Index

Abstract

The plant cell is a very coordinated system, with a central nucleus, organelles and structures, interacting continuously to perform the tasks needed to the cell growth, survival and replication. Genomes in fact store the instructions to define the cells, according to their genetic program. Most of cell activities are in fact the result of a work performed by multiple partners, collaborating in a serial or parallel way. All these efforts require hence a strong coordination among the gene functionalities too.

This work aims to investigate the gene expression, in particular, the genes which are co-expressed, their roles and the main bioinformatics approaches exploited to analyze them. In order to perform this study, we have exploited the genome of *Arabidopsis thaliana*, a plant species very well studied and with a genome sequence made available in 2000 [The Arabidopsis Genome Initiative, 2000]. Because of this and thanks to the huge amount of data derived for this plant, *Arabidopsis thaliana* is considered the reference organism in plant genomic and constantly exploited for comparative analyses with other species. *A. thaliana* was established as a model plant genome because of its easy management properties, such as a small diploid genome, the small size and the high seeds production through self-pollination. However this plant has shown a very complex genome structure, with a hard to count number of gene duplications and chromosomes shuffles. The gathering of information about the co-expressed genes has inevitably led us to challenge some important questions in bioinformatics, as the research of

a network identification method able to group genes with a biological meaning, and the evaluation of the statistical components exploited to define genes as co-expressed. This effort has hence required the comparison of the public available platforms concerning the co-expression issue. As result of these activities, we have also tried to produce a new network profiling approach, inspired by social networks and able to underline quickly the genes in relation inside the genome. The comparison of the platforms concerning the co-expression has instead revealed a not uniform behavior of the genes according to the co-expression identification methods, underlining the need of data mining analyses not only through a single and specific dataset, but trying to compare the highest number of resources available.

Acknowledgements

I would to thank my supervisor Maria Luisa Chiusano for her support and for the possibility offered to improve myself in her laboratory. A special thank to Chiara who has helped in all my projects from the first day, and to my friends Hamed, Luca and Valentino that have shared with me these years of Phd as I were living in a second family. I would like to thank the prof. Crescenzio Gallo for his partnership and support in this work and Gennaro for his patience and technical assistance. Thanks a lot to my friend Alessandra who has always brought cheerfulness in the lab. All the results and efforts of these years are dedicated to my family and to my girlfriend Paola who has shared with me, all the good, and bad moments of this period, always with her precious smile.

List of figures

List of tables

Introduction

1.1 Gene Expression

The plant cell is a biological system with the following main tasks: growing up, surviving and reproducing. All these activities are performed by regulating and coordinating the production of specific molecules, according to the information available inside the genes. The gene expression in fact is the process which translates the information of the gene sequence in a functional molecule, generally a protein, after different steps. Each gene consists in two antiparallel and complementary strands of DNA, each one having 5' and 3'ends defined according to their chemical structures. The first process needed for the gene expression is the transcription, which exploits a RNA polymerase (RNAP) to produce a RNA sequence (pre-mRNA) complementary to the DNA gene strand oriented in 3'-5' (the "template" strand) (Fig. 1). As consequence, the pre-mRNA has the same sequence of the 5'-3' oriented DNA gene strand (the "coding" strand), with the exception of having thymines replaced by uracils.

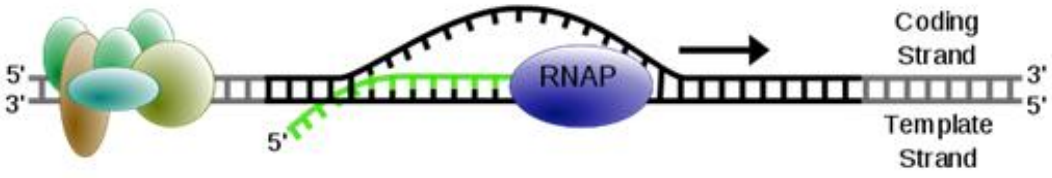


Figure 1 The green strand is a pre-mRNA sequence obtained by using the 3'-5' DNA strand as template.

There are three types of RNA polymerases in eukaryotes: RNA Polymerase I, II and III. Each one requires a special DNA sequence to start, named promoter, and a set of DNA-binding proteins called transcription factors. The RNA polymerase I is involved in the transcription of ribosomal RNA (rRNA) genes, while the RNA

polymerase II transcribes all protein-coding genes and some non-coding RNAs (snRNAs, snoRNAs) too. The RNA polymerase III produces 5S rRNA, transfer RNA (tRNA) genes, and some small non-coding RNAs. As seen for the start, the transcription process ends when the polymerase reaches a DNA sequence called terminator. In the eukaryotic cell the pre-mRNA produced during the transcription has to undergo several steps of maturation before becoming mature RNA (mRNA), the RNA sequence needed to guide the final gene product synthesis. First of all, in order to protect the final mRNA from exonucleases degradation, a 7-methylguanosine (m^7G) cap is added to the 5' end of the pre-mRNA. At the 3' end, instead, the pre-mRNA is cleaved close to a "polyadenilation signal sequence" (5'--AAUAAA-3'), located between the protein-coding sequence and terminator. Then, at 3' end, a tail of ~200 adenines (A), called poly(A) is added to protect RNA from degradation and improves the translation start. Another important maturation process is the RNA splicing: in the mRNA in fact, we defines two important regions: the exons, which remains present within the final mature RNA, and the introns that are not available in the final mRNA (Fig. 2). The splicing consists in the introns removal by a protein complex called splicesome. Although the transcription represents already an event of gene expression modulation, one gene can produces different mRNAs according to the number of exons kept in the final mRNA (alternative splicing). In fact, if one or more exons are settled between two introns, they can be removed too by the splicesome, producing different transcripts and hence different proteins.

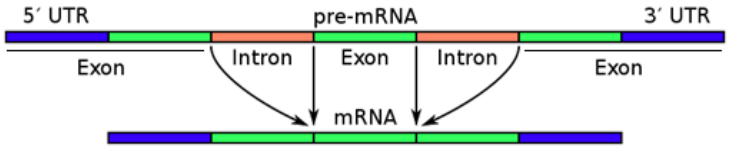
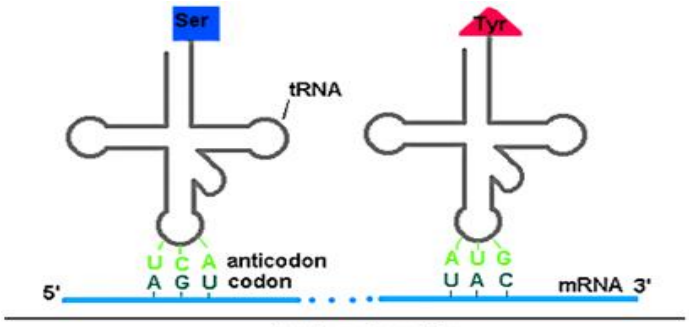


Figure 2 Introns, are removed during the pre-mRNA maturation in eukaryotic organism.

After all these maturation processes, the final mRNA structure is the following: a 5' untranslated region (5' UTR), a protein-coding region (CDS) and a 3' untraslated region (3' UTR). The final step to synthesize a protein from a gene is the translation, which is performed on particular cell structures called ribosomes. The central part of the mRNA, the coding region, stores the information for protein synthesis encoded through a genetic code, which matches to each triplets of nucleotides inside the mRNA (codon), the binding site complementary to an anticodon triplet in a transfer RNA (tRNA) (Fig. 3) [Weiner AM., 2009]. Each tRNA is bound to a specific aminoacid, so, through the help of the ribosome, the reading of the mRNA triplets sequences matching for a tRNA produces a chain of aminoacids. This chain is completed when the translation reaches the stop codon triplets (Fig. 3), and after several maturation steps, it becomes a fully working functional protein.



2nd base in codon

1st base in codon	U	C	A	G	3rd base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Figure 3 According to the genetic code shown in this table, each triplet of nucleotides (codon) in the mRNA is complementary matched to an RNA triplet (anticodon) on the tRNA, which is bound to a specific amminoacid.

In the eukaryotic cell, the transcription, the splicing and the translation are hence the possible checkpoints to manage the gene expression [Zaidi SK., 2004]. The regulation of this latter is fundamental: by managing the quantity and the time of survival of a gene product, the cell has the control over all structures and functions, obtaining the suppleness needed to adapt to different environments and stimuli.

According to their gene expression, the genes can be classified in:

- constitutive genes, which are always transcribed

- facultative genes, transcribed only when there is the need.

- housekeeping genes, needed for the cell maintenance, with a permanent expression level unaffected by the experimental conditions

- inducible genes, whose expression is responsive to environmental stimuli or to the cell cycle status.

In particular, the transcription regulation can be achieved in different ways:

- in the genes, there are several protein binding sites close to the coding region, which regulate the transcription. They are divided in enhacers, insulators and silencers, according to their influence on the transcription. These sites are in fact recognized by proteins, called transcription factors that can promote or avoid the RNA polymerase binding on the template strand of a gene.
- transcription factors activity can be modulated by post-translational modification on them, as phosphorylation, acetylation, or glycosylation, which can be due to cascades of intra cellular signals or due to environmental stimuli or endocrine signals [Nguyen T., 2009].

This addition of small chemical groups can change the transcription factor's ability to bind the promoter, or to recruit the RNA polymerase, or to promote the elongation of a new RNA molecule.

- as described for the transcription factors, also some proteins called histones, which keep organized the genome structure, can be modified with the adding or the removing of small molecules, usually methyl or acetyl group, changing the DNA accessibility for transcription.

- epigenetics: DNA methylation is a very common mechanism of transcription regulation. The adding of methyl group to cytosine or adenine nucleotides inside the gene sequence can physically avoid the binding of transcriptional proteins to the DNA strand or create binding sites for proteins able to change the DNA conformation to a worse shape for the transcription.

The time lasting of the final gene product defines the expression level of the gene too, in fact a not stable product results in a low expression level. As already described during the pre-RNA maturation, capping and poly(A) tail can protect the final mRNA from exonucleases, extending the half life of this RNA molecule. However, post transcriptional gene expression regulation can be also aimed to reduce the expression, as seen for the epigenetic mechanism of the RNA interference (RNAi). In this process, specific mRNAs are degraded or slowed during the translation, by a protein structure called RISC. RISC binds itself to a non coding single strand RNA molecule of 19-21 nucleotides (siRNA, short interfering RNA), able to match complementary the mRNAs that must be suppressed. The siRNA are the product of a processed dsRNA (Fig. 4), a short RNA transcript produced in particular regions of the genome.

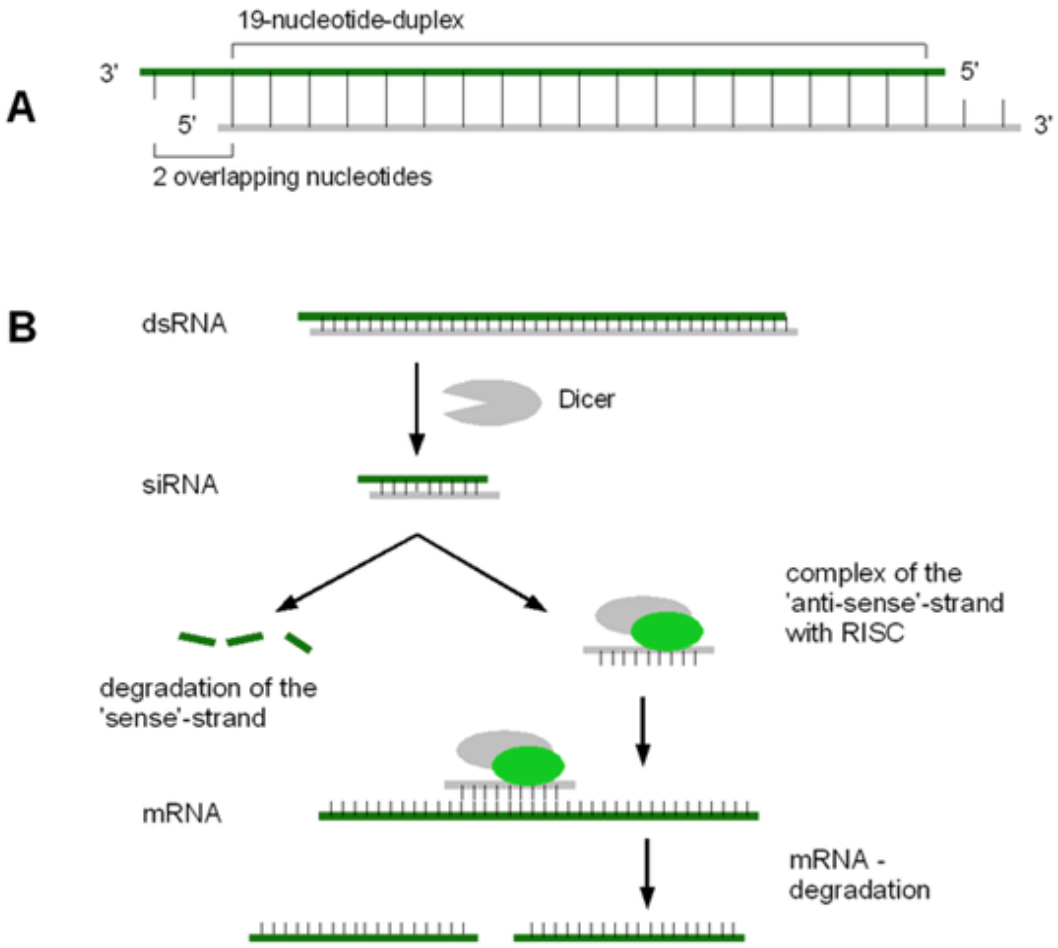


Figure 4 A scheme of the siRNA maturation from dsRNA and its suppressive activity on complementary matching mRNA, through the DICER protein activity.

1.2 Gene co-expression networks

The term ‘co-expression’ indicates a similarity of gene expression patterns across different experimental conditions and it is considered a clue about the presence of possible functional relationships among genes [Aoki et al., 2007]. The coordinated gene expression of several genes in fact is a process: required for the building of protein complexes by different subunits; required for coordinated pathways [Ihmels et al., 2004] and it represents a footprint of the transcription factors’ activity. This functional relationship has been summarized in the “guilt-by-

association" principle, which declares if two genes are co-expressed, the encoded proteins should be involved in the same, or related, cellular functions [Oliver S., 2000]. Because not all the genes maybe co-expressed, to describe the co-expression we have exploited the graph theory and its study of graphs, mathematical structures used to model pair wise relations between objects. In this field, a graph is an ordered pair G = (V, E) comprising a set V of vertices or nodes - the genes in our case - together with a set E of edges or lines – the co-expression event - which are 2-element subsets of V (i.e., an edge is related with two vertices, and the relation is represented as an unordered pair of the vertices with respect to the particular edge). The networks proposed in this work are graphs defined "undirected", meaning that there is no distinction in the direction of the edge linking two vertices. In the graph theory the networks can be described according to their topological properties such as the degree distribution (number of relations made by a node, Fig. 5a), network density (the ratio of the established number of relations to all possible relations, Fig. 5b) and clustering coefficient (the ratio of the observed number of relations between one node and its neighbors, to the number of all possible relations between the node and its neighbors, Fig. 5c).

