and *PSB* gene families which are components of photosystem I and II complexes (Figure 4.4C, Table 4.4; Nelson and Yocum, 2006). In addition, several other genes not directly associated with the photosystem complexes, but with ATP generation, such as ATP synthase, glyceraldehyde 3-phosphate dehydrogenase and triose phosphate transporter family were also present in the network. Interestingly, two Domain of Unknown Function families, DUF566 and DUF477, were present in four and three of the co-expression networks

analyzed, respectively, suggesting association of those families with the biological function of the PSA-D gene products.

	Arabidopsis	Barley	Medicago	Poplar	Rice	Soybean
Description:	at4g02770	contig4015_at	Mtr.19267.1.S1_at	Ptp.5240.1.S1_s_at	Os03g09220.3	Gma.10852.1.S1_a_at
PetM family of cytochrome b6f complex subunit.	at2g26500			PtpAffx.6372.1.S1_s_at		
PsaD. This family consists of PsaD from plants and cyanobacteria	at4g02770 at1g03130	contig4015_at	Msa.1655.1.S1_at Mtr.19267.1.S1_at	Ptp.5240.1.S1_s_at	Os03g09220.3	Gma.10852.2.S1_at
Copper binding proteins, plastocyanin/azurin family.	at1g76100 at1g20340	contig2142_s_at contig2141_s_at	Mtr.37317.1.S1_at		Os06g01210.1	
Photosystem I reaction centre subunit VI.	at3g16140 at1g52230	contig2247_s_at contig2247_at	Mtr.42999.1.S1_at		Os05g48630.2	Gma.16838.1.S1_at Gma.15376.1.A1_s_at
Family of unknown function (DUF566).		contig2859_s_at			Os04g33830 Os06g15400	
Domain of unknown function (DUF477).	at1g54780		Mtr.12223.1.S1_at	Ptp.805.2.S1_a_at		Gma.6887.1.S1_at

Table 4.4. Detailed identity information from the NetworkComparer analysis of the PSA-D related genes using At4g02770 as query. *The table is truncated due to space limitations.*

While the common network depicts enrichment and associations between protein families in the analyzed co-expression networks for the individual species, the associated table provides detailed information regarding the identity of probesets associated with the families (Table 4.4).

Apart from predicting functionally related genes and putative functional homologs across species, the table can also reveal functional redundancies. For example, using *AtPSA-D1* from *Arabidopsis* as query for the analysis, the pipeline detected *AtPSA-D2* to be present in the VN of *AtPSA-D1* (Table 4.4). Literature search reveals that whilst mutations in *AtPSA-D1* affect the photosynthetic electron flow, disruption of *AtPSA-D2* results in no observable phenotype (Ihnatowicz et al., 2004). This could perhaps be due to functional redundancy between the two gene products. Indeed, *atpsa-d1 atpsa-d2* double mutants result in an additive phenotype, i.e. seedling lethality (Ihnatowicz et al., 2004), supporting this hypothesis.

4.4.2 Flavonol and flavonoid synthesis - Chalcone Synthases

The *PSA-D* gene family is relatively small and we therefore also chose to approach a considerably larger gene family; the Polyketide Synthase family (PKSs; Austin and Noel, 2003; Abe and Morita, 2010). From within this family we chose the Chalcone Synthase (CHS) subfamily, which is one of the larger subfamilies of the PKSs. Whilst *Arabidopsis* only contains four *CHS* related genes, rice and Medicago have at least 20 CHS homologs each (Figure 4.5).

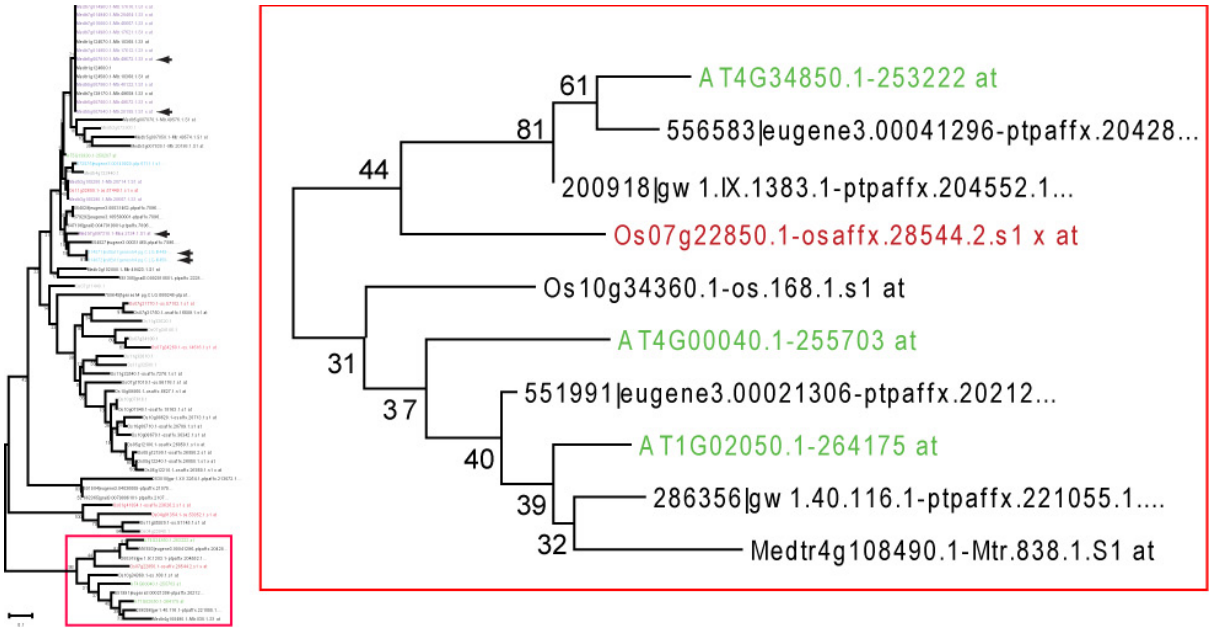

Figure 4.5. Phylogenetic tree of CHS-related proteins in Arabidopsis, Poplar, Rice, and Medicago. *Colour coded genes/probe sets corresponds to; Grey; gene not present on the array, Green; Arabidopsis genes with secondary metabolism related VNs, Red; Rice genes with secondary metabolism related VNs, Light-blue; Poplar genes with secondary metabolism related VNs, Purple; Medicago genes with secondary metabolism-related VNs, and Black; Genes with no significant secondary metabolism related ontology terms for their VNs. The boxed area is blown up and displays, together with genes/probesets indicated with arrow heads, putative CHS related genes associated with floral tissues for the four species. Protein sequences were aligned in MEGA4 (Tamura et al., 2007) using ClustalW and phylogenetic trees were constructed using bootstrapped Neighbour-Joining method (1000 runs). Values on the branches indicate bootstrap support in percent.*

The CHS gene products are associated with flavonoid-related biosynthesis pathways in which they catalyze the conversion of coumaroyl-CoA into naringenin chalcone (Austin and Noel, 2003; Abe and Morita, 2010; Figure 4.6). One of the more prominent *CHS* members in

Arabidopsis (At5g13930; *TT4*) has been experimentally associated with the main flavonol/flavonoid biosynthesis route (Feinbaum and Ausubel 1988), and is co-expressed with many of the genes for which the gene products work either up- and downstream of the CHS (Tohge et al., 2005; Yonekura-Sakakibara et al., 2008; Tohge and Fernie, 2010). These relationships may readily be seen in Figure 4.6, in which the general flavonoid biosynthetic genes, e.g. *CHS* (*TT4*, At5g13930), *CHI* (*TT5*, At3g55120), *F3H* (*TT6*, At3g51240), *F3'H* (*TT7*, At5g07990), *Fd3GT* (*UGT78D2*, At5g17050) and *4CL3* (*At4CL3*, At1g65060) are found in a central co-expressed cluster. In addition, this central network is connected to genes associated with anthocyanin production, such as *ANS* (*TT18*, At4g22880), *DFR* (*TT3*, At5g42800), *A3G2"XT* (*UGT79B1*, At5g54060), *A3GCoT* (At1g03940 and At1g03495; Luo et al., 2007), *A5GT* (*UGT75C1*, At4g14090) and *A5GMalT* (At3g29590), and also to genes associated with flavonol production, e.g. *FLS* (*AtFLS1*, At5g08640), *F3RT* (*UGT78D1*, At1g30530), *RHM1* (*ROL1*, At1g78570) and *MYB111*(*PFG3*, At5g49330). Interestingly, several light response genes, including *HY5*, *CRYD* and *PHR1*, are transcriptionally coordinated with the *TT4* gene (Figure 4.6).

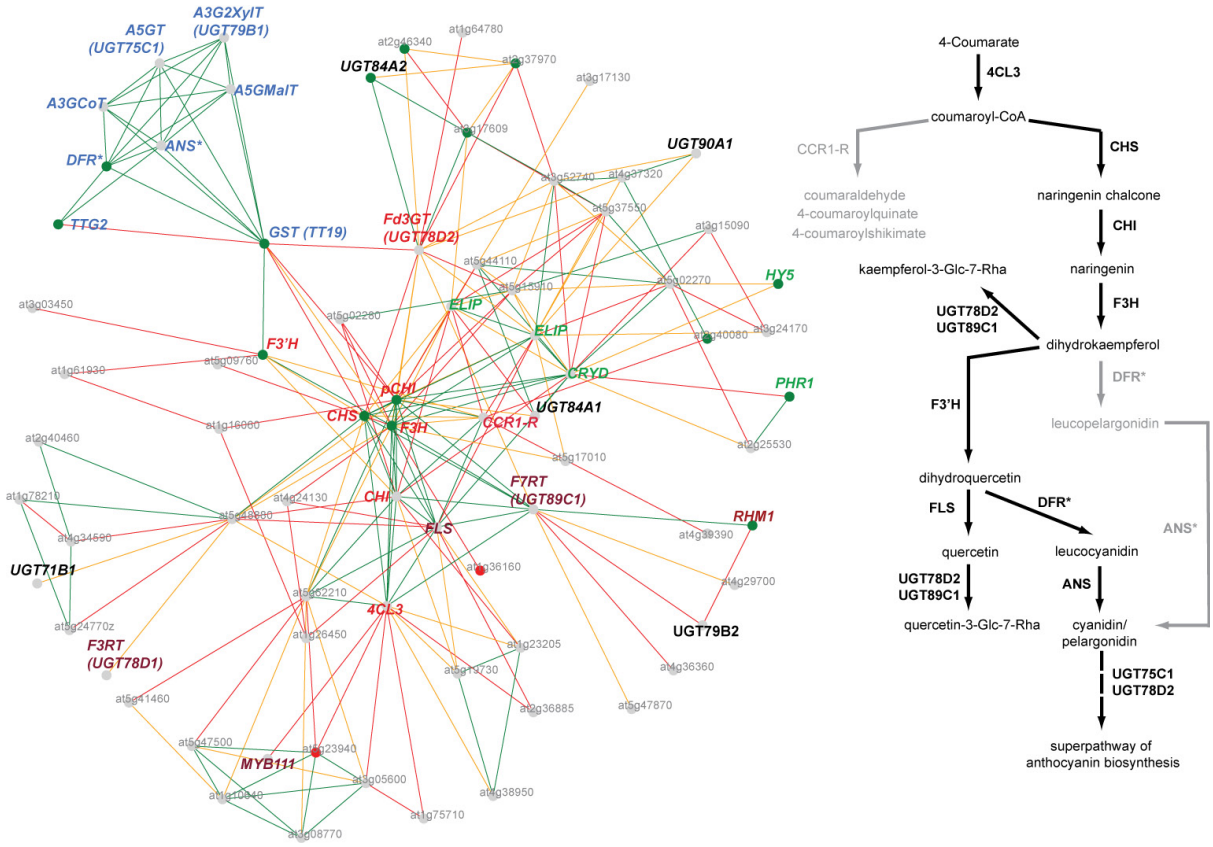

Figure 4.6. Co-expressed gene vicinity network for Chalcone Synthase (CHS; At5g13930) in Arabidopsis. *Many of the genes co-expressed in this VN (left) participate in the flavonol/flavonoid pathway (right) leading up to anthocyanin production. Network genes*

indicated in red, brown and blue are related to flavonoid, flavonol and anthocyanin production, respectively, and green are related to light responsive genes.