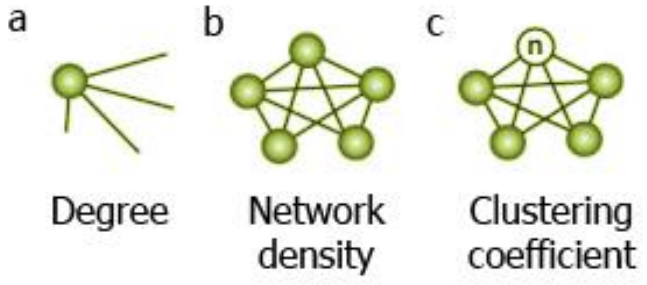


Figure 5 a) Degree: number of relations made by a node; b) Network density: the ratio of the established number of relations to all possible relations; c) Clustering coefficient: the ratio of the observed number of relations between one node (n) and its neighbors, to the number of all possible relations between the node and its neighbors

From a biological viewpoint, we have defined our networks, considering the Arabidopsis protein-coding gene collection as an undirected graph with the vertex representing genes, and edges representing the co-expression, described by the value of the statistical Pearson's correlation test [Pearson K., 1936]. This test declares genes as co-expressed, i.e. statistically correlated, according to the gene expression trend shared through different experimental conditions. This approach hence is based on the hypothesis that if two genes are co-expressed and functionally correlated, the change in the expression level due to environment or whatever stimulus of the first gene, must affects also in the second one, although their expression does not have necessarily the same level. In this way we can not only distinguish among the co-expressed and not co-expressed genes (in the graph a vertex may exists and not belongs to an edge), but also identify "modules", sub graphs with a particular topology corresponding to biologically patterns of co-expression (Fig. 6).

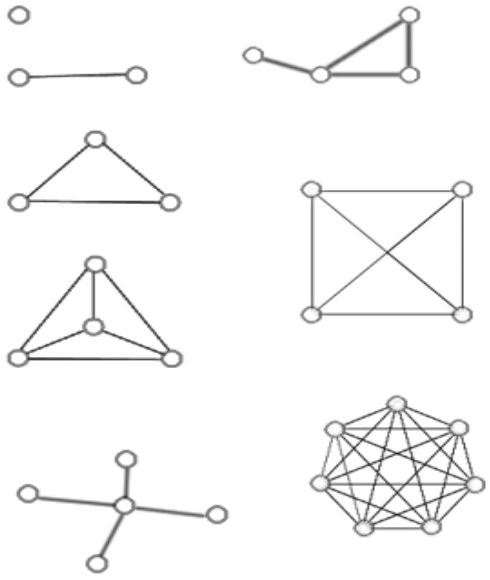


Figure 6 Examples of possible graphs. Some of them have a very specific structures defined as "modules", other a not organized one.

Module identification within a graph representing a network is a great goal because it should represent a collection of genes having a well defined function underlined through their reciprocal interactions [Hartwell et al., 1999]. To identify these modules we have exploited some strategies such as the clique detection, coming from the investigation of not biological system as Internet and society [Barabasi et al., 1999].

1.3 Network and cliques

Given a network with node not all connected by edges like example in figure 7, we define a clique as a group of elements inside a network, where each member is related to each other member of the same group. From the graph theory, this sentence describes the research of complete subgraphs in a graph, i.e., groups of elements where each pair of members is connected. Applying this concept to network of co-expressed genes means to look for group of genes where each gene is correlated with each other gene, and hence, each gene has a number of correlations equals to n-1, where n is the size of the group (clique). In computer science this approach brings two possible descriptions: the maximum clique and the maximal clique. The maximum clique is a clique with the largest number of elements, while the maximal clique is a clique that cannot be enlarged (Fig. 7) [Bron C. et al., 1973]. So, the information obtained through the maximum clique in biology aims to identify the main scaffold networks inside a graph, while the maximal clique detects specific subgroup of co-expressed genes within a graph.

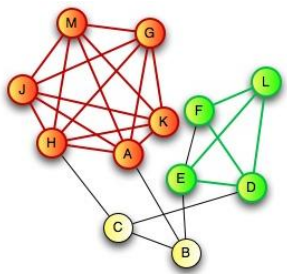


Figure 7 In this undirect graph, following the color scheme, there are 3 maximal cliques. Among these, the orange one is also the maximum clique since it is the largest clique available.

1.4 *Arabidopsis thaliana*

Arabidopsis thaliana, belonging to the *Brassicaceae* family, was the first plant species whose genome was sequenced in 2000 [The Arabidopsis Genome Initiative, 2000]. This plant has been studied for a long time because of its small diploid genome, with only five chromosomes, and its cultivation properties, such as the small size, that limits the requirement for growth facilities, the short life circle and the high seeds production through self-pollination. All these attributes fit properly the researchers' needs, and have led to consider this plant as the best organism to exploit as model plant [Koornneef M. et al., 2010]. However, beyond these important features, its genome has showed an unexpected complexity: probably, three rounds of genome duplications have occurred during its evolution, followed by an undefined number of gene duplications and reassortments [Blanc G. et al, 2003]. This lacking of conserved gene order has hence made difficult the exploitation of this species for studies of phylogenetic relationships among species and moreover, the lacking of an exhaustive annotation underlines how *Arabidopsis thaliana* is still far to be the perfect model organism. A huge number of resources is available for this plant: the Arabidopsis Information Resource (TAIR) [Rhee S.Y. et al., 2003], which is the main gate of all the data related to this plant; the Nascarrays website [Craigon D. J., 2004], the reference platform for expression data in *A. thaliana;* twelve different main websites dedicated to the gene expression profiling, with different datasets and investigations method (see chapter 2.3); the pATsi database (http://cab.unina.it/athparalog/) [SanGiovanni M. et al., 2013] dedicated to the investigation of duplicated genes and realized in our laboratory.

1.5 TAIR

The Arabidopsis Information Resource (TAIR) [Rhee S.Y. et al., 2003] available at http://arabidopsis.org, is a genome database dedicated to *Arabidopsis thaliana* and works as a central access to Arabidopsis gene function, expression patterns, assembly and annotation data. In this latter are present all the information about the Arabidopsis' genes, such as the kind of gene product (protein, trasposon, pseudogen, etc), their positions on the chromosomes and their predicted functions according to the presence of protein domains or signal sequences. Several genome releases where published in the last years. All the analyses in this work were performed on the TAIR9 genome release (ftp://ftp.arabidopsis.org/Genes/TAIR9_genome_release/)

1.6 Microarray

The detection of genes involved in specific cellular processes is an important challenge, lacking of a universally recognized approach. In the work "Genomics. Microarrays--guilt by association" [Quackenbush J., 2003] it has been proposed the "guilt by association" principle which asses that genes co-expressed may be related to accomplish the same task. To fulfill this concept, nowadays one of the most common techniques to identify genes involved in a functional network consists of data mining on massive collections of expression data. Since its release in 1983 [Chang T., 1983], the microarray technology has developed and has been considered the best approach to achieve a snapshot of the expression profiles from thousands of genes, covering the whole genome too in some cases [Craigon D. J., 2004]. So, despite the appearing of new techniques, such as the RNA-seq [Wang Z. et al. ,2009], microarray technology is still considered one of the main tool to investigate the cell transcriptome at low price. This technique is based

on a chip (also named biochip), covered by microscopic DNA spots called probes (or *reporters or oligos*). Each spot can store picomoles (10^{-12} moles) of a specific DNA sequence, with the aim of hybridizing, in high stringency conditions, the target, i.e. a cDNA or cRNA extracted from a sample. These two molecules are obtained exploiting a reverse transcriptase [Brooke-Powell, 2004] on the sample mRNA and during this process they are labeled with a fluorophor, silver-, or chemiluminescent compound. The signals produced by the labels are detected and collected in an image, in order to quantify, and so, to define the abundance of target inside the sample. However these signals may be wrongly detected because of the "noise background", a measurement of signal intensity caused by auto fluorescence of the array surface and non specific binding, and usually background correction and normalization processes are needed after the analyses. Microarrays exploited in our analyses are from the brand Affymetrix [Craigon D. J., 2004], and consist of a number of probe cells where each probe cell contains a unique probe. Probes are tiled in probe pairs as a Perfect Match (PM) and a Mismatch (MM). The sequence for PM and MM are the same, except for a change in the middle of the MM probe sequence that avoid the perfect match with the target sequence. A probe set consists of a series of probe pairs and represents an expressed transcript (Fig. 8).

Figure 8 Probe set structure in the Affymetrix chip

The reason for including a MM probe in the microarray is to provide a value that comprises most of the background cross-hybridization and

stray signal affecting the PM probe. If the MM value is less than the PM value, it is a physically possible estimate for background, and can be directly used. If the MM value is larger than the PM value, it is a physically impossible estimate for the amount of stray signal in the PM intensity. Since each microarray chip can be covered by a collection of different thousands probes covering all the know transcriptome of a species, forming the "microarray slide", in one shot it is possible to evaluate the signal level of each spot, measuring the expression levels of all the probe matching sequences in a sample, usually obtained by the transcription of gene sequences. Microarray technology has been considered for a longtime one of the most powerful approach to study the transcriptional activity of a biological sample, whether it is represented by a tissue, cells, or a mixture, in specific conditions, physiological, stress or pathological ones [Slonim et al., 2009]. Since its power to give a consistent snapshot of the expression of many different genes, though with some technical limits [Hoheisel J., 2006], microarray is still a relevant technology, not only to detect patterns of high or low expressed genes from experiments, but also to describe expression profiles, i.e. the variability of the expression of a multitude of genes, during the evolving of a specific process and measured in different stages.

1.7 Nascarrays

The improving of the microarray technology has offered the possibility to analyze, with one microarray, the whole transcriptome of an organism in a specific condition [Craigon D. J., 2004]. In fact, thanks to an advanced probes production, supported by the availability of a complete sequenced genome, a large collection of *Arabidopsis thaliana* microarrays has been produced in the last years. As a consequence, a huge number of microarrays collecting databases have appeared, and in

particular the Nascarrays website (http://affymetrix.arabidopsis.info/) [Craigon D. J., 2004] is considered the reference site for all the public Affymetrix ATH1 and AG 'GeneChip' microarrays performed on *Arabidopsis thaliana*. At this moment, the platform in fact collects 706 experiments and 5364 slides. All data are described by the sample information, hybridization, normalization and scanning protocol exploited. For each gene in a slide, the expression is defined by:

- the signal: raw adjusted intensity,

- stat pairs: number of probe pairs to interrogate each gene,

- stat pairs used: number of pairs used to calculate signal,

- detection call: presence or absence of transcript,

-detection p-value: p-value used to determine presence or absence of transcript,

and generally the original probe measures of the CEL files are available too.

Nascarrays allows to the user to find microarray experiments in two ways: i) searching by sample condition as tissue, development stage, growth conditions, treatment, etc, or by keywords, PO term, genetic background or genechip exploited; ii) browsing through the experiments indexed by date, author, MGED classification, experiment design with two specific folders for the microarrays belonging to the AtGenExpress Project and AFGC project. The Atgen express project [Schmid M., 2005] in fact represents the main producer of microarray data for *A. thaliana*, since they have depicted through their analyses the expression profiles of all the genes activated in specific conditions. Data mining section completes the Nascarrays platform with several useful tools: the spot history tool, which shows the expression profile of a studied gene

over all the available experiments; the two gene scatter plot tool, which shows in a scatter plot the expression values of two considered genes through all the experiments; the gene swinger tool which shows the experiments whose expression value for a desired gene has changed significantly compared to the ones inside the other experiments; last, the bulk gene download tool let the user to download all the expression profiles of a gene (or of all the genes, with the superbulk command) over all the experiments of the database. All the information can be downloaded in the common file formats as CSV, Excel, Gnumeric spreadsheet or TAB-delimited.

1.8 Objectives

This work aims to analyze mainly co-expressed genes in *Arabidopsis thaliana*, according to the complexity of its genome. In fact, after its sequencing in 2000 [The Arabidopsis Genome Initiative, 2000] researchers have understood that probably in its small diploid genome with only five chromosomes, three rounds of genome duplications have occurred during its evolution, followed by a huge number of gene duplications and reassortments [Blanc G. et al, 2003]. So, *Arabidopsis thaliana*, which before its genome sequencing was already exploited commonly for analyses, has concentrated even more the attention on itself, as we have done with our study.

The identification of co-expressed gene groups represents an important objective, needed to describe the networks of interactions available among the genome components. The identification of these interactions is the key to understand the main characters of the cell functions and to draw the scheme exploited by the cell to perform its tasks. Hence we want to understand if co-expressed genes are involved in the same specific function and their degree of organization. The gathering of

information about the co-expressed genes has inevitably led us to challenge some important questions in bioinformatics, as the research of a well fitting method able to group genes with a biological meaning, and the evaluation of the statistical components exploited to define genes as co-expressed. With this work hence we want also to evaluate the distribution of co-expression networks inside the genome, and the public available platforms concerning the co-expression issue. During these activities, we have also tried to produce a new approach to identify co-expressed gene networks, inspired by social networks and able to underline quickly the genes in relation inside the genome. In fact, while the definition of gene families according to their DNA or protein sequence similarity has become one of the main tool of bioinformatics [Lobo I., 2008], the detection of genes involved in specific cellular processes is an important challenge, lacking of an universally recognized approach.

Materials and methods

2.1 Dataset

The "Developmental Series Expression atlas of Arabidopsis development" subfolder (http://affymetrix.arabidopsis.info/narrays/experimentbrowse.pl) in the Nascarrays "Atgen express project" folder, stores a huge collection of experiments describing the genes expression in each tissue and according to the plant development stage, growth condition, and presence of mutations inside the genome.

From this subfolder we have downloaded the gene expression values of 79 experiments with samples taken from several tissues, in physiological conditions and repeated in triplicate, for a total number of 237 microarray slides (Tab. 1). Each slide was based on the ATH1 Affymetrix chipset, able to detect 22810 probes and normalized through MAS 5.0 protocol (top and lowest 2% of each signal in each experiment were removed, than all the values in each experiment were transformed in order to have an average value of 100). Only 21769 probes had signals and from these latter we have removed the following probes: 387 know as multiple genes matching, 53 no gene matching (trasposon, miRNA, etc), 107 similar to unrelated sequences (x_at probes) (Redman et al., 2004), 27 shared probes (s_at), 3 "sequence family" probes (f_at), 1 "rules dropped" probes. Moreover, the expression signal of 224 genes was defined by more than one probe (totally 458 redundant probes), so we took the average of these ones. The final step was to filter out all the probes with an expression level under the 5^{th} percentile in each sample, in all the experiments, bringing the final number of gene specific probes exploited in this work to 20908. The average signal of each probe in each experiment was calculated in Excel taking the average of the three replicates. A log2 transformation on all the signals has been applied as

final step: in this way only genes with a huge difference in the signal will be treated and considered as differently expressed

Sample ID	Genotype	Tissue	Age	Photoperiod	Substrate
ATGE_1	wild type	cotyledon	21+ days	continous light	soil
ATGE_2	wild type	hypocotyl	21+ days	continous light	soil
ATGE_3	wild type	root	21+ days	continous light	soil
ATGE_4	wild type	shoot apex, vegetative + young leaves	21+ days	continous light	soil
ATGE_5	wild type	leaves 1+2	21+ days	continous light	soil
ATGE_6	wild type	shoot apex, vegetative	21+ days	continous light	soil
ATGE_7	wild type	seedling, green parts	21+ days	continous light	soil
ATGE_8	wild type	shoot apex, transition (before bolting)	21+ days	continous light	soil
ATGE_9	wild type	roots	21+ days	continous light	soil
ATGE_10	wild type	rosette leaf #4, 1 cm long	21+ days	continous light	soil
ATGE_11	gl1-T	rosette leaf #4, 1 cm long	21+ days	continous light	soil
ATGE_12	wild type	rosette leaf #2	21+ days	continous light	soil
ATGE_13	wild type	rosette leaf #4	21+ days	continous light	soil
ATGE_14	wild type	rosette leaf #6	21+ days	continous light	soil
ATGE_15	wild type	rosette leaf #8	21+ days	continous light	soil
ATGE_16	wild type	rosette leaf #10	21+ days	continous light	soil
ATGE_17	wild type	rosette leaf #12	21+ days	continous light	soil
ATGE_18	gl1-T	rosette leaf #12	21+ days	continous light	soil
ATGE_19	wild type	leaf 7, petiole	21+ days	continous light	soil
ATGE_20	wild type	leaf 7, proximal half	21+ days	continous light	soil
ATGE_21	wild type	leaf 7, distal half	21+ days	continous light	soil
ATGE_22	wild type	developmental drift, entire rosette after transition to flowering, but before bolting	21+ days	continous light	soil
ATGE_23	wild type	as above	21+ days	continous light	soil
ATGE_24	wild type	as above	21+ days	continous light	soil
ATGE_25	wild type	senescing leaf	21+ days	continous light	soil

ATGE_26	wild type	cauline leaf	21+ days	continous light	soil
ATGE_27	wild type	stem, 2nd internode	21+ days	continous light	soil
ATGE_28	wild type	stem, 1st node	21+ days	continous light	soil
ATGE_29	wild type	shoot apex, inflorescence (after bolting)	21+ days	continous light	soil
ATGE_31	wild type	flower, stage 9	21+ days	continous light	soil
ATGE_32	wild type	flower, stage 10/11	21+ days	continous light	soil
ATGE_33	wild type	flower, stage 12	21+ days	continous light	soil
ATGE_34	wild type	flower, stage 12, sepals	21+ days	continous light	soil
ATGE_35	wild type	flower, stage 12, petals	21+ days	continous light	soil
ATGE_36	wild type	flower, stage 12, stamens	21+ days	continous light	soil
ATGE_37	wild type	flower, stage 12, carpels	21+ days	continous light	soil
ATGE_39	wild type	flower, stage 15	21+ days	continous light	soil
ATGE_40	wild type	flower, stage 15, pedicels	21+ days	continous light	soil
ATGE_41	wild type	flower, stage, 15, sepals	21+ days	continous light	soil
ATGE_42	wild type	flower, stage, 15, petals	21+ days	continous light	soil
ATGE_43	wild type	flower, stage, 15, stamen	21+ days	continous light	soil
ATGE_45	wild type	flower, stage, 15, carpels	21+ days	continous light	soil
ATGE_46	clv3-7	shoot apex, inflorescence (after bolting)	21+ days	continous light	soil
ATGE_47	lfy-12	shoot apex, inflorescence (after bolting)	21+ days	continous light	soil
ATGE_48	ap1-15	shoot apex, inflorescence (after bolting)	21+ days	continous light	soil
ATGE_49	ap2-6	shoot apex, inflorescence (after bolting)	21+ days	continous light	soil
ATGE_50	ap3-6	shoot apex, inflorescence (after bolting)	21+ days	continous light	soil
ATGE_51	ag-12	shoot apex, inflorescence (after bolting)	21+ days	continous light	soil

ATGE_52	ufo-1	shoot apex, inflorescence (after bolting)	21+ days	continous light	soil
ATGE_53	clv3-7	flower, stage 12; multi-carpel gynoecium; enlarged meristem; increased organ number	21+ days	continous light	soil
ATGE_54	lfy-12	flower, stage 12: shoot characteristics; most organs leaf-like	21+ days	continous light	soil
ATGE_55	ap1-15	flower, stage 12: sepals replaced by leaf-like organs, petals mostly lacking, 2° flowers	21+ days	continous light	soil
ATGE_56	ap2-6	flower, stage 12: no sepals or petals	21+ days	continous light	soil
ATGE_57	ap3-6	flower, stage 12: no petals or stamens	21+ days	continous light	soil
ATGE_58	ag-12	flower, stage 12: no stamens or carpels	21+ days	continous light	soil
ATGE_59	ufo-1	flower, stage 12; filamentous organs in whorls two and three		continous light	soil
ATGE_73	wild type	mature pollen	6wk	long day(16/8)	soil
ATGE_76	wild type	silique, with seeds stage 3; mid globular to early heart embryo	8wk	long day(16/8)	soil
ATGE_77	wild type	silique, with seeds stage 4;early to late heart embryo	8wk	long day(16/8)	soil
ATGE_78	wild type	silique, with seeds stage 5	8wk	long day(16/8)	soil
ATGE_79	wild type	seed, stage 6; mid to late torpedo embryos	8wk	long day(16/8)	soil

ATGE_81	wild type	seed, stage 7; late torpedo to early walking-stick embryo	8wk	long day(16/8)	soil
ATGE_82	wild type	seed, stage 8; walking-stick to early curled-cotyledons embryo	8wk	long day(16/8)	soil
ATGE_83	wild type	seed, stage 9; curled-cotyledons to early green-cotyledons embryo	8wk	long day(16/8)	soil
ATGE_84	wild type	seed, stage 10; green cotyledons embryo	8wk	short day (10/14)	soil
ATGE_87	wild type	vegetative rosette	7 days	short day (10/14)	soil
ATGE_89	wild type	vegetative rosette	14 days	short day (10/14)	soil
ATGE_90	wild type	vegetative rosette	21 days	short day (10/14)	soil
ATGE_91	wild type	leaf	15 days	long day (16/8)	soil
ATGE_92	wild type	flower	28 days	long day (16/8)	soil
ATGE_93	wild type	root	15 days	long day (16/8)	soil
ATGE_94	wild type	root	8 days	continuos light	soil
ATGE_95	wild type	root	8 days	continuos light	soil
ATGE_96	wild type	seedling, green parts	8 dyas	continuos light	soil
ATGE_97	wild type	seedling, green parts	8 days	continuos light	soil
ATGE_98	wild type	root	21 days	continuos light	soil
ATGE_99	wild type	root	21 days	continuos light	1x MS agar, 1% sucrose
ATGE_100	wild type	seedling, green parts	21 days	continuos light	soil
ATGE_101	wild type	seedling, green parts	21 days	continuos light	1x MS agar, 1% sucrose

Table 1 List of samples exploited in our analyses and collected from the "Developmental Series Expression atlas of Arabidopsis development", on Nascarrays' website.

2.2 Pearson's correlation

In biology, one of the most used criteria to establish a relation between two gene expression partners is the Pearson's correlation [Pearson K., 1936]. Generally, this statistical test evaluates the linear correlation between two variables, *X* and *Y*, giving a value between +1 and −1 inclusive, where 1 is total positive correlation, 0 is no correlation, and −1 is total negative correlation. This value is defined as "ρ", *population correlation coefficient* or *population Pearson correlation coefficient*, when applied on a population, and is obtained through the following formula:

$$\text{I)} \quad \rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

where cov is the covariance, σ_X is the standard deviation of X, μ_X is the mean of X, and E is the expectation. On limited samples, as it happens in our study, by substituting estimates of the covariances and variances based on a sample into the formula in I), the value obtained is *r* (or R), i.e. the s*ample correlation coefficient* or the *sample Pearson correlation coefficient*, described in detail with the formula in II).

$$\text{II)} \quad r = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}$$

with *n* equals to the number of samples.

As for "ρ", the range of "*r"* is between -1 and 1, defining the strength of the relationship between two variables as *r* becomes close to its boundaries.

The Pearson's correlation is at the base of most of our analyses.

2.3 Public co-expression platforms

Arabidopsis thaliana transcriptome has been largely studied exploiting the microarray technologies. Indeed an overwhelming amount of experiments based on the Affymetrix platform have been conducted and collected at the Nacarrays website (http://affymetrix.arabidopsis.info/) [David J. Craigon, 2004]. As a consequence, this pushed the flourishing of web based resources including both microarray data collections and dedicated tools for co-expression analyses. In this chapter we have listed the main platforms considered.

2.3.1 ACT

ACT offers a wide package of useful tools to perform co-expression analyses, although its dataset is relatively small. It includes 54 experiments from the Nascarrays website [http://affymetrix.arabidopsis.info/narrays/experimentbrowse.pl], organized in 122 AtGenome1 (8k probes) and 422 ATh1 (22k probes) arrays, covering several tissues, biotic and abiotic stress conditions, mutants and a range of developmental stages experiments too. In this latter, normalization and preparation of samples stored in Ath1 chip experiments follow a standard Affymetrix processing protocol based on MAS 5.0 software, with an added step fixing to 20 the lowest signal unit detected in the experiments with at least one sample above this number. By this way, genes with an expression values under 20 in each experiment, have been excluded. Correlations are expressed through the Pearson Correlation (PC) coefficient (r) [Pearson K., 1936] which assumes value between -1 and 1, expressing a strong or a weak correlation between pair of genes when ranging from|1| to 0. About the interface, ACT offers seven different tools to analyze correlations among *Arabidopsis* genes: i.e., it is possible to evaluate the co-

expression values of a gene not only with the ones belonging to the same slide or same experiments, but with a user defined specific list of slides or experiments chosen among the ones available in the website dataset too. The Co-correlation Scatter Plot and Clique Finder tool are the most interesting ones. Using two gene probes as query, with the first tool it is possible to visualize in a scatter plot the correlation values of all genes against the inserted ones, offering in this way, a global image of the found correlations. The Clique Finder tool, instead, proposes a list of, at most, top 100 correlated genes with the one requested by the user. These ones are gathered in specific clusters, where each element is related to every other member of the cluster, if the shared correlation among them has an r-value above a threshold fixed by the user. So the Clique finder should offer the most related genes in a very stringent way.

2.3.2 ATCOECIS

AtCOECiS [Vandepoele et al., 2009] is an online platform exclusively for *Arabidopsis thaliana* that allows the user not only to identify co-expressed genes but also gene co-expression neighborhoods associated with cis-regulatory motifs or GO categories. The microarray data used for co-expression analyses, were retrieved from the Nottingham Arabidopsis Stock Centre (http://arabidopsis.info/) and they include a total of 322 (48x3 AtGenExpress Development and Tissue slides + 68x2 AtGenExpress Stress slides + 42 Birnbaum Root slides) Affymetrix ATH1 microarray slides in different tissues and under different experimental conditions. The raw data were then normalized with RMA and a custom-made chip description file (CDF) that can reliably measure and discriminate between the expressions of both copies of duplicated gene pairs with valid probe sets. With the aim of verifying on predefined functional sets of genes the guilty-by-

association relationship, which establishes a link between gene expression trend and the gene function, they quantified the level of expression similarity using the expression coherence (EC). EC is a measure of expression similarity level in a gene set, ranking between 0and 1, reporting the fraction of gene pairs per GO category that shows elevated co-expression. Hereafter, the PC coefficient has been used as a measure to describe the similarity between expression profiles and tree different thresholds have been applied: higher than 0.63, 0.72 and 0.83. The tool offers different types of search options for co-expressed genes: the user can use a seed gene or a GO category as query. The output shows to the user the gene annotation of the query followed by the associated GO categories, the proprieties of co-expression neighborhoods, the cluster size and the clustering coefficient. Next to the cluster proprieties there is the link for the complete co-expressed genes list, even in txt mode, that shows all the genes correlated with the query, ranked by the PCC. In the output there is also information, like GO and motif enrichment, that is the ratio of the GO/motif frequency in the cluster over the background GO/motif frequency (in the complete genome).

2.3.3 ATTED-II

ATTED-II [Obayashi et al., 2007] was released in 2007 and it is a co-expression database for *Arabidopsis thaliana,* rice, soybean, maize, grape, medicago and poplar, although, in this moment it is possible to query only *Arabidopsis* genes. The *Arabidopsis* microarray dataset collects a huge number of Affymetrix microarrays data from Array Express, with a total of 11171 slides in 737 experiments, including several tissues and biotic or abiotic stress conditions. Moreover, the dataset collects RNA-seq data too with 328 slides in 28 experiments. Raw RMA normalization has been applied to the series and then the

gene expressions values were further normalized by subtracting the average expression value across all chips in pre-defined subseries of experiments. ATTED-II is organized primarily in two sections called “Search” and “Draw”.

-“Search” section offers four tools to obtain information about genes functions and about their expressions variation using different and global microarray datasets or user defined correlations list. These correlations are ordered by the Mutual Ranking (MR) algorithm: as first step, a group or a single driver gene is chosen by the user, and its correlations with each other gene in the genome are ranked by their Pearson value; as we were using each gene obtained in this list as query, the rank positions of their correlations with the first gene driver are found, and then, the final value of the mutual ranking is equal to the geometric average between the rank position of “gene driver”-“gene in the list” correlation and the rank position of the “gene in the list”-“gene driver” correlation. In this way, the result of a co-expression query for a specific gene in ATTED-II is the list of the first 300 genes ordered by their decreasing MR. The main benefit of Mutual Ranking value, in comparison with the most used PC values is its lower sensitiveness to the differences within the tissues and experimental conditions of microarrays dataset.