HY5 can activate the transcription factor Production of Flavonol Glycosides (*PFG*) in response to ultraviolet-B light (Stracke et al., 2010). Thus, this co-expressed gene cluster reveals links between various natural products, and also between transcriptional activators and biosynthetic genes

To assess similarities across the different species for CHS-related processes we used *TT4* as query gene for the NetworkComparer platform. The output from the tool resulted in approximately 30 co-expression groups, with varying degrees of VN similarities to the VN of the query gene (Figure 4.7). We selected the most similar VN from each species compared to the query gene VN (boxed in red and blue in Figure 4.7A), which yielded a combined co-expression network for the CHSs across the seven species (Figure 4.7B). From this network it is apparent that many of the VNs contain genes associated with the general flavonoid biosynthesis, including *CHI*s, *F3H*s, *FLS*s and *DFR*s (Figure 4.7C). Also, several genes that transport and modify flavonoids, such as OMTs and glycosyltransferases, and ABC-transporters, glutathione-S-transferases and sugar-transporters, are present in the VNs in multiple species (Figs. 4.7B and 4.7C). Flavonoids are generally accumulated as glycosylated forms in the vacuole. The conservation of both glycosyltransferses and transporters in the VNs across species suggest that both glycosylation events and the vacuolar transporting systems occur in all the species we studied here. Furthermore, genes that encode epimerases and certain transcription factors are also included in the combined network. NDP-sugar converting enzymes, such as UDP-rhamnose synthases can provide substrates for the glycosylation of flavonoids. Transcription factors, on the other hand, transcriptionally activate the biosynthetic pathway genes. It is important to note that several of these annotations have previously been directly associated with flavonoid-related biosynthesis in *Arabidopsis*.

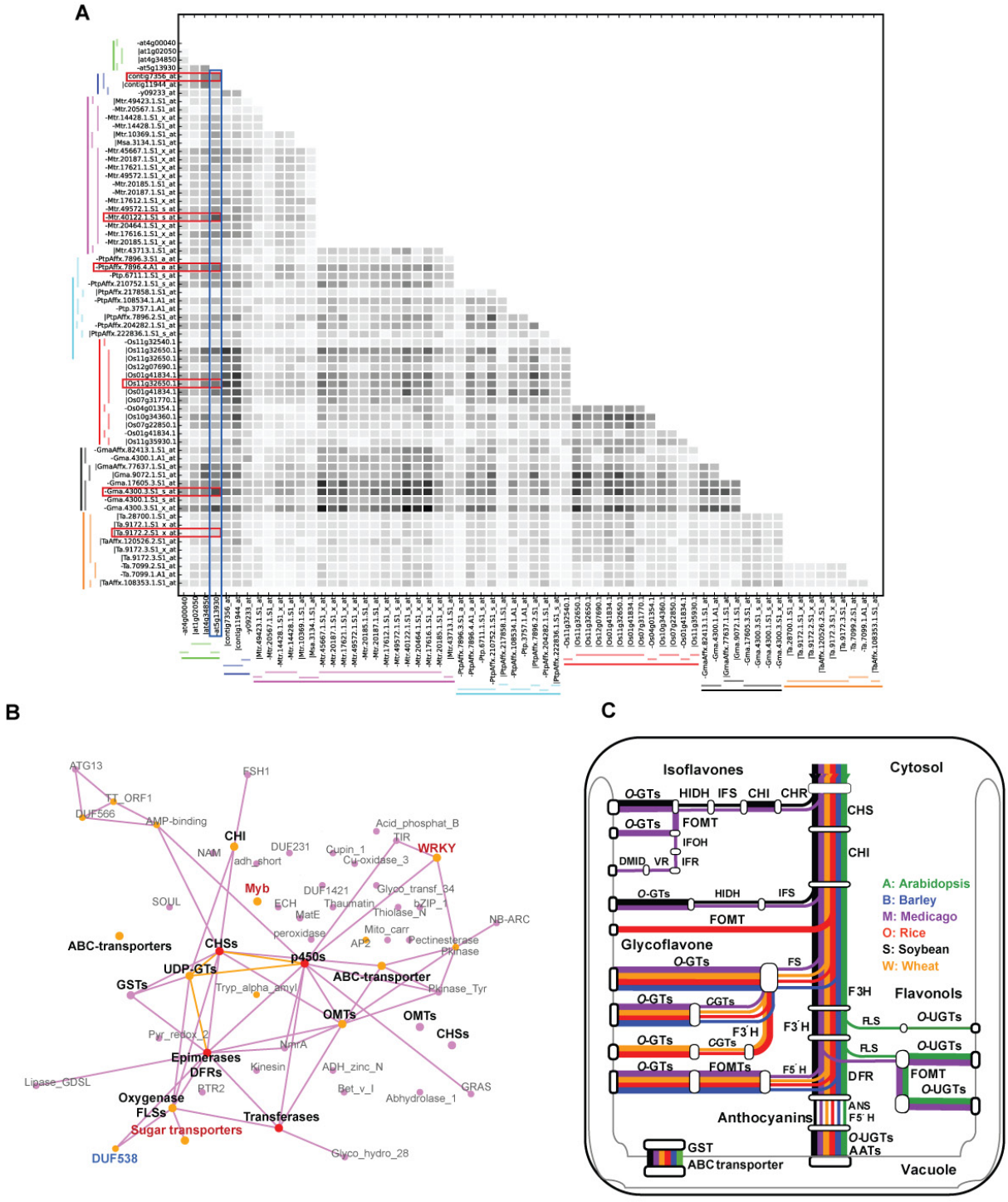

Figure 4.7. Combined co-expressed gene vicinities for CHS-related genes across seven species. A. *Heatmap depicting similarities of gene vicinity networks for the CHS family in seven species. CHS genes are represented by probeset IDs. Higher opacity of intersecting areas indicates higher gene VN similarity. Overlapping gene network vicinities are marked by alternating - and | sign and by alternating color bars. The column and rows containing VNs selected for the combined network are boxed in blue (query gene) and red (selected*

*VNs). **B.** Combined gene VNs for CHS-related genes across the seven species using CHS (At5g13930) as bait. Different colored nodes and edges correspond to number of species in which a homolog is found in the gene vicinity network (see explanation in Figure 4.4). Genes that directly correspond to pathway members are highlighted in black. Genes that are functionally related to the pathways are indicated in red. The blue gene indicates a DUF538 containing gene of unknown function for which homologs occur in the CHS gene network vicinity in at least three species. **C.** Schematic pathway structure of the anthocyanin/flavone biosynthesis in the different species. Different colors correspond to the different species as indicated. See box I for acronym annotations.*

For example, over-expression of the MYB transcription factor *PAP1* resulted in accumulation of cyanidin and quercetin derivates, and led to the activation of genes associated with the anthocyanin production (Tohge et al., 2005). Figure 4.7C shows a schematic pathway outline of the conserved flavonoid biosynthesis pathway, including anthocyanin, flavonol, glycoflavone and isoflavone synthesis, based on literature survey and KEGG pathways, for *Arabidopsis* (Tohge et al., 2005, Tohge et al., 2007, Yonekura-Sakakibara et al., 2008), barley (Nørbaek et al., 2003; Brazier-Hicks et al., 2009; Klausen et al., 2010), Medicago (Nørbaek et al., 2003; Kowalska et al., 2007; Farag et al., 2008), rice (Han et al., 2009; Kim et al., 2009), soybean, (Steele et al., 1999; Choung et al., 2001; Latunde-Dada et al., 2001) and wheat (Ioset et al., 2007). Sub-classes of flavonoids, and anthocyanins have been detected and reported in all six plant species, but none of the flavonoid subfamilies flavonol, glycoflavone and isoflavone has been reported. By comparing Figures 4.7B and 4.7C it is clear that many of the enriched protein family annotations in Figure 4.7B are prominent in the flavonoid pathway structure.

To obtain further information about the specific genes in the different VNs for the *CHS*-related genes we looked at the respective gene pages. One prominent example is the *CHS*-related gene (*AtPKS-B*, At4g34850, Mizuuchi et al., 2008) in *Arabidopsis*. The VN for this gene contains some genes that could be associated with flavonoid-related processes (Figure 4.8), such as a dihydroflavonol reductase (At4g35420), a 4-coumarate CoA ligase (At1g52940), and a glycosyltransferase (At1g33430). However, the VN does not contain the charactarized flavonoid biosynthetis genes nor the flavonol arabinosyltransferase *F3AbT* (*UGT78D3*, At5g17030), which convert flower specific flavonol. In addition, flavonoid profiling of *tt4* mutant flowers showed that many flavonoids were not detected (Yonekura-Sakakibara et al., 2008).