- Even in the “Draw” section of ATTED-II are available four tools to visualize gene relations networks, hierarchical clustering, gene-to-gene co-expression and Gene Ontology (GO) classification. The first tool is called Network Drawer and displays the genes correlated with the ones chosen by the user, selecting, for each query gene, the three elements with the highest Mutual Ranking values that share at least two links among the other members of the network. The outputs of this tool are networks built in Cytoscape [Shannon et al., 2003]. Mutual Ranking

values is also exploited as clustering distance measure in the H cluster, the second tool developed for hierarchical clustering, which offers even a clustering based on the classical Pearson value (1-r) distance, in order to compare the two methods. The Coexviewer tool instead depicts in a scatter plot the contribution of each specific sample in defining as co-expressed a pair of genes and moreover, it offers the chance to evaluate the co-expression stability by PCA analysis. The last one, the GO path drawer aims to show the relationships and the GO terms hierarchy of the GO keywords chosen as query.

2.3.4 BAR

The BAR (Bio-Array Resource for Plant Biology) [Toufighi et al., 2005], from University of Toronto, is an on-line platform that offers several tools for the management and exploration of the expression data, primarily in *A. thaliana*. Several microarrays datasets, all based on Ath1 Affymetrix chip, are available for correlation study: i) a self made microarrays database of about 175 experimental samples; ii) the NASCarrays database collecting 392 samples and iii) the Atgen Express project microarray slides, organized in Hormone, Stress, Pathogen, Tissue, Seed, Root and Extend Tissue dataset; iv) a dataset organized according to *A. thaliana* ecotype. The philosophy beyond BAR website is to offer simple and smart tools, developed with the most user friendly interface. The Expression Angler tool shows the best (or the worst) correlated genes with the query one, according to their PC value, calculated using one of the dataset described in the table or exploiting a customized one. Query results are available not only as text, but with a really practical heat mapping visualization format too which let the user to have genes ranked by their PC values, while showing their expression profiles distribution among the experiment slides considered for the correlation analysis. In a quick way, weak or strong predicted

interactions (*interolog*) among these ranked genes are shown by arcs, and moreover to complete this powerful graphic scheme, a functional classification depicted by a color system attends each gene involved in a significant correlation and not last, the "promomer" tool, allows to find cis elements within the promoters of the genes of interest. Sample Angler is a tool aimed to detect a shared expression trend between two or more samples, chosen from a particular dataset or from a self made one. As it happens for genes correlation, microarray slides similarity is expressed with PC value too and, according to this latter, a short ranking list with the heat mapping graphics is shown. Similarly, slide sharing similar experimental conditions as tissue, mutant status, time and treatment, are also pinpointed in the heat mapping graph with an easy color scheme annotation. In the main result page, the user can find also other information: two scatter plots describing gene expression trend among the two most correlated samples; a complete list of the most correlated samples ranked by Pearson value; a list containing the top 10 genes and the worst 10 ones according to their expression similarity, which have contributed to assign or avoid a correlation between the samples compared and, last, a gene score distribution graph showing the frequency and the grade of this expression similarity among the genes in the samples. The same visualization tool is available for the Expression browser tool. Expression browser let to analyze up to 125 gene expression values in a defined dataset selected from the ones proposed by BAR or from one of the seven AtGenExpress series cited above. The results are shown by the already described heat mapping tool in an unclustered or a clustered version which clusters and collects all the available details about sample conditions, gene expression, function classification or protein interactions. The electronic Fluorescent Pictograph Browser (eFP) offers for a specific probe a very impressive scheme of the most involved in gene expression plant

tisuees. In the query form, user can define the sample dataset exploited for the visualization of the gene expression distribution, and in the results page it is shown the change of expression of a desired gene, described by the color variation in experiment illustrating pictures. Moreover, as expected, a more traditional chart tool, describing gene expression variation alone or in comparison with another gene is available too. With the similar graphic approach exploited in the eFP tool, the Cell eFP browser goes in depth, showing through the usual color index already seen before, the most probable sub cellular organ(s) containing a query protein, according to computationally predictions and experiments. In order to realize an exhaustive analysis on *A. thaliana* gene expression, a species comparison tool, called Expresslog Tree Viewer, shows the best orthologs of a query gene found in poplar, medicago, soybean, rice, barley and maize, keeping in consideration not only the sequence similarity among them but their expression profiles too. However, the most impressive tool for gene co-expression evaluation on BAR is the ”Arabidopsis Interaction Viewer”: this tool offers a really detailed landscape of protein interactions, showing in one graph all the relations within a proteins query list, defined by PC, experimental results and computational predictions obtained by associating ortholog proteins interaction behaviors in yeast (*Saccharomyces cerevisiae*), nematode worm (*Caenorhabditis elegans*), fruitfly (*Drosophila melanogaster*), and human (*Homo sapiens*). The whole information depicted in this graph, is also listed in a complete table containing, for each gene, its pc value with each other involved genes, its sub cellular localizations, pubmed recording, annotations, and, if present, its interolog relations, specifying which ortholog organism is involved too. Other smaller, but not less important instruments complete the BAR tools suite for gene expression analyses. For example, AGURR (Arabidopsis Genetic Uniqueness and

Redundancy Revealer) aims to discover and isolate those genes expressed exclusively only in one dataset, across a multitude of experimental conditions, while the Sample Angler identifies samples that exhibit similar pattern of expression for all genes or for a subset of genes in BAR database, in several AtGenExpress datasets from the AtGenExpress Consortium, and an in-house generated chemical genomics dataset.

2.3.5 COP

CoP (Co-expressed biological Process) [Ogata et al, 2010] is an online platform with the main proposal of associating genes with similar expression profiles and biological information. This database contains the expression data from several plants, included *Arabidopsis*. Microarray data for *Arabidopsis* were taken from GEO and ArrayExpress, where 5257 chips (CEL files or MAS5-processed data files from Affymetrix Gene Chip) have been selected. Then the CEL files have been analyzed using the Bioconductor package 2.3.13 with R version 2.8.1 to obtain MAS5-processed text data, that after became standardized, they are used to calculate gene-to-gene correlation. For the correlation, they have considered only positive gene-to-gene correlation, not using the most used PC, but exploiting the cosine correlation (CC), which considers only correlation between 0 and 1. The main approach to analyze gene co-expression on Cop website is choosing "the gene-co-expression information" from the main page, using AGI code, probe id or gene name in the query form. In the result page, genes are listed not only by CC, but primarily by their Vertex F-measure (VF) [http://webs2.kazusa.or.jp/kagiana/cop0911/pages/ terms. html], ranged 0-1, which indicates the strength of a gene to belong to a group of genes. It is calculated as VF= 2 x E / (D + N - 1), with E equals to the number of edges within a network module, D equals to

connectivity degree of the module members to any genes (the number of gene-to-gene links of the module members to any genes) and N the number of modules member). So genes with the highest VF values are chosen as the most co-expressed ones. In the Cop website Network, modules of co-expression are identified through the "Confeito" algorithm which produces and ranks network modules according to the Network F-measure (NF), ranged 0-1, which is the harmony mean between the Network Recall (NR) and the Network Precision (NP) [http://webs2.kazusa.or.jp/kagiana/cop0911/pages/terms.html] where NR=E/((N x (N - 1) / 2) and NP=E/D. Network modules obtained in this way are statistically supported by comparing the percentile score of the NF index of the module obtained to the NF indices of all other modules. Over the expression analysis section, Cop offers other useful information pages too, like the "specific expression" one i.e., which shows the microarray samples with the highest level of expression for a requested gene, or vice versa, through "the microarray experiments" page it is possible to check the most expressed genes of a specific GSM sample or GEO Series. As an added value, the homology, gene ontology and metabolic pathways sections, respectively let the user to discover the paralogs and orthologs of the query gene through a BLASTn algorithm, identify the genes according to their gene ontology or visualize the genes involved in a specific metabolic pathways among the ones stored on KEGG.

2.3.6 CORNET

CORNET (CORrelation NETworks) [De Bodt et al., 2010], released in 2009, is another on line microarray platform specific for *Arabidopsis thaliana*. The dataset of experiments stored in the database is very huge, in fact it includes: 425 experiments from all AtGenExpress (http://www.arabidopsis.org/portals/expression/microarray/ATGenExpr

ess.jsp); 454 and 192 experiments from Microarray compendium 1 and 2, respectively; 256 experiments for abiotic stress series (cold, drought, genotoxic, heat, osmotic, oxidative, salt, UV-B, wounding); 69 for biotic stress series (*Botrytis*, *Pseudomonas*, *Phytophthora*, etc.) and, 336 ones with a combination of the abiotic and biotic stress; 235 experiments for developmental series (different tissues, developmental stages, developmental mutants); 72 experiments in flower, 212 in leaf, 258 in root, 83 in seed and 83 for the whole plant; 313 experiments with genetic modifications, in which transgenic lines are profiled (gene overexpression (knock-in), gene knock-out, transient transgene expression); 140 experiments with hormone treatment (ABA, brassinosteroids, GA, cytokinin, etc. and inhibitors). The set was normalized with RMA procedure, i.e. background correction, normalization, summarization and then the output has been log2 transformed. The site offers two tools, namely co-expression and PPI tool. The co-expression tool allows identifying genes with similar expression profiles with the query gene/s. The user can select both the predefined expression datasets (one or more than one, among those described before) or compile a user-defined dataset. The correlations can be calculated either with Pearson or Spearman test, and the threshold can be fixed by the user. It is also possible to know the localizations of the proteins translated by the genes co-expressed and the output of this tool is a Cytoscape view of the correlations or, if preferred, a downloadable text file.

The two tools of the site can be exploited together, with only one query. In this way, the output will include not only the correlations between the genes but also the possible interactions between the proteins translated by them. Gene annotations are provided both in Cytoscape output and in text format output.

2.3.7 CressExpress

While the major part of co-expression databases provides a single oriented dataset viewpoint, CressExpress [Srinivasasainagendra et al., 2008] offer a more customizable approach in this field. Available since 2008, in fact this resource allows choosing not only different microarray datasets collected from NASC website, but it offers the chance to select also the preferred chip platform and normalization method. As reported in the table below (Tab. 2), 4 microarrays dataset releases are selectable for co-expression analyses:

Release	Data sources	Number arrays (slides)	Array processing/normalization	Nr. of experiments
2.00	Affywatch I,II	486 ATH1, 80 AG	RMA	147
3.00	Affywatch I,II,III	1779 ATH1, 80 AG	RMA	190
3.01	Affywatch I,II,III	1779 ATH1	GCRMA	190
3.02	Affywatch I,II,III	1779 ATH1	MAS5/log2 transformation, divide by average	190

Table 2 Experiment releases available on CressExpress database

CressExpress customization features become clearer using its query interface: after the selection of the microarray release to exploit, a Kolmogorov-Smirnov quality-control filter is available, which allows the dataset filtering by keeping for co-expression analyses only robust or weak statistical microarrays slides. Moreover, during the last query steps, it is possible to focus analyses not only to a specific class of tissue microarrays, but also to experiment in a specific condition too. Co-expression among genes is expressed through r^2, the square of the common used PC: its outcoming p values and slope numbers show the positive or negative nature of the correlation. In this way, correlations having a really low p-value, and a correlation value a little bit higher than the low default threshold of r^2 equals to 0.36, still assume a strong

statistical meaning. In addition, CressExpress offers a pathway co-expression density analysis, defined as PLC (pathway level co-expression), which allows the user to have a ranking of the most co-expressed genes in an Aracyc pathway, with the ones chosen in the query, according to an user defined r^2 threshold, p-value and numbers of connections established by each gene. Another characterizing feature of CressExpress is its ready-to-use delivery system: in fact all the information obtained from this database are not showed directly in the browser, but they are available as text and excel data files stored in a single and compressed file, which is delivered directly by e-mail in order to speed up the data manipulation with the most common bioinformatics software. So, contrarily as what happens in many databases, the basic idea of CressExpress is not to manage results in a not optimized environment, but it offers directly the data in the most quick and practical way. Hence users will find in their mail a compressed file containing: a .csv and a .txt list file specific for each probe chosen at the query beginning, containing the list of the correlation values with all the other genes in the dataset; an "ALL" file merging in one document all the co-expression values established among the query genes and the other ones in the dataset; an "Experiment Microarray Map" describing the experiments and microarray slides chosen in the query, with their statistical Kolmogorov-Smirnov value; a "PLC Annotation Id" document with GO terms, Gene description and Gene Symbol for each PLC analysis involved gene; a "PLC Co-regulated Genes List" with the average r^2co-expression values of each gene against the query; the "Query Ids", showing only the correlations between the query genes and the "PLC Results" with the PLC ranking, p-value, slope and r^2values of each gene against the query ones. All these results are released in Cytoscape format too within the initial compressed file.

2.3.8 CSB.DB

The Comprehensive Systems Biology Project (CSB) [Steinhauser et al., 2004] website hosted at the Max Planck Institute of Molecular Plant Phisiology was developed with the purpose of containing transcriptional correlations databases of key model organisms as *A. thaliana*, *Saccharomyces cerevisiae* and *Escherichia coli*. The first one, AthCoR@CSB.DB offers a co-expression querying platform based on four base 2 logs normalized primary microarrays collection, one from the NASC's International Affymetrix Service and the other three from the AtGenExpress consortium. The first one, the nasc0271 (m0271) dataset, contains 51 experiments covering several treatments and mutant characterization for 9694 genes; the atge0100 dataset instead collects correlation profiles of 12200 valid measures of genes through 63 experiments about the development stages of several tissues and without mutants; III) the atge0200 dataset stores 13197 gene expression profiles in 60 experiments about aboveground abiotic stress and the last one, the atge0250 offers info about 15377 genes through 60 experiments about root abiotic stress. Three tools are available on AthCoR@CSB.DB . The first one is the Single Gene Query (sGQ). As the name suggests, sGQ tool allows obtaining the most correlated genes for a query one, according to the expression profiles of one of the dataset described above. Correlations among genes can be defined with a Pearson correlation test, or with a Spearman or Kendall one, while the query output can be customized in order to have genes shown according to their correlation value, statistical meaning or if belonging to a particular cell process or categories. In the query results, genes are ranked by their degree of correlation with the one of interest and moreover, for each gene in the list it is available a short description, the number of pairs, the confidence interval, the power, the mutual

information and their normalized Euclidean distances according to the one chosen in the query. Two visual representations of the most important biological process and terms associated with the best co-expressed genes are available too. The second tool, the multiple gene query, follows the same interface of sGQ but, however, it shows the correlations established only among a list of 60 genes of interest at most. In this way, user can identify and rank the most co-expressed elements of a customized gene pool. The last tool of AthCoR@CSB.DB is the Intersection Gene Query (isGQ) and let the user to choose two or three genes of interest and identify the most co-expressed ones shared by the ones stored in one of the four dataset defined during the query. Two or three lists of genes (according to the number of inputs inserted in the query form) ranked by their shared correlation degree are shown in the result page, each one with all the statistical and biological information described as in SGQ. Moreover if only two genes are selected as input query, results can be customized in order to have the best positive correlations with the first gene and the most negative ones with the second gene and vice versa.

2.3.9 GENECAT

GeneCAT (gene co-expression analysis toolbox) [Mutwil et al., 2008] is a multispecies database, released in 2008, containing the gene expression values for *Arabidopsis thaliana*. The microarray dataset stored on GeneCAT consists of 351 RMA normalized Ath1 slides from TAIR [http://www.arabidopsis.org]. After choosing one or more genes for the query in the "Data entry" section of the site, it is possible to analyze the desired genes using 5 different tools. The "Expression profiling" tool allows to visualize and compare the expression values of query genes among several tissues and conditions, while below the "Expression tree" tool offers the degree of co-expression among the

selected genes, in an UPGMA phylogenetic tree output, built on a Pearson's correlation distance matrix. As the name suggests, "Co-expression analysis" tool is the core of co-expression investigation in GeneCAT: this latter compares the expression profile of the query gene to every other gene in the database, with a ranked by PC elements list as output, which can be moreover filtered by a specific r-value threshold too, if preferred. With the aim to point out a common biological role among the co-expressed genes shown in the list, the result page offers also i) a co-expressed gene network built by measuring mutual co-expression ranks in a pair-wise manner between the 50 most correlated with the query term genes, and ii) a table containing the top 150 genes from the list, clustered by sequence similarity. Another useful tool of GeneCAT is Map-o-matic which, after declaring a dataset of defined genes, allows visualizing the Pearson values distribution of the correlations between these latter and the genes chosen for the query.

2.3.10 Genemania

Genemania [Mostafavi et al., 2008] can be considered the Swiss army knife of gene function and interaction prediction. Released in the 2010, this platform integrates PPI, literature, genomic and proteomic information from several on line datasets with the purpose to develop or to define by *the novo*, the functional roles, the relations and the possible interactions of a single or multiple genes in A. thaliana and other organisms. The result of this investigation collapses in a graphical representation of a gene network built by different edges, each one describing the nature and the weight of the relation shared by two or more elements. The collection of datasets exploited origins from literature and publicly databases as PFAM, Expression Omnibus (GEO), BioGRID, I2D, Pathway Commons, Memorial Sloan-Kettering Cancer Center, Human Protein Reference Database, HumanCyc,

Systems Biology Center New York, IntAct, MINT, INTERPRO, NCI-Nature Pathway Interaction, iRefIndex and Reactome. A personal dataset can be uploaded too for users who want to customize or integrate their output results. So, the first step of Genemania query form is the definition of the dataset(s) to exploit in order to infer the relations among a group of genes. About 215 resources among these just described above are available for *Arabidopsis*. The next step is to define the network weighting and Genemania offers 3 different set of choices: a query dependent weighting, a Gene Ontology based method or a "based on equal weighting" set of preference. Results page offers the previously described gene network with each edge colored according to its criteria of relation and with a percent value describing its contribution in gene-to-gene association. Genes tab on the right shows the cellular function(s) associated to each element of the network, with a list of possible synonymous genes, while the function tab allows to visualize globally in the graph, all the genes associated with one or more cellular process. Gene function association is statistically supported with an FDR value and for each process a coverage indication is available too, which is equal to the number of query elements found in the network compared to the size of the full list of genes associated to that particular function.

2.3.11 Genenvestigator

Probably the most cited resource as bioinformatic resource since 2004, Genenvestigator [Zimmermann et al, 2004] collects biomedical and plant biology genomics data of the most studied organisms, like Hordeum vulgare, Oryza sativa, Zea mais, Triticum aestivum, Solanum lycopersicum, Physcomitrella patens, Nicotiana tabacum and in particular Arabidopsis thaliana. Datasets collected for expression analysis of this latter come from Atgen Express, FGCZ, GEO, TAIR,

ArrayExpress, Nasc, Gruissem Lab and other, while the chip platforms used for microarrays are the Ath1 22k, with 9848 experimental microarray slides available, the Ag 8k with 92 and the Agro1 whole genomic tilling array with 53 slides respectively. In order to compare these different studies, all experiments have been preprocessed by RMA normalization. So, The first step of a co-expression analysis in the Genenvestigator query is the definition of a fully customizable list of platforms and datasets assortments, with the possibility to choose and relate single tissue or experiment combinations too if preferred. The instruments available for gene co expression analyses are organized in three main collections: a condition search tools set, which allows visualizing the expression variation of the query genes among tissue, developmental studies or targeted experiments; a gene search tools set which instead extracts gene expression variation by defining the particular experiment conditions in study and a similarity set, focused on finding co-expressed genes within a specified dataset, subsets or comparing a customized gene list.

The Condition search tools set

This suite of tools let the user visualize the expression value of the query genes through a scatter plot or a heat map representation measured with a log2 or a linear scale. In the sample tool, the expression values of each gene of interest is shown according to the samples and datasets chosen during the query; in the Anatomy tool the expression values are organized in tissues instead, while in the Development tool, data follows the plant development stages. The last tool, the Pertubations one, shows the highest expression changes among the genes and samples of interest, indexed according to the type of stress, chemical, mutation or stimulus present among the experiments

considered. All the tools just described are followed and filterable with statistical info as p-value, fold change and number of samples involved.

The Gene search tools

All the tools of this section aim to identify one or a group of gene with a relevant expression in a specific condition. The RefGene tool in particular helps to define stable reference genes for qt-r PCR analysis, choosing the sample(s) with the most similar condition(s) to the gene(s) of interest and defining a signal intensity range in log 2 scales, needed to filter out the results. The Anatomy section displays the most expressed genes in one or more tissues, and moreover can underline the ones with the highest different expression values among a customized tissue dataset. Following the same scheme, but applying it on development stages, the Development tool identify the most expressed genes in one or more stages of growth, with the chance to define a comparison among these latter too. The perturbation tool in this set works in the opposite way from the Condition one described previously, showing the most up regulated or down regulated genes in a user defined list of experiments or in comparison between themselves.

The similarity search tools set on Genenvestigator allows identifying group of genes gathered by their expression profiles. The hierarchical clustering i.e. tool offer several ways to visualize genes association according to the distribution of these latter among samples, tissues and development stages or perturbations schemes. In a similar manner, user can cluster factors instead of gene, in order to identify expression trend shared by two or more samples and, moreover, genes and factors can be clustered together too to obtain the elements with the most similar expression profile for both aspects. All results can be measured with the common clustering parameters like the Euclidean and the Manhattan

distance or the PC value. A sharper approach to cluster query genes in relation to the biological aspect considered is available in the biclustering tool and is based on the Bimax algorithm. After choosing the factor to investigate in a user defined dataset, it is possible to search for cluster able to satisfy desired conditions as the smallest number of genes to hold within (min. probe sets), the smallest number of samples or factor elements to consider (min. factors), a minimum expression value and a minimum up or down regulation degree if a perturbation scheme is the chosen factor at the beginning. Co-expression tool completes the similarity search suite aiming to identify the most co-expressed genes with a query one. Co-expression is measured with the Pearson correlation on the log2 transformed values of the dataset chosen and the results are depicted with a circular hierarchical clustering collecting the query gene in the center and the co-expressed ones around, with distances from the former defined by their Pearson value. Moreover, as for the clustering tool, a factor defined subset can be chosen to restrict co-expression analyses only to genes characterizing specific tissues, samples or conditions, and another added values it is the chance to filter out co-expressed genes according to their mutual correlation value.

2.3.12 PlaNet

PlaNet (Planet Network) [Mutwil et al., 2011] is a network website for *Arabidopsis thaliana* and other eight species (barley, wheat, rice, poplar, medicago, soybean, brachypodium and tobacco). It stores a very large microarrays database for *Arabisopsis*, with 468 dual channel and 606 single channel microarrays slides obtained from TAIR. According to their microarray experimental conditions, all these data have been arranged in specific sets gathering in total 308 experiments, each one with at least 10, 5 arrays for dual channel and single channel

essay respectively. On PlaNet this website, there are a lot of useful features to evaluate *Arabidopsis* co-expressions: after choosing one or more genes for the query, as already seen in the previous databases, it is possible to observe the expression values variation among several tissues and/or experimental conditions. But the core of PlaNet database is its network tools package, based on the Highest Reciprocal Ranking (HRR) and on the Heuristic Cluster Chiseling Algorithm (HCCA). HRR (Highest Reciprocal Rank) is a variant of the Mutual Ranking algorithm seen in ATTED-II and it expresses the correlation strength between two genes, not through the geometric average between their rank positions in a mutual PC list, but by the highest rank position in these latter. So correlations between genes are graphed in a network by green, orange or red edges if HRR<=10, 10<HRR<=20, 20<HRR<=30, respectively. By keeping in a graph all genes within *n* steps away from the query gene, PlaNet offers a simple but powerful cluster representation, called node vicinity network (NVN). This latter offers a quick graph of the most related genes with the query and, together with the HRR, this is the core of the HCCA: this algorithm is articulated in five steps: 1) a NVN is filtered out for each elements of an HRR network, by keeping only the elements with a neighborhood of n (3 generally) steps away;2) then the genes showing more connections to the outside of their NVN than within this latter are removed, 3) the so called stable putative clusters (SPC) are generated which in turn are 4) selected and filtered out according to their size, and ratio between their "in and out of NVN" number of edges. The main result of HCCA is the Meta Network page on PlaNet, which shows, in a comprehensive manner, pre-calculated best fitted clusters of correlated genes, in order to explore *Arabidopsis* transcriptome in the fastest way, or let the user to individuate the best pre calculated cluster which contains a query gene.

2.4 Software for clique analysis

The software exploited to perform the clique clustering was available as Matlab script at (http://www.mathworks.com/matlabcentral/fileexchange/30413-bron-kerbosch-maximal-clique-finding-algorithm). The analysis was performed on a dataset composed only by genes having at least one significant Pearson's correlation with $r \geq |0.95|$ exploiting the dataset described in chapter 2.1. Tha machine exploited was a Fujitsu Celsius server m450 (4 cores, 2.33Ghz each one, 4 Gb RAM, 300Gb drive).

2.5 Social network algorithm

The algorithm proposed [Gallo C. et al, 2011] has been exploited in our analyses to identify key genes in groups of co-expressed genes. Here follows a description of the algorithm used:

1) The Pearson's coefficients calculated among 8 genes, for example, are collected in the matrix "C" of 8x8 cells, as shown in figure 9.

	a	b	c	d	e	f	g	h
a	1	0.8	0.45	0.51	0.82	0.74	0.37	0.22
b	0.8	1	0.91	0.92	0.81	0.3	0.31	0.33
c	0.45	0.91	1	0.77	0.83	0.34	0.26	0.23
d	0.51	0.92	0.77	1	0.94	0.41	0.42	0.1
e	0.82	0.81	0.83	0.94	1	0.35	0.33	0.05
f	0.74	0.3	0.34	0.41	0.35	1	0.28	0.07
g	0.37	0.31	0.26	0.42	0.33	0.28	1	0.07
h	0.22	0.33	0.23	0.1	0.05	0.07	0.07	1

Figure 9 10x10 matrix "C" storing the Pearson's correlation values obtained in our example

2) A threshold, defined as "θ_1", equivalent to the Pearson coefficient "*r*" has been chosen to filter the pairs of genes from the matrix "C". In this example θ_1 has been set at $\theta_1 \geq |0.70|$, hence only the pairs of genes with a correlation above this value have been considered. Drawing the genes as nodes, and the correlations as edges, we have

hence produced the followed unweighted graph built with the nodes: a,b,c,d,e,f (Fig. 10).

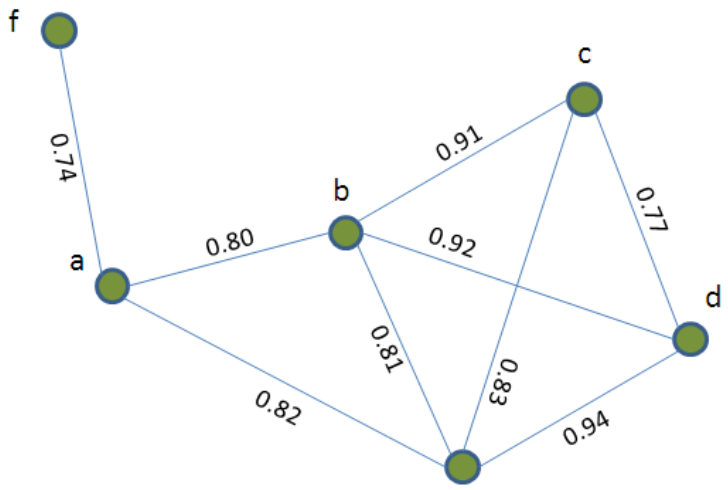


Figure 10 Graph obtained establishing the edges between the nodes, when $\theta_1 \geq |0.70$.