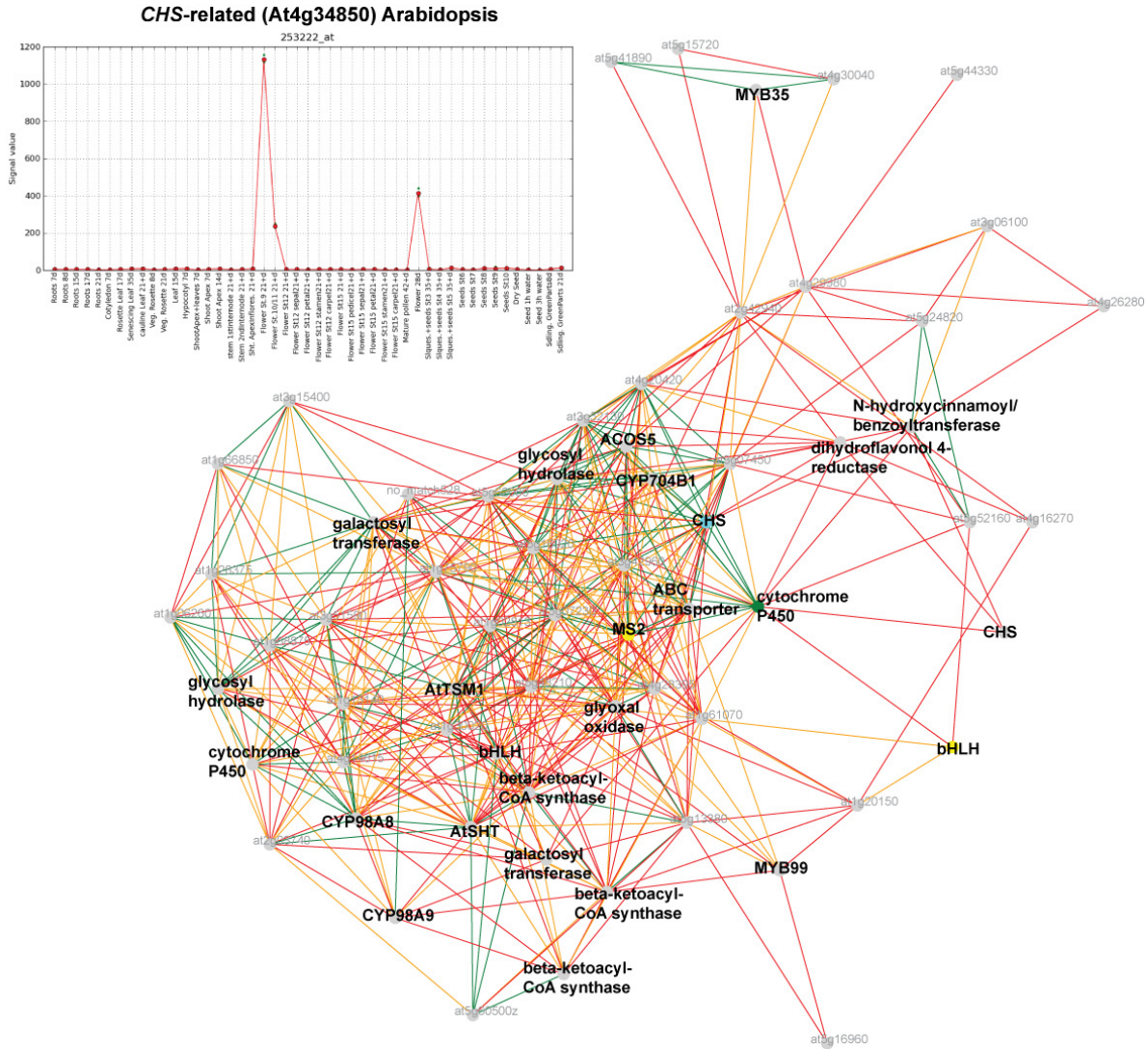

Figure 4.8 Co-expressed gene vicinity network for CHS-related gene (At4g34850) in Arabidopsis. *Nodes in the network resemble individual genes and the connecting edges represent co-expressed links. Expression of the CHS-related gene across different tissues is displayed upper left.*

These results indicate that the majority of flavonoids are produced via TT4 in *Arabidopsis* flowers. This in turn suggests that *AtPKS-B* is part of a different biosynthetic pathway. Indeed, Mizuuchi et al. (2008) concluded that *AtPKS-A* (At1g02050) and *AtPKS-B* could accept fatty acyl CoAs as a starting substrate. Furthermore, recent studies have shown that several of the genes in the VN of *AtPKS-B* participate in the synthesis of polyamines, such as N^1,N^5-di(hydroxyferuloyl)-N^{10}-sinapoylspermidine, being a part of the sporopollenin surrounding the pollen grains (Ehlting et al., 2008; Matsuno et al., 2009; Dobritsa et al., 2010). Several *MYB* and *bHLH* transcription factor encoding genes are also present in the *AtPKS-B* VN and may be good candidates for transcriptional regulators of the pathway

(Figure 4.8). Interestingly, Os07g22850 is a close rice homolog for AtPKS-B (Figure 4.5), making it a good candidate for related functions in rice. Indeed, this gene appears to be exclusively expressed in floral tissues and the VN contains genes that are associated with polyamine-related processes (http://aranet.mpimp-golm.mpg.de/ricenet/r48241).

Since many of the flavonoid-related processes have been relatively well characterized in *Arabidopsis* we inspected VNs from Medicago with high similarity to the *Arabidopsis TT4* VN. The highest scoring Medicago VN is associated with the probe ID Mtr.40122.1.s1_s_at (Figure 4.7A) and contains many genes with annotations related to isoflavone/flavonoid synthesis, including cytochrome P450s, *IFR*s, and *O*-methyltransferases (data not showed). Interestingly, another high-scoring VN surrounds the probe ID (Mtr.45667.1.s1_x_at), which is exclusively expressed in roots and root nodules (Figure 4.9). The flavonoid-derived metabolite medicarpin is a phytoalexin that is utilized by plant roots as protection from fungus and insects (Dakora et al., 1993; Dixon and Sumner, 2003; Naoumkina et al., 2007). Closer inspection of the VN revealed that many genes that encode proteins tentatively involved in the synthesis of medicarpin-conjugates were found co-expressed with the *TT4*-related Medicago gene. These genes represent gene products for virtually all the pathway steps from isoliquiritgenin to medicarpin and its downstream conjugates (Figure 4.9), and include *CHI*s, *IFS*s, *4O'MT*s, *HIDH*s, *I2'H*s, *IFR*s, *VR*s, *DMID*s, *UGT*s, *GST*s, transporters, *WRKY*s and *bHLH*s. While several of the proteins responsible for the catalysis of the constituent pathway steps have been identified or anticipated (Naoumkina et al., 2007) we propose that many of the genes associated with the VN may qualify as good candidates for biosynthetic and regulatory gene products for this pathway.

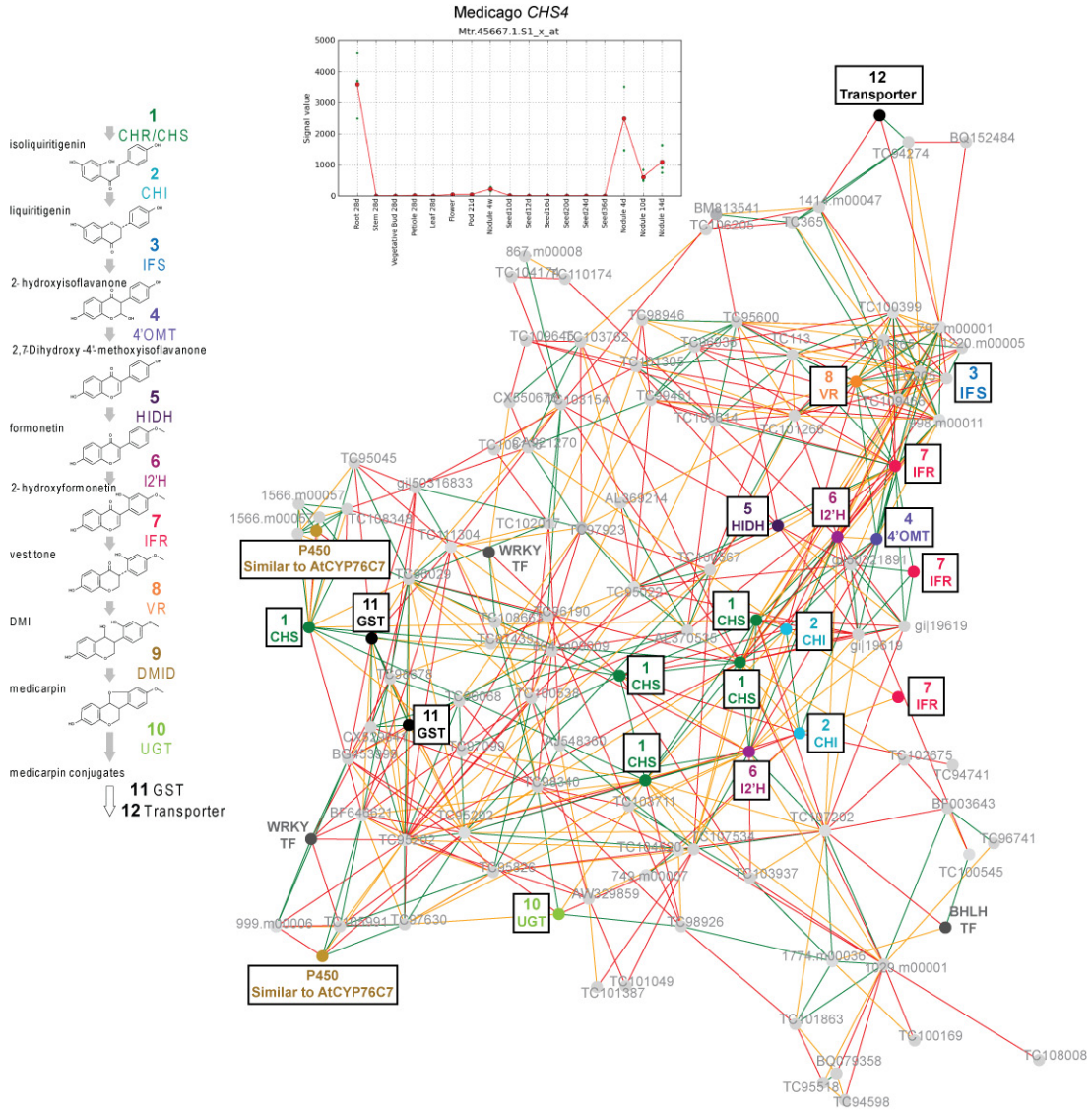

Figure 4.9. Co-expressed gene vicinity network for putative CHS-mediated (CHS4) Medicarpin biosynthesis. *Nodes in the network resemble individual genes and the connecting edges represent co-expressed links. The coloration of nodes and edges are explained in Figure 4.2. The different steps in the medicarpin biosynthetic pathway are indicated in different colors according to the pathway structure on the left. The P450-related genes indiacted in gold could encode the missing DMID step in the pathway. The interactive networks may be found at http://aranet.mpimp-golm.mpg.de/medinet/m40731. The expression profile for the CHS4 is indicated above the network. Acronyms stand for: CHS; Chalcone Synthase, CHR; Chalcone Reductase, CHI; Chalcone Isomerase, IFS; Isoflavone Synthase, 4'OMT; 4' O-methyltransferase, HIDH; 2-hydroxyisoflavanone dehydratase, I2'H;isoflavone 2'-hydroxylase, IFR; Isoflavone Reductase, VR; Vestitone reductase, DMID; 7,2 -dihydroxy-4'-methoxyisoflavanol dehydratase, UGT; UDP-Glycosyl transferase, GST; Glutathione S-transferase, WRKY and BHLH; Different types of transcription factors.*

4.5 MapMan ontology networks

Although co-expression analysis can suggest gene function and help unravel novel components of biological machineries, another important quest in biology is to understand how different biological functions are orchestrated to fulfill cellular processes. We have observed that genes involved in distinct yet related biological functions are often associated in co-expression networks. For example, ontology analysis of *PSA-D* genes used in one of the above case studies revealed that genes associated with photosystem I/II complexes, ATP synthesis, and Calvin cycle are co-expressed in Arabidopsis (http://aranet.mpimp-golm.mpg.de/aranet/a12253). To similarly evaluate co-occurence of ontology terms, we examined the HCCA clusters generated for the different organisms for MapMan ontology terms annotated to genes in the VN. Ontology terms showing a significant enrichment or depletion (Fisher test $p<0.05$) were then extracted. Subsequently, the co-occurrence of pairs of terms was determined for all clusters, and was tested for overrepresentation using a Fisher's exact test. Pairs of terms that were overrepresented were then connected, and the resulting networks were visualized as interactive networks for the seven species. Many ontology terms that were anticipated to be functionally connected also occured in close vicinity in the networks. For example, mitochondrial ATP synthesis/electron transport and TCA cycle related ontology terms, and terms such as photosynthesis, Calvin cycle, and tetrapyrrole synthesis are connected in the network for Medicago (http://aranet.mpimp-golm.mpg.de/medinet/ Mapman_network).