3) The next step aims to refine the structure. For each node in the graph we have calculated its "social index", defined as "i" which is equal to the numbers of triangles available considering the node as one of the vertex, against the total number of possible triangles considering the node and its two closest edges around. For example the social index of node b is i=4/6=0.66 because 4 triangles are available out of 6 possible (Fig. 11), the social index I of thenode c is i=3/3=1.
In this way the social indexes i for each node in the graph are:
c=3/3=1; b=4/6=0.6; d=3/3=1; e=4/6=0.6; a=1/3=0.3; f=0;

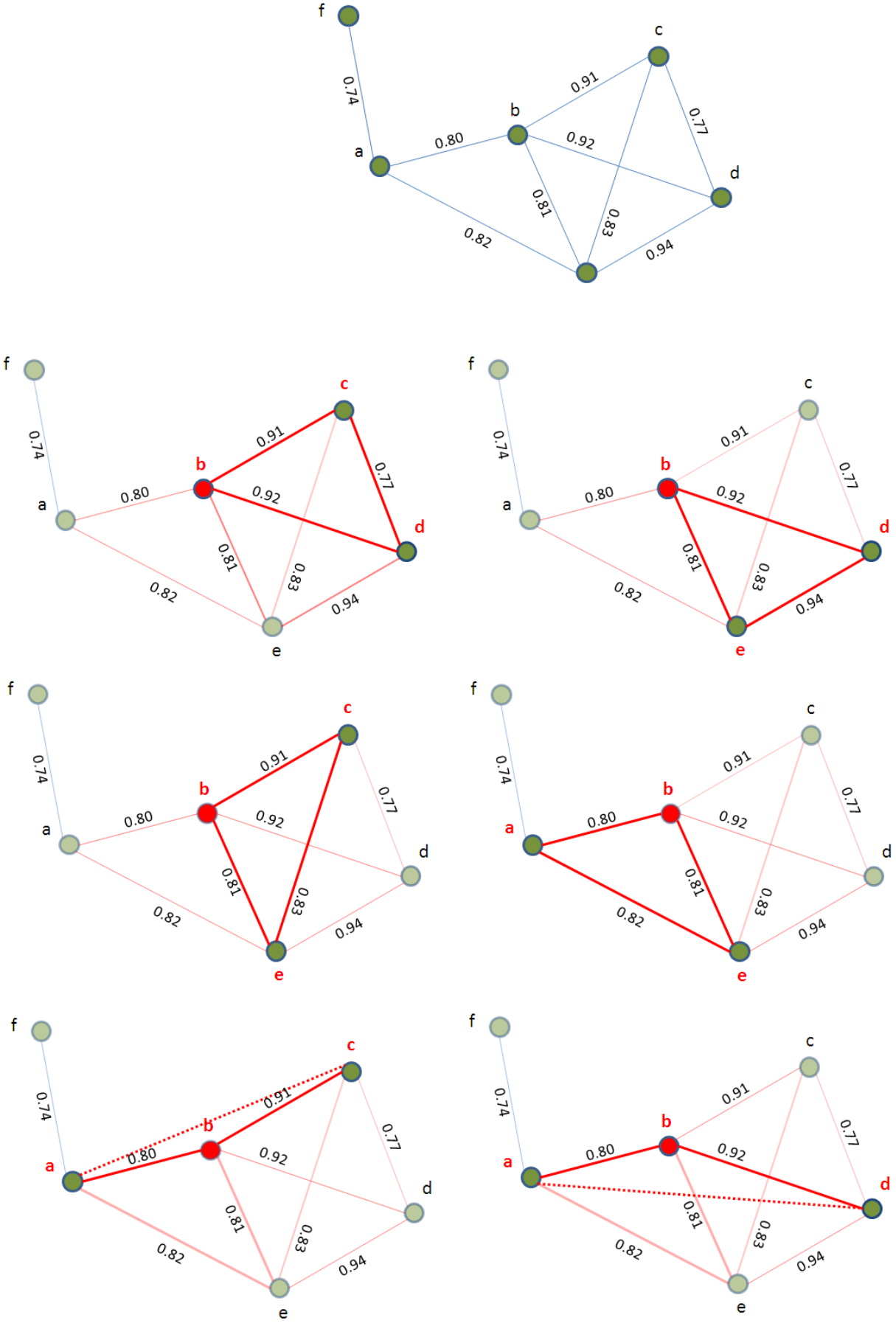


Figure 11 The node b has a social index i=4/6=0.66.because it has 4 established triangles in the graph out of 6 possible considering the node b and its two closest edges around.

4) We remove all the nodes with a social index below a specific threshold, indicated as θ_2, which in our example is $\theta_2 \geq 0.5$ (Fig. 12)

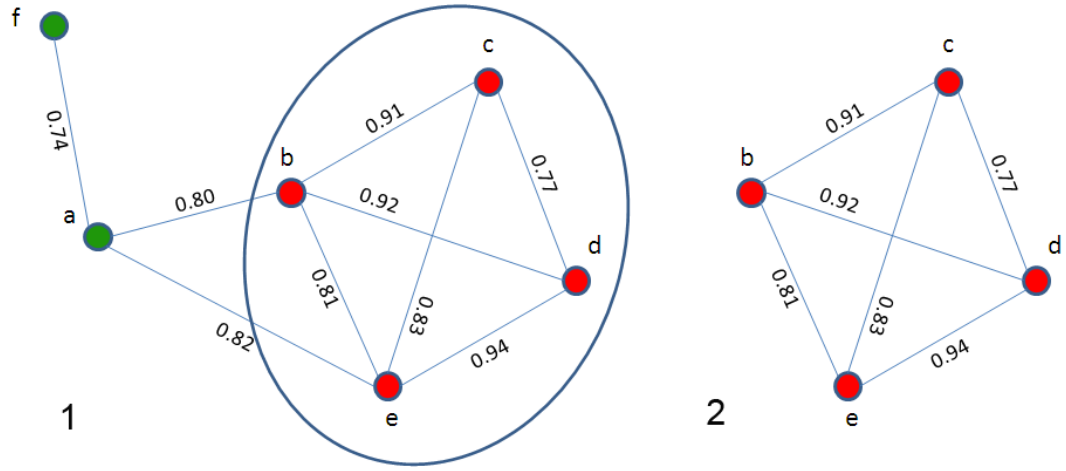


Figure 12 The threshold θ_2=0.5 removes the genes a,f

As shown in figure 10.2 the refined graph is built only with the nodes b, c,d,e. Summarizing this variable i should measure the consistence of interactions with its co-expressed partners, reflecting its ”popularity”, and for this reason we define this algorithm the “social network algorithm”. Considering the nodes as genes and the edges as co-expression, we defined the graphs obtained with this approach as networks of co-expression.

2.6 GO terms enrichment

The research of specific functions inside the group of co-expressed genes was performed through the GOrilla software [Eden E., 2009]. This platform exploits the GO terms dictionaries, collections of terms describing the biological functions and assigned to specific genes of an organism, to underline the presence of a significant concentration of genes sharing the same functions.

On the Gorilla platform the enrichment is calculated as:

Enrichment = (b/n) / (B/N)

where the cited variables are:

N - the total number of gene dataset

B - the total number of genes associated with a specific GO term

n - the number of genes inside the group in which we are calculating the enrichment

b - the number of genes inside the analyzed group associated with a specific GO term

The p-value for each GO term enrichment is available in the results, and it ranks the significance of the enrichment found in five levels: $>10^{-3}$, 10^{-3} to 10^{-5}, 10^{-5} to 10^{-7}, 10^{-7} to 10^{-9}, $<10^{-9}$. Lower is the p-value, more significance has the result. This is also underlined with the help of a color scheme in the results (Fig. 13)

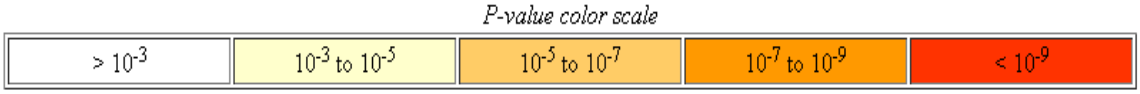


Figure 13 Color schemes describing the p-value significance in the GOrilla results.

Only the enrichments with a p-value below 10^{-7} were considered significant in this work.

Results

3.1 Validation of microarray platform

Arabidopsis thaliana has been chosen as a reference species in plant biology because of its genome described as fully notated in the 2000 [The Arabidopsis Genome Initiative, 2000]. While checking the Affymetrix microarray probes exploited during our works we noticed an important discrepancy between the available probes and the number of the protein gene described in the Tair 9 annotation (ftp://ftp.arabidopsis.org/ Genes/TAIR9_genome_release/). As shown in table 3 the number of genes in the TAIR 9 are 33239, with a total coding protein number of 27169 (81.74%), however the number of the probes with a signal detected in our experiments, on the reference chip Affymetrix,, is only 21769/27169, only the 80% of the protein coding set. Hence we are still a little bit far from having the complete snapshot of the whole transcriptome. Moreover, probes are generally designed planning to assign for each gene, only one specific probe, but on the chipsets exploited, 387 probes from the 21769 available, are not gene specific, so one gene can be matched by more than one probes and this fact decreases the number of gene expression profiles checkable to 21382. This is relevant because some analyses cannot be exhaustive as hoped due to the lacking of important probes as the one for the CESA8 gene (AT4G18780), whose co-expression with CESA7 (AT5G17420) and CESA4 (AT5G44030) to perform the plant cell wall synthesis, is experimentally confirmed [Eckardt N. A. et al, 2003] and often exploited for data validation. Hence the results aiming to obtain a complete information from the transcriptome must be accepted with some reserves, since 20% of the protein coding genes are without probes in *A. thaliana*.

Genes annotated in TAIR9	**33239**	**(%)**
-Total protein coding genes	27169	81.7383
-Other genes (transposable, unknow, misc_rna…)	6070	18.2617
Total protein coding genes	**27169**	**(%)**
-Genes with probes	22810	83.9559
-Genes without probes	4359	19.8755
Probes with signal	**21769**	**(%)**
-Specific probes	21382	98.2222
-Not specific probes (multiple genes)	387	1.7777

Table 3 Distribution of probes availability in *Arabidopsis thaliana* transcriptome and according to the paralog structure of the genome.

3.2 Analysis of genes correlated by expression patterns

In order to discover the genes straightly or inversely related inside the genome of *Arabidopsis thaliana*, as first step we have calculated the Pearson correlation coefficient (*r)* between each pairs of genes available from our collection of 20908 genes, exploiting the expression profiles of all the 79 samples inside the Developmental Series Expression atlas of Arabidopsis development" (http://affymetrix.arabidopsis.info /narrays/experimentbrowse.pl) (see "Data collection in material and methods). In this way, we have obtained the *r* values for 437144464 pairs of genes.

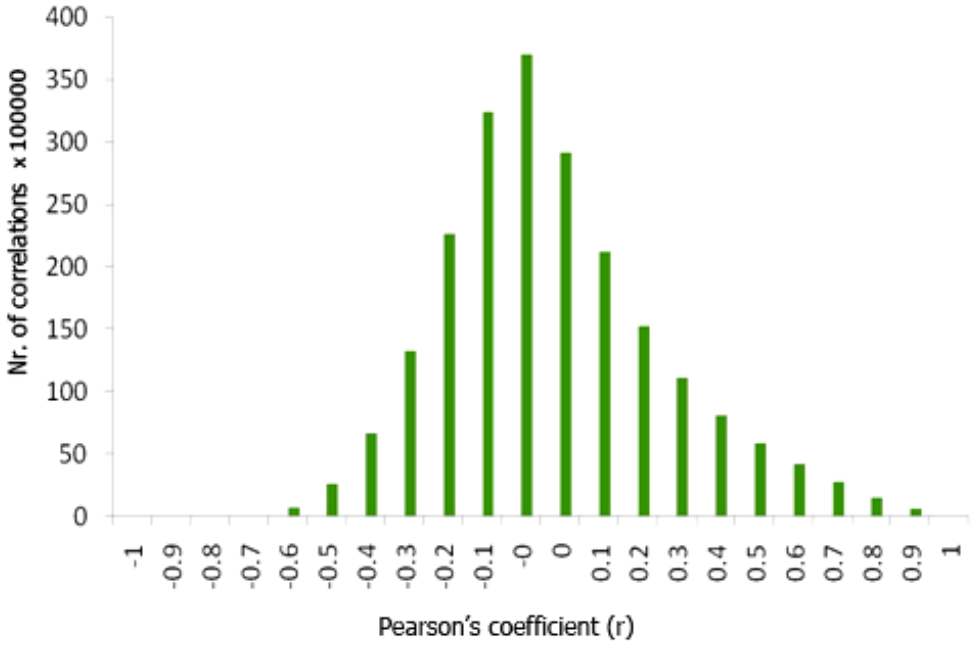


Figure 14 Distribution of Pearson coefficients among the correlations obtained with a matrix of 20908x20908 genes

In the Pearson test, an *r* value close to 1 asses that two genes are expressed together, while *r*=-1 asses that the presence of one is associated to the absence of the other. However, as shown in figure 14

most of the correlations calculated has an *r* value close to 0, which means that correlations for these genes are not present. With the aim of extracting only the meaningful relations among the genes, we have filtered from the Pearson's test results only the correlations with $r \geq |0.7|$. We confirmed the significance of this threshold according to the exploitation of the 5^{th} percentile strategy [Blanc G. et al, 2004], based on the fact that the 95% of gene correlations in our results has an *r*-value under |0.7| and hence there is a probability of mistake of 5% (α=0.05), which is commonly accepted. The Pearson test on our collection of 20908 genes has found 3861300 correlations with $r \geq |0.7|$. Although this value represents 1,77% of the whole correlations (437144464), it is interesting to underline that 18136/20908 elements, 87% of the considered genes, establish at least one relationship with another gene. This information underlines that the *Arabidopsis* genome expression is very complex as expected, with a huge number of coordinated patterns of expression.. Checking the distribution of the number of correlations per gene, we have investigated the relationship between number of correlations and functionality, by splitting our datasets in six percentiles (Fig. 15), each one containing 3022 genes. So as shown also in figure 15, in the first percentile the number of correlations per gene reaches the maximum value of 26, while in the second percentile the number of correlations per gene reaches the maximum value of 107. A GO enrichment analysis on these percentiles through the GOrilla software [Eden E., 2009] (see chapter 2.6) has highlighted that the "rhythmic process" and "transport activities" are the most shared terms among these two groups (Fig. 16-17). This result underlines that basilar cell processes do not require a strong regulation, hence very few genes are cooperating to perform these tasks.

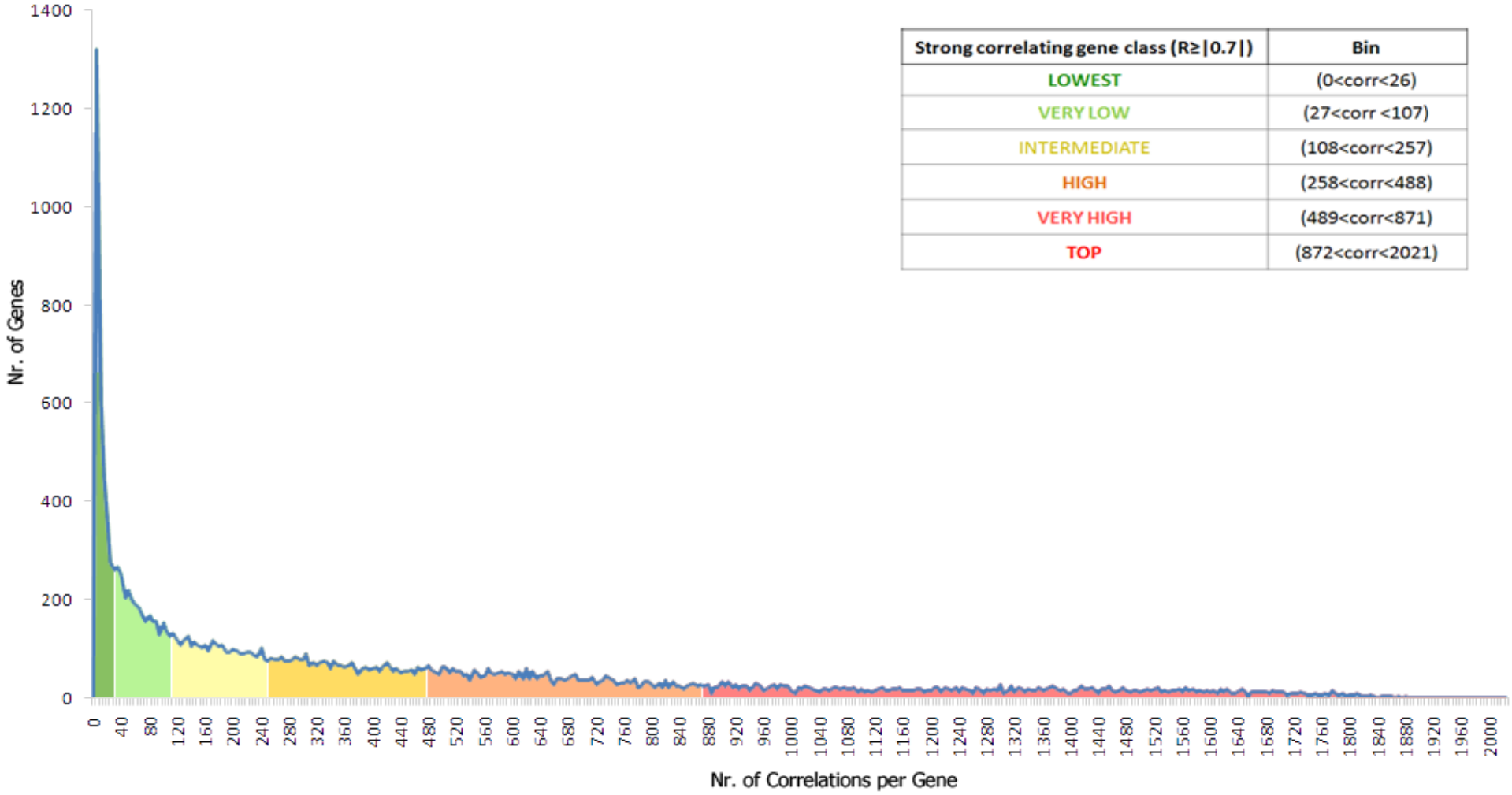


Figure 15 Distribution of correlations per gene. We have kept 18136 correlated genes (R≥|0.7|) and then divided them in 6 percentiles based on the distribution of the number of correlations established by each gene, shown in colors.

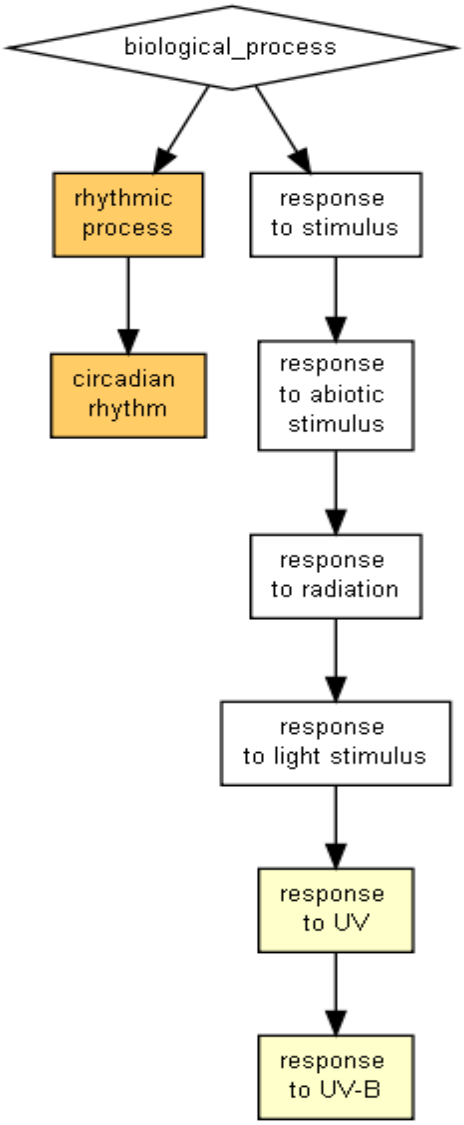
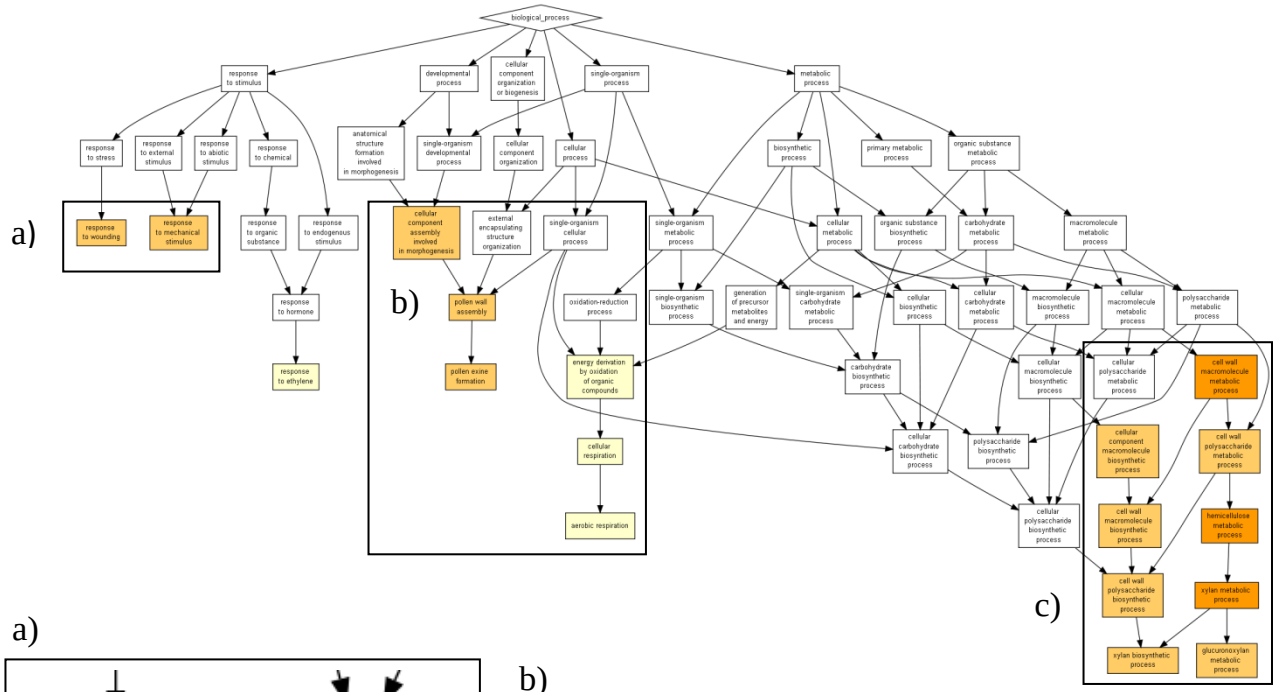


Figure 16 Process GO Terms in the 1° Percentile.



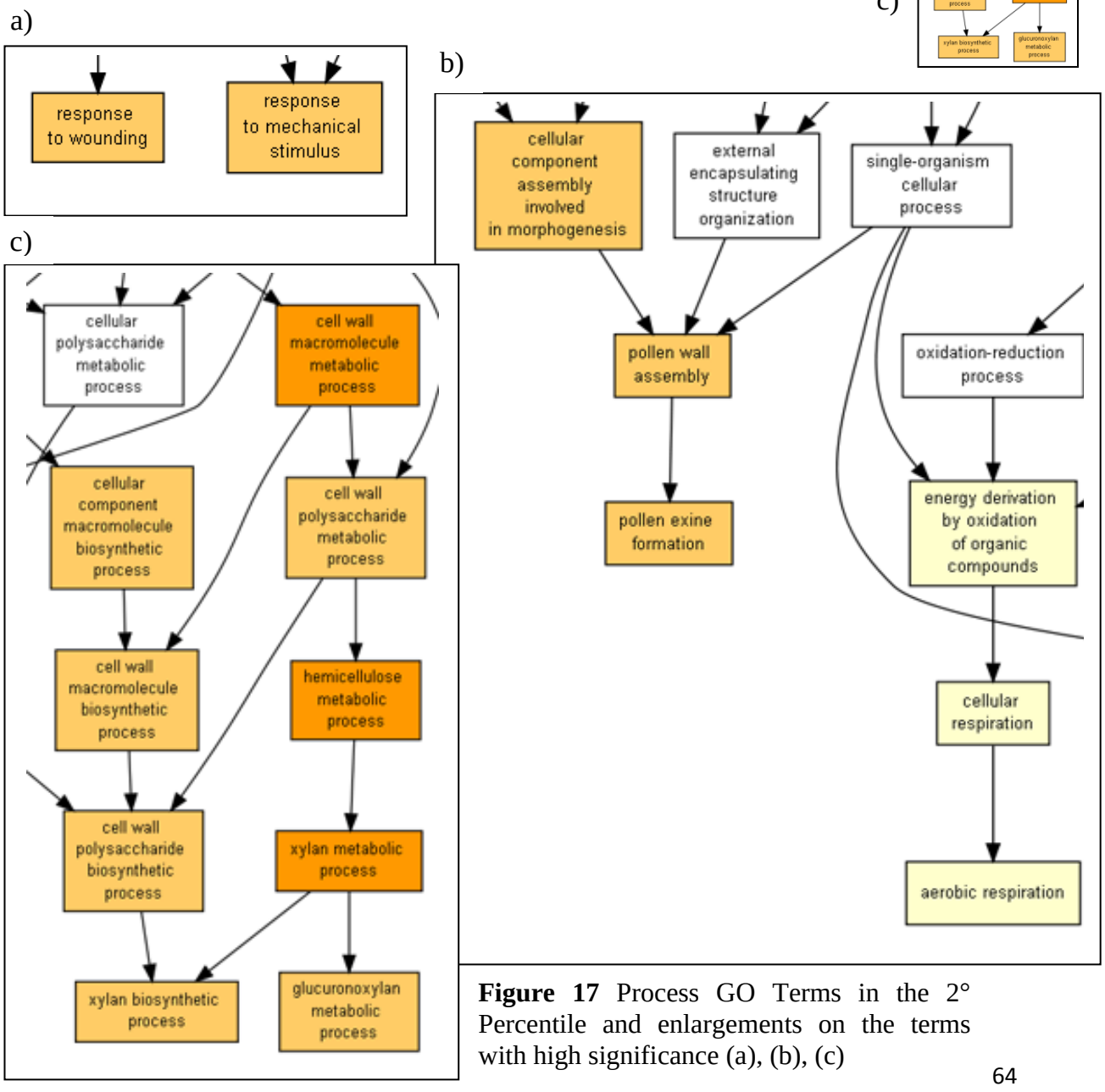


Figure 17 Process GO Terms in the 2° Percentile and enlargements on the terms with high significance (a), (b), (c)

We expected that genes with simple functionalities need less cooperation and therefore the number of correlations per gene could be lower. As the number of correlations per gene increases, there is a constant growing up of GO terms related to “nucleic acids metabolism”, “methilation” and “nuclear localization”. The GO terms enrichment of the sixth percentile, shown in figure 18, is referring to the ribosome complexes and transcriptional factors, underlining how fundamental cell life regulation activities are managed by genes interacting with many others in the cells (Fig. 18-20).

A whole list of GO terms for the six percentile is provided in the Appendix.

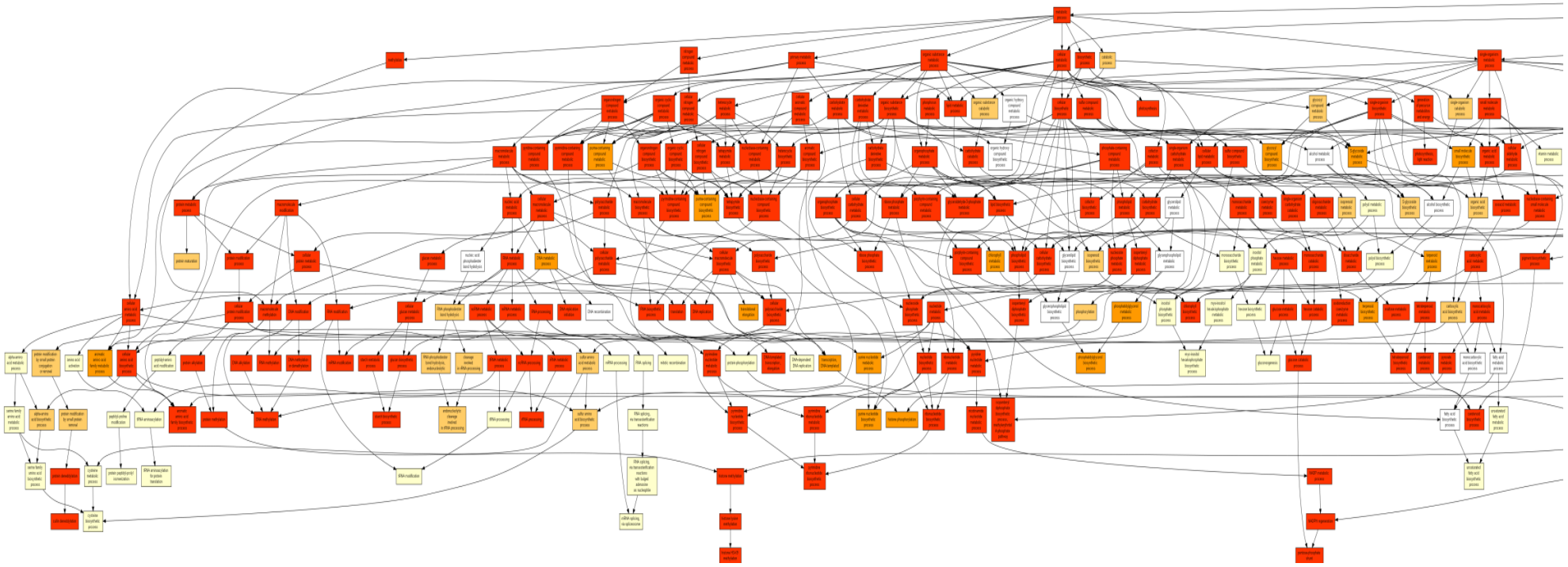


Figure 18.a First part of the landscape of GO terms enrichment for process in the 6° percentile (827<corr per gene<2027). Red boxes are associated to GO terms with high significance (p-value<10^{-9}) in the enrichment.