Similar to the comparative network approach for individual gene networks described above, we postulate that ontology terms that are connected in two or multiple species may more reliably reflect noteworthy links between the terms. To produce such network we identified terms which were associated in at least two monocots and two dictots. The resulting network, therefore, represents conserved ontology term associations across at least four of the plant species (Fig 4.10). Visual inspection of the network reveals that related processes are readily connected, and often form clusters. For example, photosynthesis related terms are associated with terms such as Calvin cycle, and tetrapyrrole biosynthesis, but is also associated with glycolysis, TCA cycle, and various mitochondrial processes. The latter associations could be viewed as reflecting the cross-talk between the chloroplast and mitochondria, for example in the form of different redox-related metabolites, such as malate and oxaloacetate, and in the exchange of ATP (Sweetlove et al., 2006; Raghavendra and Parmasree, 2003; Nunes-Nesi et al., 2008). In addition, terms associated with cell division,

e.g. various cell cycle related terms, histone biosynthesis, and chromatin structure, are closely linked to various vesicle trafficking terms, e.g. p- and v-ATPases, G-protein signalling, dynamins (Figure 4.10). Finally, flavonoid synthesis-related terms are directly connected to several stress related terms, which in turn connects to cytochrome P450s, GSTs, peroxidises and to jasmonate biosynthesis. The latter relationship is substantiated by the induction of anthocyanin production and of flavonoid-related genes by methyl jasmonates (Franceschi and Grimes, 1991). Hence, the ontology associations captured in the networks recapitulate known biological connections, and may therefore also be used as a guide to discover and establish new relationships between different biological pathways and functions.

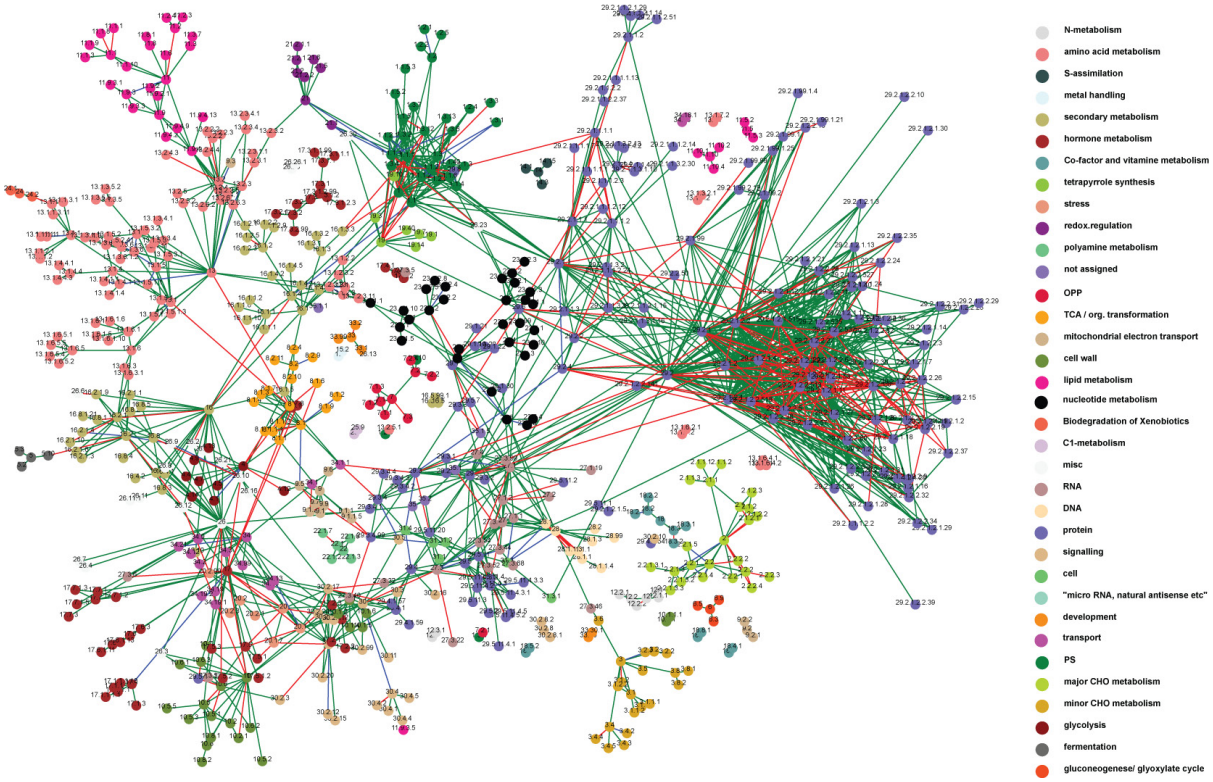

Figure 4.10. Combined MapMan ontology network for the seven species. *Co-expressed gene MapMan ontology terms in at least four of the investigated species. Key to the MapMan ontology is displayed on the right.*

4.6 Summary and future prospects.

The PlaNet platform integrates transcriptomic, genomic, phenomic and ontology terms for seven plant species with the aim to rapidly transfer knowledge across the species. Current functional homology predictions rely heavily on sequence comparisons, i.e. phylogentic relationships. While such inference in some instances may be accurate we contest that the

combination of sequence information with transcriptional relationship is likely to considerably improve such inferences. We have exemplified this using genes involved in assembly of the photosystems and in flavonoid-related processes. These analyses revealed many gene products and whole pathways that we predict would be closely linked to the respective pathways but currently are uncharacterized. Zooming out from an individual gene level we took advantage of co-expression between ontology terms and provided networks displaying how different terms are transcriptionally linked. We propose that the combination of the tools presented here will allow researchers to predict genes involved in highly diverse pathways and processes across diverse species, and to contextualize biological processes by ontology term associations. This tool may readily be extended to also include other species, provided that sufficient expression datasets exist. Inclusion of even more plant species may also allow for more detailed analyses, for example a more detailed assessments of transcriptional differences between monocots and dicots. That said even in its current format PlaNet should represent a highly useful resource for the many groups currently attempting to transfer knowledge gleaned from Arabidopis to species more vital for human circumstance.

5. General discussion

In the introductory chapter, I outlined the topics and tools used in this thesis, with the main focus set on co-expression analysis. Co-expression is based on the observation that transcriptionally coordinated genes tend to be functionally related. As stated in Chapter 1, a bioinformatic method is as useful as its accessibility, and quite commonly many prospective bioinformatical methods are either not implemented at all, or require considerable background in programming, e.g. R or Perl. A major objective for this thesis has therefore been to create bioinformatic-based tools that relatively easily can address biological problems.

Despite the great potential of co-expression analysis, it is important to keep in mind that correlation does not prove causal relationships. It only provides useful leads for establishing hypotheses, which may lay as ground for experimental procedures. While a potent tool, co-expression analysis has also several caveats:

(a) Low temporal resolution of gene expression contributes to false negatives. Some genes are only expressed in a specific tissue or stimuli, which may or may not be covered by available microarray data. If the specific condition is not covered by the arrays the genes would be tagged as not expressed, and would either show no correlation to any gene, or show correlation to other "not expressed" genes.

(b) False positives can arise due to the fact that samples used for microarray study often contain whole organs. For example, if one uses a microarray dataset containing expression profile of whole flowers, excluding specific data for petals, sepals, stamens and carpels, genes from these separate tissues would appear strongly co-expressed, even when this is not the case.

(c) At least 40 high-quality arrays are needed to conduct a sound co-expression analysis (Manfield et al., 2006). The ATH1 microarrays that fulfil those criteria for *Arabidopsis* cover approximately 63% of predicted genes (~21000 genes on microarray, with ~30000 genes in *Arabidopsis* genome). It is therefore clear that certain transcriptional relationships are not revealed using microarrays, resulting in further false negatives.

These caveats should prompt caution by biologists in over-reliance, or at least over-interpretation, of “whole-genome” expression analyses. I attempted to address this issue by introducing comparative transcriptomics for seven plant species. Comparative analysis could, in response to the above points, improve: a) poor temporal transcriptomic resolution of one organism, using a rich resolution of another, b) separate bulky, whole tissue transcriptomic snapshots, using fine grained measurements of another, and c) suggest missing co-expression relationships in one species by using discovered relationships from another species. Indeed, several studies have indicated that comparative transcriptomic analyses can improve co-expression analysis, by highlighting conserved co-expression relationships, e.g. Oti et al., 2008)

At the end of Chapter 1, I stated that co-expression analysis is a potent tool that can be used for prediction of: (i) gene function, (ii) organization of biological processes and (iii) functional homologs. With those three problems in mind, the web-tools GeneCAT, AraNet and PlaNet were created. This chapter briefly summarizes the three web-tools and output from them. This chapter also discusses how comparative co-expression analysis can be applied to improve co-expression analysis, and also presents an outlook for future work.

5.1 Prediction of gene function

The lack of knowledge of gene function is still the major bottleneck in understanding how a complex system, such as the cell, functions. *Arabidopsis*, as the most thoroughly investigated plant in science, has ~50% genes without any functional annotation. Those genes do not share sequence similarity with annotated genes, and are often described as "expressed protein" or "protein of unknown function". Functional annotation can be suggested by basic co-expression analysis, as suggested in chapter 2.1 of this thesis. Examples from literature that used co-expression analysis as predictor of gene function in *Arabidopsis* include Horan *et al.* (2008), who associated 1,541 proteins of unknown function (PUFs) with clusters containing proteins with known function (PKFs), thus suggesting putative annotations for the PUFs. Similar analysis was performed in Mutwil et al., (2009, not presented in this thesis), where I investigated association of cell wall related gene families with all families present in the genome. I have observed logical associations between families involved in cell wall biosynthesis and also with families not associated with cell wall biosynthesis, forming a potential basis for future investigations.

5.2 Organization of biological processes

Any biological process or pathway, for example photosynthesis, requires coordinated effort of many gene products. While uncovering the components of a pathway is essential for understanding the underlying mechanism, it is equally vital to understand how the different biological processes are associated with one another. For example the biological process of cell division could be roughly divided into biological sub-processes of DNA synthesis, DNA methylation, histone biosynthesis, cell cycle control, cell wall biosynthesis and others.

By combining Mapman ontology analysis with co-expression, I have investigated the association of hierarchies of biological processes, as shown in Chapter 3 and 4. The results presented in those chapters revealed known, and logical associations, e.g. ontologies constituting photosynthesis related processes were strongly connected with one another, but also with plastid protein synthesis and tetra-pyrrole biosynthesis (Figure 4.9). Another interesting associations were found between vesicle trafficking, G-proteins, post-translational modification of proteins and proteasome, possibly reflecting a pathway for tagging, transportation and degradation of proteins

Several unknown, but interesting associations were also found in the Mapman ontology association network. For example, plant defensins were connected to sphingolipid biosynthesis. As the mode of action of plant defensins often seems to be mediated by sphingolipids of the attacking pathogen (Thevissen et al., 2000; 2005; Ramamoorthy et al., 2009), it could be speculated, that plant sphingolipids might play a role in this mechanism as well.

Thus, the combination of ontology associations and co-expression analysis can capture known biological connections, and may therefore also be used as a guide to discover and establish new relationships between different biological pathways and functions.

5.3 Prediction of functional homologs.

Arabidopsis is characterized by short generation time, large seed production, convenient size and a small, fully sequenced genome, but with minor economical value, making it a good basis for studying dicots. Basic research done in *Arabidopsis* assumes that many biological pathways are also conserved in other plants, important for the society. While this assumption probably is true in many cases, identifying the corresponding pathways in economically valuable plant is a non-trivial task, largely due to large gene families generally found in plants. As a consequence, sequence comparison of a gene from the knowledge donor to the

genome of the acceptor can return a large list of possible candidate genes, as discussed in chapter 4.4.

Intuitively, a functional homolog should be present when the relevant biological process occurs. Thus, a functional homolog may be identified by combined sequence and co-expression approaches. Several studies have observed that biological pathways are conserved across different species (Zhou et al., 2004, Stuart et al., 2003, Li et al., 2009). However, those studies investigated conservations across highly divergent species such as yeast, mouse and human, and most discoveries belong to already well characterized, primordial pathways, such as protein synthesis. Thus, a study comparing less divergent species would constitute a more powerful analysis.

In Chapter 4.4, I have introduced NetworkComparer, a pipeline that can identify and display the most similar co-expression network regions in three monocots (barley, rice, wheat) and four dicots (*Arabidopsis*, medicago, poplar and soybean). As mentioned above, an advantage of such approach is that comparative transcriptomic analyses can highlight conserved co-expression relationships. Importantly, the pipeline both suggests the corresponding pathways, and also reveals the *identity* of the associated transcripts.

Another interesting observation during the investigation of co-expression networks relating to flavonoid associated pathways was made in Chapter 4.4. The analysis revealed two distinct network regions, a ubiquitously expressed flavonoid associated pathway, and a flower specific polyamine pathway. Thus, while the two pathways synthesize different compounds, they employ similar molecular machinery. While it is not clear from the analysis whether those two pathways evolved via convergence or divergence, this finding suggests that not only genomic material can be subjected to copy/pasting events (e.g. gene or genome duplication), but whole pathways. Another clear example of such duplication of biological pathways can be observed for cellulose biosynthesis, where two distinct co-expression network regions are associated with synthesizing primary and secondary cell walls (discussed in Chapter 2).

5.4 Future work

5.4.1 Improved algorithm for comparing network structures

While NetworkComparer pipeline can be effectively used to predict functional homologs and identify duplicated co-expression network regions, it could be improved in two aspects: (i)

the comparative algorithm does not return any value indicating the significance of similarity between co-expressed vicinities, and (ii) the results are not pre-calculated, resulting in relatively slow analysis.

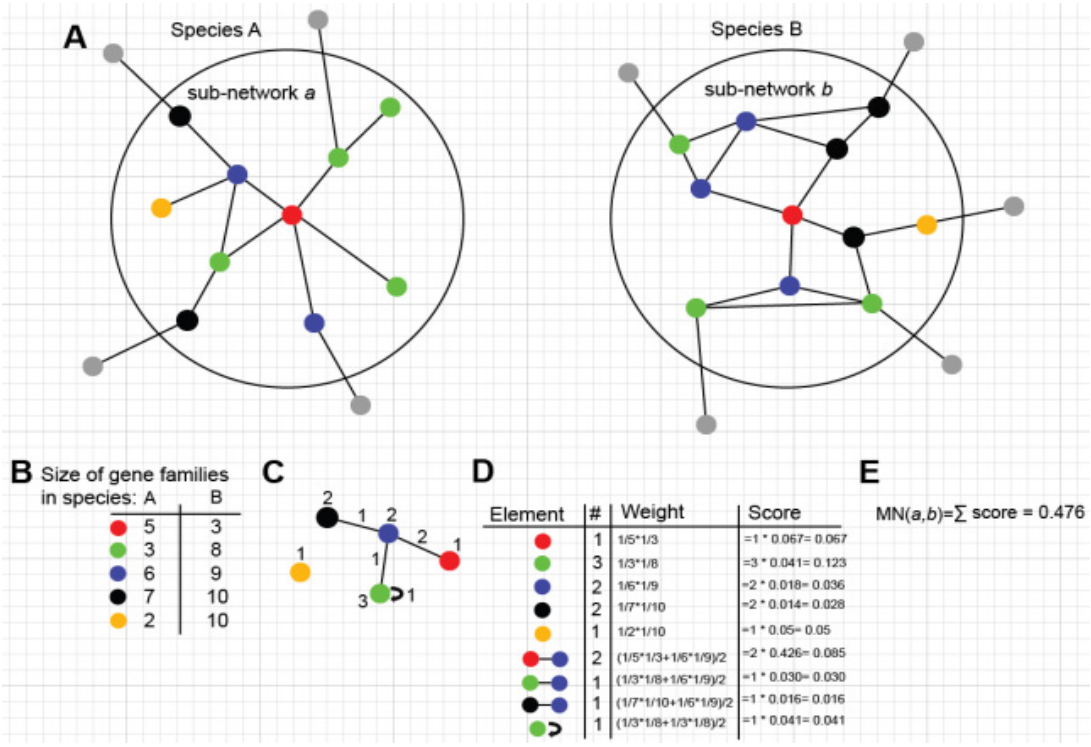

Figure 5.1. Workflow of "Mutwil-Nikoloski" algorithm with example of comparing two sub-networks from two species. ***A)*** *Nodes represent genes, node colors represent different gene families, while edges depict significant co-regulations between genes. Sub-networks are generated by taking 2 steps out away from the seed node (red). Gray nodes represent the remaining parts of network that are outside of the sub-network and thus not compared.* ***B)*** *Table of gene family sizes in species A and B.* ***C)*** *Minimal common network of sub-network a and b. For example, there are 2 and 3 black nodes in sub-networks a and b, so there are minimum 2 black nodes in the minimal common network. There are 2 red-blue edges in sub-network a, and 3 red blue-edges in sub-network b, therefore there are 2 red-blue edges in the minimal common network.* ***D)*** *Scoring the nodes and edges present in minimal common network. Weight of nodes is a product of reciprocal values of gene family sizes in species A and B. Weight of edges is an average of weight of nodes connected by the edge. Score of an element is calculated by multiplying weight of an element with its occurrence in minimal common network.* ***E)*** *Finally, the similarity score between sub-networks a and b is a sum of scores of elements found in D).*

To remedy those issues, Zoran Nikoloski and I have developed an algorithm that can return a p-value of a comparison, by shuffling gene family associations (Figure 5.1). The method consists of five steps: A) The algorithm accepts two sub-networks as query, and extracts a minimal common network (MN) from the sub-networks. B) Construction of MN. For example, there are two and three genes belonging to blue gene family in sub-network a and b, respectively, and therefore there are minimum two blue nodes shared between the sub-networks. There are two and three blue-red connections in sub-network a and b, respectively, resulting in two red-blue connections in MN. C) The nodes and edges of the MN are then weighted according to the sizes of the gene families found in respective species, and the final score is a sum of products of inversed sizes of gene families.

An advantage of this algorithm is that it weights co-presence of the different gene families according to the size of the families. In addition, the algorithm also interrogates the structure of the network by analyzing the connections between nodes.

5.4.2 Further comparative analyses

The accumulation of microarray data permits construction of high quality co-expression networks for many more organisms than has analyzed by any study. For example, during creation of PlaNet, microarray datasets comprising different tissues from economically important tobacco and tomato were released (Edwards et al. 2009, Ozaki et al., 2010). PlaNet and its tools are easily expandable and additional species will soon be included, permitting a comprehensive comparison of the different species. Pertinent information about gene functions also comes from available phenotype data, as shown in Chapter 2. Most genes in yeast and E.coli. have been characterized by knock-out analyses, and this information could be readily transferred to other species. Table 5.1 shows one row from NetworkComparer analysis for At1g53850 (20S proteasome alpha subunit E1) from *Arabidopsis* in which yeast microarray datasets were included. The analysis detected highly similar networks in all seven plant species and yeast. The table below shows one row for Rpn3 family, which is associated with the 20S subunit. Interestingly, functional homologs from *Arabidopsis* and yeast show similar phenotypes. While yeast and Arabidopsis have completely different anatomies, I argue that this type of comparative analyses, i.e. inclusion of sequence similarity, co-expression vicinity and phenomics, may be very powerful when comparing for example species fungal, mammalian and plant species.

pfam ID	Description	*Arabidopsis*	Barley	Medicago	Rice	Poplar	Yeast	Soy bean	Wheat
Rpn3_C	Proteasome regulatory subunit C-terminal	at1g20200 **embryo defective**	contig 6657_at	N/A	N/A	PtpAffx. 163275.2	YER021W **Phenotype: inviable**	N/A	Ta.23012.1

Table 5.1 NetworkComparer analysis of At1g20200 (regulatory subunit of proteasome) across the seven plant species and yeast.

Additional species will be added to the PlaNet platform, and features, such as ontology analysis and available phenotypic data, will be integrated into the database. This study will hopefully reveal many previously unknown functional pathway and gene function homologies, and uncover duplicated network regions.

5.4.3 Transcriptomic associations between gene families.

Association of ontology terms presented in this study can also be applied to gene families. Previous comparative studies have largely focused on functional predictions of single genes. An equally worthwhile quest is to gain more general knowledge about gene families. Certain gene families are specifically associated with certain biological processes, while others are not. For example, it is unlikely that photosystem I subunit protein is a component of any other process than photosynthesis. On the other hand, two protein kinases from same family can be involved in different biological processess and phosphorylate vastly different substrates.

To investigate the transcriptional association on a gene family level, one could apply the same calculation as described in Chapters 3 and 4, but replace gene ontologies with gene families. Indeed, premilinary results indicate many logical associations between different gene families (Figure 5.2)

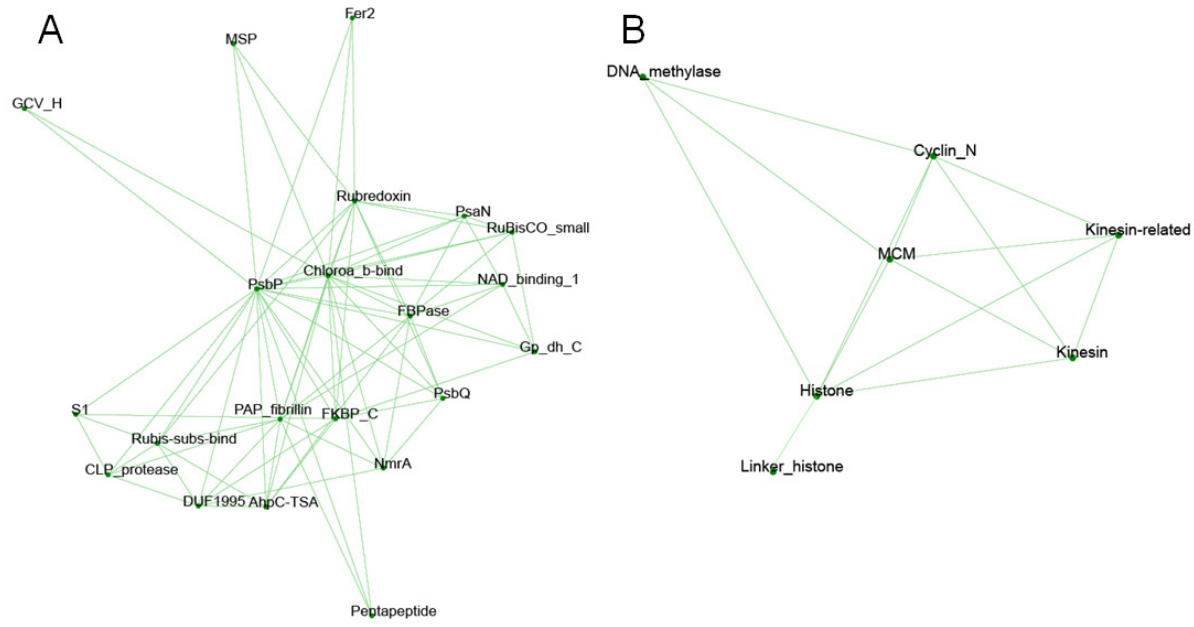

Figure 5.2 Associations between pfam families in Arabidopsis, barley, Medicago, poplar and rice. *Any two families are connected if they are co-regulated with one another (p-value>0.05) in at least two monocots and two dictos. A. Family association network centered around PsbP. B. Family association network centered around Cyclin_N.*

Figure 5.2A is centered around *PsbP*, which gene product participate in the photosystem II complex. The network contains other subunits of photosystem I and II, but also several unexpected proteins, e.g. DUF1995. Figure 5.2B is centered around the *Cyclin_N* family, which is co-regulated with families broadly associated with cell division. While these results are preliminary, they suggest that transcriptional association can be used to predict general function for a gene family.

5.5 Conclusion

The search for alternative methods for investigating biological phenomena can provide new perspectives and discoveries. Work presented in this thesis focus on three problems: prediction of gene function, understanding organization of biological processes and finding functional homologs between species.

Several available co-expression tools for plants use transcriptional coordination of genes to prioritize genes associated with a specific biological function. In this thesis, I have demonstrated how co-expression analysis can transcend the standard applications. However, co-expression analysis suffers from incomplete genomic representation of genes on available

microarrays, and poor spatio-temporal resolution of measured gene expression. Several studies have indicated that comparative transcriptomic analyses across species can improve co-expression analysis. GeneCAT was the first plant web-tool to introduce comparative transcriptomics analyses for Arabidopsis and Barley.

Co-expression analysis can be further augmented by including additional information. AraNet was first web-tool in plant community to combine available phenotypic data with co-expression networks, permitting a phenotypic prediction of gene knock-outs. For comprehensive visualization of genome scale co-expression networks, I developed a novel, efficient Heuristic Cluster Chiseling Algorithm. To investigate the transcriptional wiring of biological processes, I also investigated coordinated expression of ontological processes. Finally, I confirmed the predictive power of AraNet by identifying seven previously uncharacterized genes as being essential for plant growth.

To augment AraNet platform with comparative transcriptomics, six additional plants, Barley, Medicago, Poplar, Rice, Wheat and Soybean were included under the banner PlaNet. I constructed a NetworkComparer pipeline that permits comparative analyses across the plants, and that predict conserved regulatory relationships. Importantly, this pipeline returns the identity of functional homologs which is essential for transfering biological knowledge from a model organism to organisms important for society.

These webtools are user-friendly and are freely available online.

Publications

2010

PlaNet: Combined sequence and expression comparisons across seven plant species. Mutwil M, Tohge T, Giorgi F, Fernie A, Usadel B, Persson S (under review).

Large-Scale Co-expression Approaches to Dissect Secondary Cell Wall Formation across Plant Species. Ruprecht C, Mutwil M, Saxe F, Eder M, Nikoloski Z, Persson S (under review)

Assembly of an interactive correlation network for the *Arabidopsis* genome using a novel heuristic clustering algorithm. Mutwil M, Usadel B, Schütte M, Loraine A, Ebenhöh O, Persson S. Plant Physiol. 2010 Jan;152(1):29-43.

Analyzing Gene Coexpression Data by an Evolutionary Model. M. Schütte, M. Mutwil, S. Persson and O. Ebenhöh . Genome Informatics (2010)

2009

Transcriptional wiring of cell wall-related genes in *Arabidopsis*. Mutwil M, Ruprecht C, Giorgi FM, Bringmann M, Usadel B, Persson S. Mol Plant. 2009 Sep;2(5):1015-24. Epub 2009 Jul 30.

Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. Usadel B, Obayashi T, Mutwil M, Giorgi FM, Bassel GW, Tanimoto M, Chow A, Steinhauser D, Persson S, Provart NJ. Plant Cell Environ. 2009 Dec;32(12):1633-51. Epub 2009 Aug 27. Review.

2008

Laying down the bricks: logistic aspects of cell wall biosynthesis. Geisler DA, Sampathkumar A, Mutwil M, Persson S. Curr Opin Plant Biol. 2008 Dec;11(6):647-52. Epub 2008 Sep 23.

Functional analysis of the cellulose synthase-like genes CSLD1, CSLD2, and CSLD4 in tip-growing *Arabidopsis* cells. Bernal AJ, Yoo CM, Mutwil M, Jensen JK, Hou G, Blaukopf C, Sørensen I, Blancaflor EB, Scheller HV, Willats WG. Plant Physiol. 2008 Nov;148(3):1238-53. Epub 2008 Sep 3.

Cellulose synthesis: a complex complex. Mutwil M, Debolt S, Persson S. Curr Opin Plant Biol. 2008 Jun;11(3):252-7.

GeneCAT--novel webtools that combine BLAST and co-expression analyses. Mutwil M, Obro J, Willats WG, Persson S. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W320-6.

Curriculum Vitae

Education

2007-2010 PhD student, AG Persson, University of Potsdam, Max-Planck Institute for Molecular Plant Physiology, Germany. Thesis topic: “Integrative transcriptomic approaches to analyzing plant co-expression networks”

2005-2007 M.Sc Biochemistry, University of Copenhagen, Denmark. Thesis topic: “Analysis of Arabidopsis thalian co-expression networks.”

2003-2005 B. Sc. Biochemistry, University of Copenhagen, Denmark. Thesis topic: “Functional analysis of the cellulose synthase-like genes CSLD1, CSLD2, and CSLD4 in tip-growing Arabidopsis cells."

Teaching activities

2010 Seminar supervisor in lecture series "Systems Biology and Mathematical Modeling"

Selected seminars

2008 First progress seminar, Max-Planck Institute for Molecular Plant Physiology, Golm, Germany.

2009 Second progress seminar, Max-Planck Institute for Molecular Plant Physiology, Golm, Germany.

2010 Third progress seminar, Max-Planck Institute for Molecular Plant Physiology, Golm, Germany.

2010 "Differences and similarities of transcriptional programs of cell wall biosynthetic genes across 7 plant species". XII Cell Wall Meeting, Porto, Portugal

Selbständigkeitserklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbständig und unter Verwendung keiner anderen als den von mir angegebenen Quellen und Hilfsmitteln verfasst habe.

Ferner erkläre ich, dass ich bisher weder an der Universität Potsdam noch anderweitig versucht habe, eine Dissertation einzureichen oder mich einer Doktorprüfung zu unterziehen.

Potsdam, 25 July 2010

_____

Marek Mutwil

Bibliography

Abe I, Morita H. (2010) Structure and function of the chalcone synthase superfamily of plant type III polyketide synthases. Nat Prod Rep. 27: 809-38

Albert R (2005) Scale-free networks in cell biology. J Cell Sci 118: 4947-4957.

Alonso J, Stepanova A, Leisse T, Kim C, Chen H, Shinn P, Stevenson D, Zimmerman J, Barajas P, Cheuk R, et al (2003) Genome-wide insertional mutagenesis of Arabidopsis thaliana. Science 301: 653-657.

Aoki K, Ogata Y, Shibata D (2007) Approaches for extracting practical information from gene co-expression networks in plant biology. Plant Cell Physiol 48: 381-390.

Arioli T, Peng L, Betzner AS, Burn J, Wittke W, Herth W, Camilleri C, Hofte H, Plazinski J,

Birch R, Cork A, Glover J, Redmond J, Williamson RE. (1998) Molecular analysis of cellulose biosynthesis in Arabidopsis. Science. 279: 717-720

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K,

Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC,

Richardson JE, Ringwald M, Rubin GM, Sherlock G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. May;25(1):25-9.

Austin MB, Noel JP. (2003) The chalcone synthase superfamily of type III polyketide synthases. Nat Prod Rep. 20: 79-110

Bader GD, Hogue CW (2003) An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics 13;4:2

Baerenfaller K, Grossmann J, Grobei MA, Hull R, Hirsch-Hoffmann M, Yalovsky S,

Zimmermann P, Grossniklaus U, Gruissem W, Baginsky S (2008) Genome-scale proteomics reveals Arabidopsis thaliana gene models and proteome dynamics. Science 320: 938-941.

Barabási AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5: 101-113.

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

Berthomé R, Thomasset M, Maene M, Bourgeois N, Froger N, Budar F (2008) pur4 mutations are lethal to the male, but not the female, gametophyte and affect sporophyte development in Arabidopsis. Plant Physiol 147: 650-660.

Brazier-Hicks M, Evans KM, Gershater MC, Puschmann H, Steel PG, Edwards R. (2009) The C-glycosylation of flavonoids in cereals. J Biol Chem. 284: 17926-34

Brown DM, Zeef LA, Ellis J, Goodacre R, Turner SR. (2005) Identification of novel genes in Arabidopsis involved in secondary cell wall formation using expression profiling and reverse genetics. Plant Cell. 17: 2281-2295

Burton RA, Shirley NJ, King BJ, Harvey AJ, Fincher GB. (2004) The CesA gene family of barley. Quantitative analysis of transcripts reveals two groups of co-expressed genes. Plant Physiol. 134: 224-236

Carlson MR, Zhang B, Fang Z, Mischel PS, Horvath S, Nelson SF (2006) Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. BMC Genomics 7: 40-55.

Churchill GA (2002) Fundamentals of experimental design for cDNA microarrays.. Nat Genet. Dec;32 Suppl:490-5

Cooke R, Raynal M, Laudie M, Delseny M. (1997) Identification of members of gene families in Arabidopsis thaliana by contig construction from partial cDNA sequences: 106 genes encoding 50 cytoplasmic ribosomal proteins. Plant J. 11: 1127-1140

Dakora FD, Joseph CM, Phillips DA. (1993) Alfalfa (Medicago sativa L.) Root Exudates Contain Isoflavonoids in the Presence of Rhizobium meliloti. Plant Physiol. 101: 819-824

Daub CO, Steuer R, Selbig J, Kloska S (2004) Estimating mutual information using B-spline functions – an improved similarity measure for analysing gene expression data. BMC Bioinformatics 5: 118.

Davies DL, Bouldin DW (1979) A Cluster Separation Measure. IEEE Trans. Pattern Analysis and Machine Intelligence, 1: 224–227

DeRisi JL, Iyer VR, Brown PO (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 278: 680-686.

Desprez T, Juraniec M, Crowell EF, Jouy H, Pochylova Z, Parcy F, Hofte H, Gonneau M, Vernhettes S. (2007) Organization of cellulose synthase complexes involved in primary cell wall synthesis in Arabidopsis thaliana. Proc Natl Acad Sci U S A. 104: 15572-15577

Do JH, Choi DK. (2006) Normalization of microarray data: single-labeled and dual-labeled arrays. Mol Cells. Dec 31;22(3):254-61.

Dobritsa AA, Lei Z, Nishikawa S, Urbanczyk-Wochniak E, Huhman DV, Preuss D, Sumner LW. (2010) LAP5 and LAP6 encode anther-specific proteins with similarity to chalcone synthase essential for pollen exine development in Arabidopsis. Plant Physiol. 153: 937-55

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

Edgar R., Domrachev M. Lash A.E. (2002) Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Research 30, 207–210

Edwards KD, Bombarely A, Story GW, Allen F, Mueller LA, Coates SA, Jones L. (2010) TobEA: an atlas of tobacco gene expression from seed to senescence. BMC Genomics. Feb 26;11:142.

Ehlting J, Provart NJ, Werck-Reichhart D. (2006) Functional annotation of the Arabidopsis P450 superfamily based on large-scale co-expression analysis. Biochem Soc Trans. Dec;34(Pt 6):1192-8.

Ehlting J, Sauveplane V, Olry A, Ginglinger JF, Provart NJ, Werck-Reichhart D. (2008) An extensive (co-)expression analysis tool for the cytochrome P450 superfamily in Arabidopsis thaliana. BMC Plant Biol. 8: 47

Eisen MB, Spellman PT, Brown PO, Botstein D. (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A. Dec 8;95(25):14863-8.

Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 30: 1575-1584

Fagard M, Desnos T, Desprez T, Goubet F, Refregier G, Mouille G, McCann M, Rayon C, Vernhettes S, Hofte H. (2000) PROCUSTE1 encodes a cellulose synthase required for normal cell elongation specifically in roots and dark-grown hypocotyls of Arabidopsis. Plant Cell. 12: 2409-2424

Farag MA, Huhman DV, Dixon RA, Sumner LW (XXXX). Metabolomics reveals novel pathways and differential mechanistic and elicitor-specific responses in phenylpropanoid and isoflavonoid biosynthesis in Medicago truncatula cell cultures. Plant Physiol. 146: 387-402

Feinbaum RL, Ausubel FM. (1988) Transcriptional regulation of the Arabidopsis thaliana chalcone synthase gene. Mol Cell Biol. 8: 1985-92

Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer EL, Bateman A. (2008) The Pfam protein families database. Nucleic Acids Res. 36(Database issue): D281-8

Flores-Pérez U, Sauret-Güeto S, Gas E, Jarvis P, Rodríguez-Concepción M (2008) A mutant impaired in the production of plastome-encoded proteins uncovers a mechanism for the homeostasis of isoprenoid biosynthetic enzymes in Arabidopsis plastids. Plant Cell 20: 1303-1315

Franke R, Schreiber L. (2007) Suberin--a biopolyester forming apoplastic plant interfaces. Curr Opin Plant Biol. 10: 252-259

Fredslund J (2006): PHY.FI: fast and easy online creation and manipulation of phylogeny color figures. BMC Bioinformatics. 22;7:315.

Freeman TC, Goldovsky L, Brosch M, van Dongen S, Maziere P, Grocock RJ, Freilich S, Thornton J, Enright AJ (2007) Construction, Visualization, and Clustering of Transcription Networks from Microarray Expression Data. PLoS Comput Biol 3: 2032-2042.

Gachon CM, Langlois-Meurinne M, Henry Y, Saindrenan P (2005) Transcriptional co-regulation of secondary metabolism enzymes in Arabidopsis: functional and evolutionary implications. Plant Mol Biol. May;58(2):229-45.

Gandhi TK, Zhong J, Mathivanan S, Karthick L, Chandrika KN, Mohan SS, Sharma S, Pinkert S, Nagaraju S, Periaswamy B, et al. (2006) Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. Nat Genet 38: 285-293.

Geisler-Lee J, O'Toole N, Ammar R, Provart NJ, Millar AH, Geisler M. (2007) A predicted interactome for Arabidopsis. Plant Physiol. Oct;145(2):317-29

Gibbons FD, Roth FP (2002) Judging the Quality of Gene Expression-Based Clustering Methods Using Gene Annotation. Genome Res 12: 1574-1581.

Goda H, Sasaki E, Akiyama K, Maruyama-Nakashita A, Nakabayashi K, Li W, Ogawa M, Yamauchi Y, Preston J, Aoki K, Kiba T, Takatsuto S, Fujioka S, Asami T, Nakano T, Kato H, Mizuno T, Sakakibara H, Yamaguchi S, Nambara E, Kamiya Y, Takahashi H, Hirai MY, Sakurai T, Shinozaki K, Saito K, Yoshida S, Shimada Y. (2008) The AtGenExpress hormone and chemical treatment data set: experimental design, data evaluation, model data analysis and data access. Plant J. 2008 Aug;55(3):526-42.

Gregory BD, Belostotsky DA. (2009) Whole-genome microarrays: applications and technical issues. Methods Mol Biol.;553:39-56

Gupta PK, Rustgi S, Mir R. (2008) Array-based high-throughput DNA markers for crop improvement. Heredity. Jul;101(1):5-18

Hallegger M, Llorian M, Smith CW. (2010) Alternative splicing: global insights. FEBS J. Feb;277(4):856-66.

Han RM, Tian YX, Liu Y, Chen CH, Ai XC, Zhang JP, Skibsted LH. (2009) Comparison of flavonoids and isoflavonoids as antioxidants. J Agric Food Chem. 57: 3780-5

Hartigan JA, Wong MA (1979) A k-means clustering algorithm. Applied Statistics 28: 100–108.

Hegeman AD. (2010) Plant metabolomics--meeting the analytical challenges of comprehensive metabolite analysis. Brief Funct Genomics. Mar;9(2):139-48.

Hirai MY, Sugiyama K, Sawada Y, Tohge T, Obayashi T, Suzuki A, Araki R, Sakurai N, Suzuki H, Aoki K, Goda H, Nishizawa OI, Shibata D, Saito K. (2007) Omics-based identification of Arabidopsis Myb transcription factors regulating aliphatic glucosinolate biosynthesis. Proc Natl Acad Sci U S A. 104: 6478-6483

Hobo T, Asada M, Kowyama Y, Hattori T. (1999) ACGT-containing abscisic acid response element (ABRE) and coupling element 3 (CE3) are functionally equivalent. Plant J. Sep;19(6):679-89.

Horan K, Jang C, Bailey-Serres J, Mittler R, Shelton C, Harper JF, Zhu JK, Cushman JC, Gollery M, Girke T. (2008) Annotating genes of known and unknown function by large-scale coexpression analysis. Plant Physiol. May;147(1):41-57.

Hubert L, Arabie P (1985) Comparing partitions. J Classification 193–218.

Ihmels J, Levy R, Barkai N (2004) Principles of transcriptional control in the metabolic network of Saccharomyces cerevisiae. Nat Biotechnol 22: 86-92.

Ihnatowicz A, Pesaresi P, Varotto C, Richly E, Schneider A, Jahns P, Salamini F, Leister D. Mutants for photosystem I subunit D of Arabidopsis thaliana: effects on photosynthesis, photosystem I stability and expression of nuclear genes for chloroplast functions. Plant J. 37: 839-52

Ioset JR, Urbaniak B, Ndjoko-Ioset K, Wirth J, Martin F, Gruissem W, Hostettmann K, Sautter C. (2007) Flavonoid profiling among wild type and related GM wheat varieties. Plant Mol Biol. 65: 645-54

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. Apr;4(2):249-64.

Jeong H, Mason SP, Barabasi AL, Oltvai ZN (2001) Lethality and centrality in protein networks. Nature 411: 41-42.

Jupiter DC, VanBuren V (2008) A visual data mining tool that facilitates reconstruction of transcription regulatory networks. PLoS ONE 3: e1717-e1724.

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

Kim DH, Kim SK, Kim JH, Kim BG, Ahn JH. (2009) Molecular characterization of flavonoid malonyltransferase from Oryza sativa. Plant Physiol Biochem. 47: 991-7

King AD, Przulj N, Jurisica I (2004) Protein complex prediction via cost-based clustering. Bioinformatics 20: 3013-2020.

Kirschner MW. (2005) The meaning of systems biology. Cell. May 20;121(4):503-4.

Kitano H (2002) Systems biology: a brief overview. Science 295: 1662-1664

Klausen K, Mortensen AG, Laursen B, Haselmann KF, Jespersen BM, Fomsgaard IS. (2010) Phenolic compounds in different barley varieties: identification by tandem mass spectrometry (QStar) and NMR; quantification by liquid chromatography triple quadrupole-linear ion trap mass spectrometry (Q-Trap). Nat Prod Commun. 5: 407-14

Kowalska I, Stochmal A, Kapusta I, Janda B, Pizza C, Piacente S, Oleszek W. (2007) Flavonoids from barrel medic (Medicago truncatula) aerial parts. J Agric Food Chem. 55: 2645-52

Lamport DT, Kieliszewski MJ, Showalter AM (2006) Salt stress upregulates periplasmic arabinogalactan proteins: using salt stress to analyse AGP function. New Phytol 169: 479-492.

Latunde-Dada AO, Cabello-Hurtado F, Czittrich N, Didierjean L, Schopfer C, Hertkorn N, Werck-Reichhart D, Ebel J. (2001) Flavonoid 6-hydroxylase from soybean (Glycine max L.), a novel plant P-450 monooxygenase. J Biol Chem. 276: 1688-95

Lee PR, Cohen JE, Tendi EA, Farrer R, DE Vries GH, Becker KG, Fields RD. (2004) Transcriptional profiling in an MPNST-derived cell line and normal human Schwann cells. Neuron Glia Biol. May;1(2):135-147.

Li S, Armstrong CM, Bertin N, Ge H, Milstein S, Boxem M, Vidalain PO, Han JD, Chesneau A, Hao T et al. (2004) A map of the interactome network of the metazoan C. elegans. Science 303: 540-543.

Lievens S, Lemmens I, Tavernier J. (2009) Mammalian two-hybrids come of age. Trends Biochem Sci.

Liolios K, Chen IM, Mavromatis K, Tavernarakis N, Hugenholtz P, Markowitz VM, Kyrpides NC. (2010) The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res. 38(Database issue):D346-54

Ma S, Gong Q, Bohnert HJ (2007) An Arabidopsis gene network based on the graphical Gaussian model. Genome Res 17: 1614-1625.

Maier T, Güell M, Serrano L.(2009) Correlation of mRNA and protein in complex biological samples. FEBS Lett. Dec 17;583(24):3966-73

Manfield I.W., Jen C.-H., Pinney J.W., Michalopoulos I., Bradford J.R., Gilmartin P.M. Westhead D.R. (2006) Arabidopsis Co-expression Tool (ACT): web server tools for microarray-based gene expression analysis. Nucleic Acids Research 34, W504–W509

Manfield IW, Jen CH, Pinney JW, Michalopoulos I, Bradford JR, Gilmartin PM, Westhead DR. (2006) Arabidopsis Co-expression Tool (ACT): web server tools for microarray-based gene expression analysis. Nucleic Acids Res. 34(Web Server issue): W504-509

Mao L, Van Hemert JL, Dash S, Dickerson JA. (2009) Arabidopsis gene co-expression network and its functional modules. BMC Bioinformatics. 10:346

Marchler-Bauer A, Anderson JB, Derbyshire MK, DeWeese-Scott C, Gonzales NR, Gwadz M, Hao L, He S, Hurwitz DI, Jackson JD, Ke Z, Krylov D, Lanczycki CJ, Liebert CA, Liu C, Lu F, Lu S, Marchler GH, Mullokandov M, Song JS, Thanki N, Yamashita RA, Yin JJ, Zhang D, Bryant SH. (2007) CDD: a conserved domain database for interactive domain family analysis. Nucleic Acids Res. 35(Database issue): D237-40

Mariño-Ramírez L, Tharakaraman K, Bodenreider O, Spouge J, Landsman D. (2009) Identification of cis-regulatory elements in gene co-expression networks using A-GLAM. Methods Mol Biol.;541:1-22.

Matsuno M, Compagnon V, Schoch GA, Schmitt M, Debayle D, Bassard JE, Pollet B, Hehn A, Heintz D, Ullmann P, Lapierre C, Bernier F, Ehlting J, Werck-Reichhart D. (2009) Evolution of a novel phenolic pathway for pollen development. Science. 325: 1688-92

Mentzen WI, Wurtele ES (2008) Regulon organization of Arabidopsis. BMC Plant Biol 8: 99.

Millar AA, Gubler F (2005) The Arabidopsis GAMYB-like genes, MYB33 and MYB65, are microRNA-regulated genes that redundantly facilitate anther development. Plant Cell 17: 705-721.

Mizuuchi Y, Shimokawa Y, Wanibuchi K, Noguchi H, Abe I. (2008) Structure function analysis of novel type III polyketide synthases from Arabidopsis thaliana. Biol Pharm Bull. 31: 2205-10

Muraoka R, Okuda K, Kobayashi Y, Shikanai T. (2006) A eukaryotic factor required for accumulation of the chloroplast NAD(P)H dehydrogenase complex in Arabidopsis. Plant Physiol. 142: 1683-1689

Mutwil M, Obro J, Willats WG, Persson S (2008) GeneCAT--novel webtools that combine BLAST and co-expression analyses. Nucleic Acids Res (Webserver issue): W320-326.

Mutwil M, Ruprecht C, Giorgi FM, Bringmann M, Usadel B, Persson S (2009) Transcriptional Wiring of Cell Wall-Related Genes in Arabidopsis Mol. Plant 2(5):1015-1024

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

Nelson N, Yocum CF. (2006) Structure and function of photosystems I and II. Annu Rev Plant Biol. 57: 521-65

Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69: 026113

Nørbaek R, Aaboer DB, Bleeg IS, Christensen BT, Kondo T, Brandt K. (2003) Flavone C-glycoside, phenolic acid, and nitrogen contents in leaves of barley subject to organic fertilization treatments. J Agric Food Chem. 51: 809-13

Obayashi T, Hayashi S, Saeki M, Ohta H, Kinoshita K (2009) ATTED-II provides coexpressed gene networks for Arabidopsis. Nucleic Acids Res 37(Database issue): D987-991.

Obayashi T, Kinoshita K (2009,) Rank of correlation coefficient as a comparable measure for biological significance of gene co-expression. DNA Res. In press

Obayashi T, Kinoshita K, Nakai K, Shibaoka M, Hayashi S, Saeki M, Shibata D, Saito K, Ohta H. (2007) ATTED-II: a database of co-expressed genes and cis elements for identifying co-regulated gene groups in Arabidopsis. Nucleic Acids Res. 35 (Database issue): D863-9

Obayashi T., Hayashi S., Saeki M., Ohta H. & Kinoshita K. (2009) ATTED-II provides coexpressed gene networks for Arabidopsis. Nucleic Acids Research 37, D987–D991

Oti M, van Reeuwijk J, Huynen MA, Brunner HG. (2008) Conserved co-expression for candidate disease gene prioritization. BMC Bioinformatics.

Ozaki S, Ogata Y, Suda K, Kurabayashi A, Suzuki T, Yamamoto N, Iijima Y, Tsugane T, Fujii T, Konishi C, Inai S, Bunsupa S, Yamazaki M, Shibata D, Aoki K. (2010) Coexpression analysis of tomato genes and experimental verification of coordinated expression of genes found in a functionally enriched coexpression module. DNA Res. ;17(2):105-16. Epub 2010 Feb 3.

Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, Farne A, Holloway E, Kolesnykov N, Lilja P, Lukk M, Mani R, Rayner T, Sharma A, William E, Sarkans U, Brazma A. (2007) ArrayExpress--a public database of microarray experiments and gene expression profiles. Nucleic Acids Res. 35 (Database issue): D747-750

Persson S, Paredez A, Carroll A, Palsdottir H, Doblin M, Poindexter P, Khitrov N, Auer M, Somerville CR. (2007) Genetic evidence for three unique components in primary cell-wall cellulose synthase complexes in Arabidopsis. Proc Natl Acad Sci U S A. 104: 15566-15571

Persson S, Wei H, Milne J, Page GP, Somerville CR (2005) Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets. Proc Natl Acad Sci U S A 102: 8633-8638.

Premsler T, Zahedi RP, Lewandrowski U, Sickmann A. (2009) Recent advances in yeast organelle and membrane proteomics. Proteomics. Oct;9(20):4731-43.

Prieto C, Risueño A, Fontanillo C, De las Rivas J (2008) Human gene coexpression landscape: confident network derived from tissue transcriptomic profiles. PLoS One. 2008;3(12):e3911

Ramamoorthy V, Cahoon EB, Thokala M, Kaur J, Li J, Shah DM (2009) Sphingolipid C-9 methyltransferases are important for growth and virulence but not for sensitivity to antifungal plant defensins in Fusarium graminearum. Eukaryot Cell 8: 217-229.

Rhee S, Beavis W, Berardini T, Chen G,Dixon D, Doyle A ,Garcia-Hernandez M, Huala E, Lander G, Montoya M. (2003) The Arabidopsis Information Resource (TAIR) : a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. Nucleic Acids Res. 31: 224-228.

Rocca-Serra P., Brazma A., Parkinson H., et al. (2003) Arrayexpress: a public database of gene expression data at EBI. Current Research in Biology 326, 1075–1078

Saito K, Hirai MY, Yonekura-Sakakibara K. (2008) Decoding genes with coexpression networks and metabolomics - 'majority report by precogs'. Trends Plant Sci.13: 36-43

Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270: 467-470.

Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, Vingron M, Scholkopf B, Weigel D, Lohmann JU (2005): A gene expression map of Arabidopsis thaliana development. Nat Genet. 37(5): 501-6.

Shen L, Gong J, Caldo RA, Nettleton D, Cook D, Wise RP, Dickerson JA (2005): BarleyBase -- an expression profiling database for plant genomics. Nucleic Acids Res. 1;33 (Database issue): D614-8.

Sherlock G. (2000) Analysis of large-scale gene expression data. Curr Opin Immunol. Apr;12(2):201-5

Snel B, van Noort V, Huynen MA. (2004) Gene co-regulation is highly conserved in the evolution of eukaryotes and prokaryotes. Nucleic Acids Res. Sep 7;32(16):4725-31
Somerville C. (2006) Cellulose synthesis in higher plants. Annu Rev Cell Dev Biol. 22: 53-78

Srinivasasainagendra V, Page GP, Mehta T, Coulibaly I, Loraine AE (2008) CressExpress: a tool for large-scale mining of expression data from Arabidopsis. Plant Physiol 147: 1004-1016.

Staal FJ, van der Burg M, Wessels LF, Barendregt BH, Baert MR, van den Burg CM, van Huffel C, Langerak AW, van der Velden VH, Reinders MJ, van Dongen JJ. (2003) DNA microarrays for comparison of gene expression profiles between diagnosis and relapse n precursor-B acute lymphoblastic leukemia: choice of technique and purification influence the identification of potential diagnostic markers. Leukemia. Jul;17(7):1324-32

Steinhauser D, Usadel B, Luedemann A, Thimm O, Kopka J (2004) CSB.DB: a comprehensive systems-biology database. Bioinformatics 20: 3647-3651.

Steuer R, Humburg P, Selbig J (2006) Validation and functional annotation of expression-based clusters based on gene ontology. BMC Bioinformatics 7: 380-392.

Steuer R, Kurths J, Daub CO, Weise J, Selbig J. (2002) The mutual information: detecting and evaluating dependencies between variables. Bioinformatics;18 Suppl 2:S231-40.

Stracke R, Favory JJ, Gruber H, Bartelniewoehner L, Bartels S, Binkert M, Funk M, Weisshaar B, Ulm R. (2010) The Arabidopsis bZIP transcription factor HY5 regulates expression of the PFG1/MYB12 gene in response to light and ultraviolet-B radiation. Plant Cell Environ. 33: 88-103

Stuart JM, Segal E, Koller D, Kim SK (2003) A gene-coexpression network for global discovery of conserved genetic modules. Science 302: 249-255.

Sweetlove LJ, Fernie AR. (2005) Regulation of metabolic networks: understanding metabolic complexity in the systems biology era. New Phytol. Oct;168(1):9-24

Thalamuthu A, Mukhopadhyay I, Zheng X, Tseng GC. (2006) Evaluation and comparison of gene clustering methods in microarray analysis. Bioinformatics. Oct 1;22(19):2405-12

Thevissen K, Cammue BP, Lemaire K, Winderickx J, Dickson RC, Lester RL, Ferket KK, Van Even F, Parret AH, Broekaert WF (2000) A gene encoding a sphingolipid biosynthesis enzyme determines the sensitivity of Saccharomyces cerevisiae to an antifungal plant defensin from dahlia (Dahlia merckii). Proc Natl Acad Sci U S A 97: 9531-9536.

Thevissen K, Idkowiak-Baldys J, Im YJ, Takemoto J, François IE, Ferket KK, Aerts AM, Meert EM, Winderickx J, Roosen J, Cammue BP (2005) SKN1, a novel plant defensin-sensitivity gene in Saccharomyces cerevisiae, is implicated in sphingolipid biosynthesis. FEBS Lett 579: 1973-1977.

Tohge T, Fernie AR. (2010) Combining genetic diversity, informatics and metabolomics to facilitate annotation of plant gene function. Nat Protoc. 5: 1210-27

Tohge T, Nishiyama Y, Hirai MY, Yano M, Nakajima J, Awazuhara M, Inoue E, Takahashi H, Goodenowe DB, Kitayama M, Noji M, Yamazaki M, Saito K. (2005) Functional genomics by integrated analysis of metabolome and transcriptome of Arabidopsis plants over-expressing an MYB transcription factor. Plant J. 42: 218-35

Toufighi K, Brady SM, Austin R, Ly E, Provart NJ (2005) The Botany Array Resource: e-Northerns, Expression Angling, and promoter analyses. Plant J 43: 153-163.

Turner SR, Somerville CR. (1997) Collapsed xylem phenotype of Arabidopsis identifies mutants deficient in cellulose deposition in the secondary cell wall. Plant Cell. 9: 689-701

Usadel B, Nagel A, Steinhauser D, Gibon Y, Blasing OE, Redestig H, Sreenivasulu N, Krall L, Hannah MA, Poree F, et al (2006) PageMan: an interactive ontology tool to generate, display, and annotate overview graphs for profiling experiments. BMC Bioinformatics 18: 535-543.

Usadel B, Obayashi T, Mutwil M, Giorgi FM, Bassel GW, Tanimoto M, Chow A, Steinhauser D, Persson S, Provart NJ (2009) Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. Plant Cell Environ. 32:1633-1651
van Dongen S (2000) Graph Clustering by Flow Simulation. PhD thesis, University of Utrecht.

van Noort V, Snel B, Huynen MA (2004) The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. EMBO Rep 5: 280-284.

Wang T, Stormo GD. (2003) Combining phylogenetic data with co-regulated genes to identify regulatory motifs. Bioinformatics. Dec 12;19(18):2369-80.

Wasserman S, Faust K (1994) Social Network Analysis Ch, 12, 4 (Cambridge Univ. Press, Cambridge)

Wei H, Persson S, Mehta T, Srinivasasainagendra V, Chen L, Page GP, Somerville C, Loraine A (2006) Transcriptional coordination of the metabolic network in Arabidopsis. Plant Physiol 142: 762-774.

Yonekura-Sakakibara K, Tohge T, Matsuda F, Nakabayashi R, Takayama H, Niida R, Watanabe-Takahashi A, Inoue E, Saito K. (2008) Comprehensive flavonol profiling and transcriptome coexpression analysis leading to decoding gene-metabolite correlations in Arabidopsis. Plant Cell. 20: 2160-76

Yu H, Luscombe NM, Qian J, Gerstein M. (2003) Genomic analysis of gene expression relationships in transcriptional regulatory networks. Trends Genet. Aug;19(8):422-7.
Zhong R, Ye ZH (2007) Regulation of cell wall biosynthesis. Curr Opin Plant Biol 10: 564-572.

Zimmermann P, Hirsch-Hoffmann M, Hennig L, Gruissem W (2004) GENEVESTIGATOR. Arabidopsis microarray database and analysis toolbox. Plant Physiol 136: 2621-2632.

Zotenko E, Mestre J, O'Leary DP, Przytycka TM (2008) Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality. PLoS Comput Biol 4: e1000140.