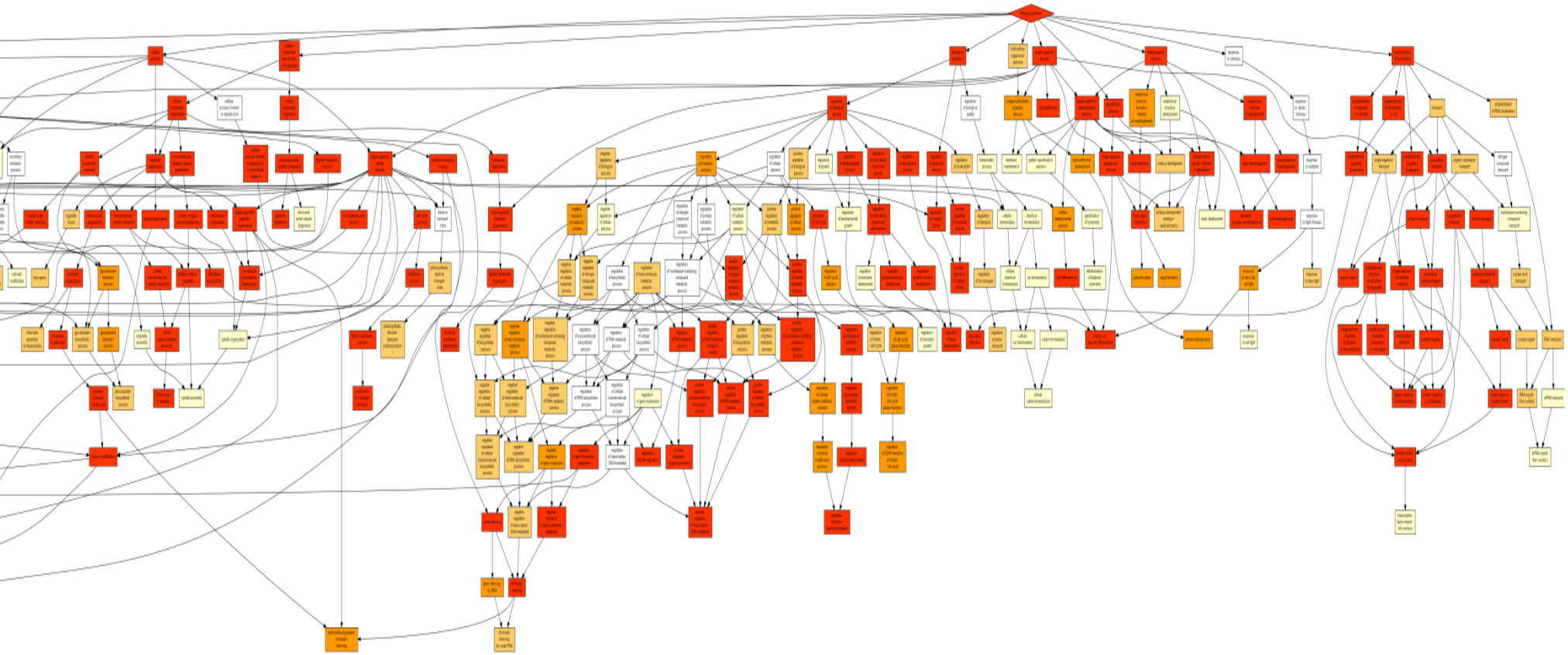


Figure 18.b Second part of GO TERMS enrichment landscape. Red boxes are associated to GO terms with high significance (p-value<10^{-9}) in the enrichment.

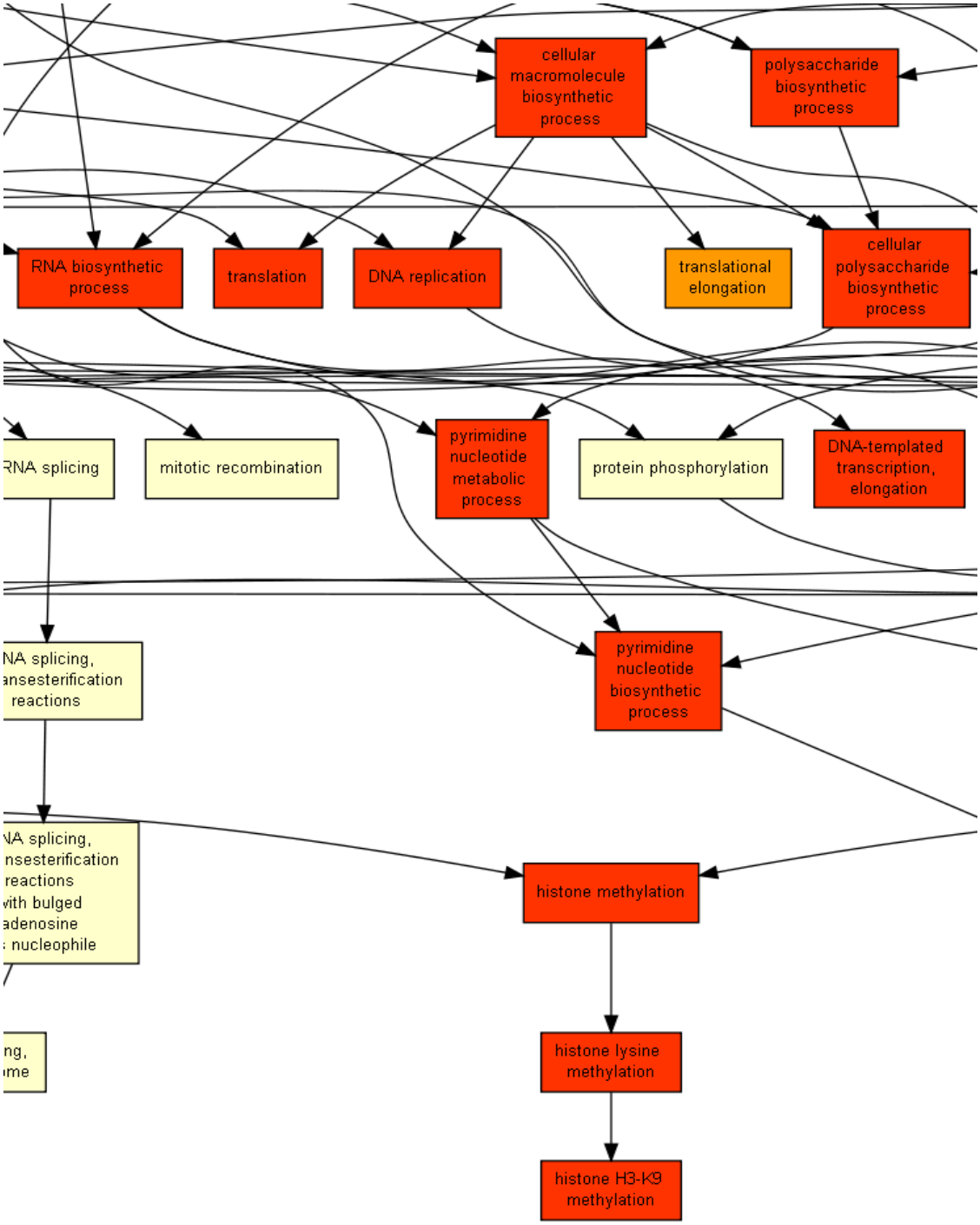


Figure 19.a In the 6° percentile of GO TERMS for process, the number of significant correlations per gene is very high. Red boxes are associated to GO terms with high significance (p-value<10^{-9}) in the enrichment. Orange boxes have a lower p-value enrichment, between 10^{-7} and 10^{-9}

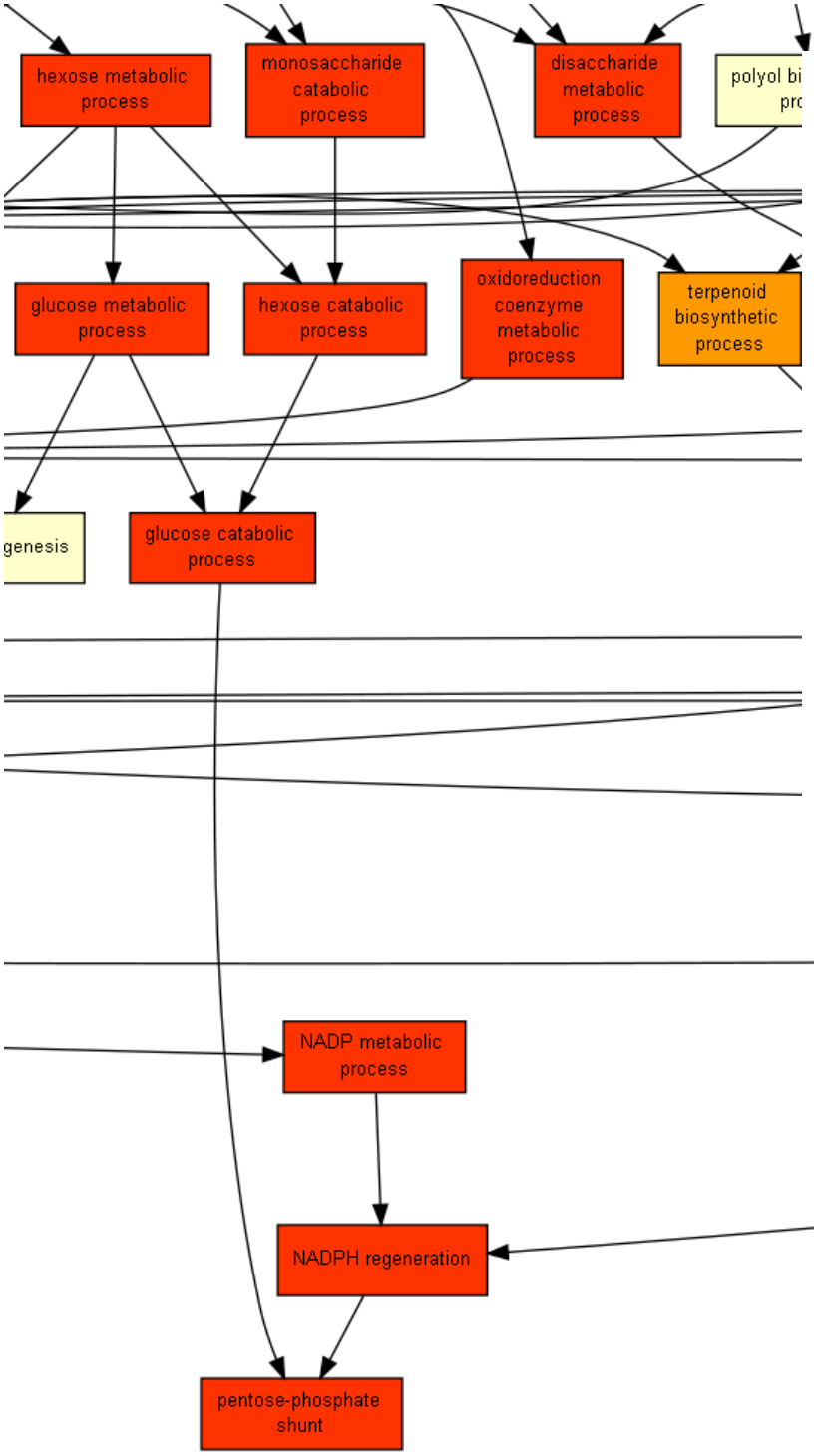


Figure 19.b Red boxes are associated to GO terms with high significance (p-value<10^{-9}) in the enrichment.

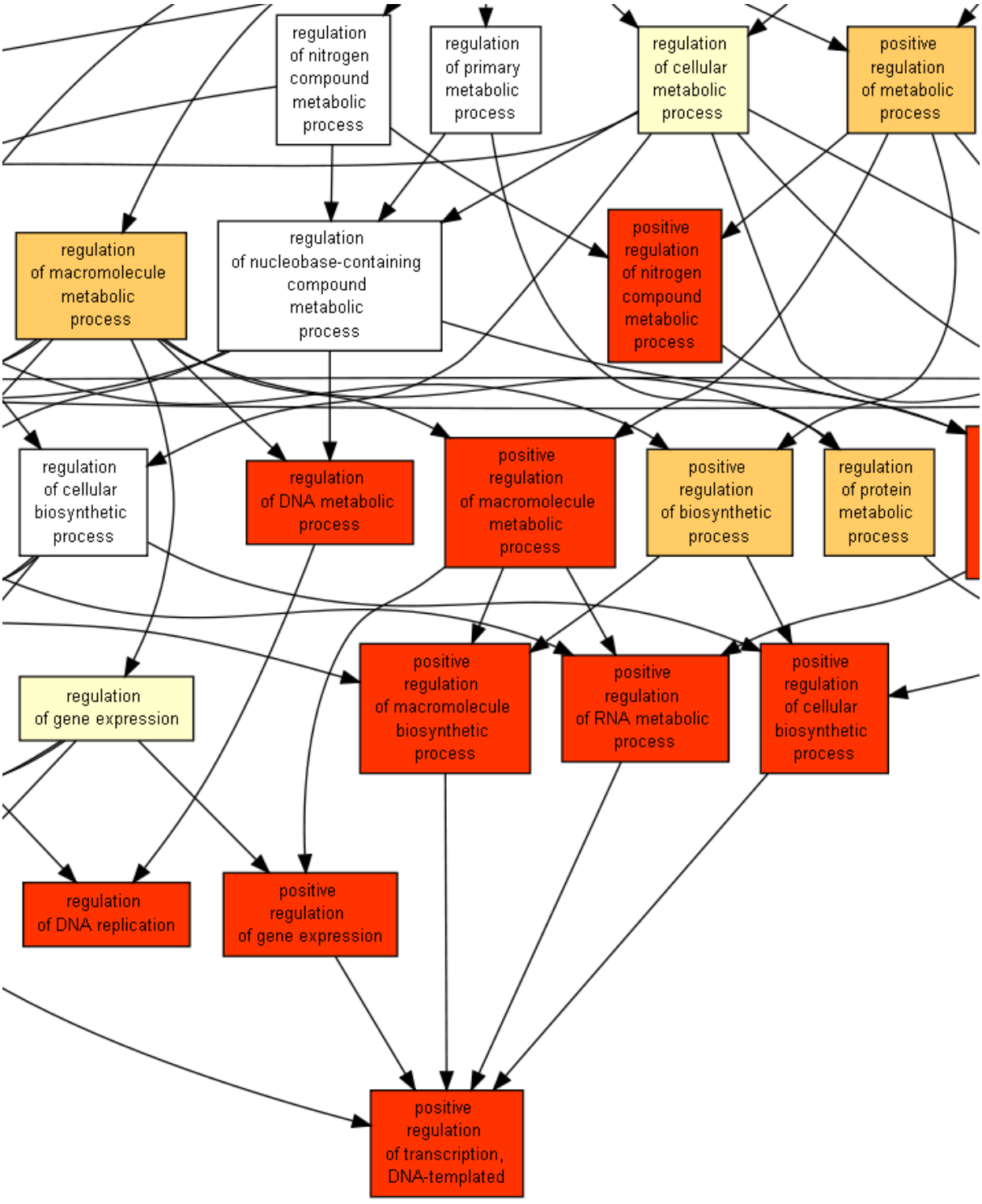


Figure 20.a Red boxes are associated to GO terms with high significance (p-value<10^{-9}) in the enrichment. Orange boxes have a lower p-value enrichment, between 10^{-7} and 10^{-9} and so on.

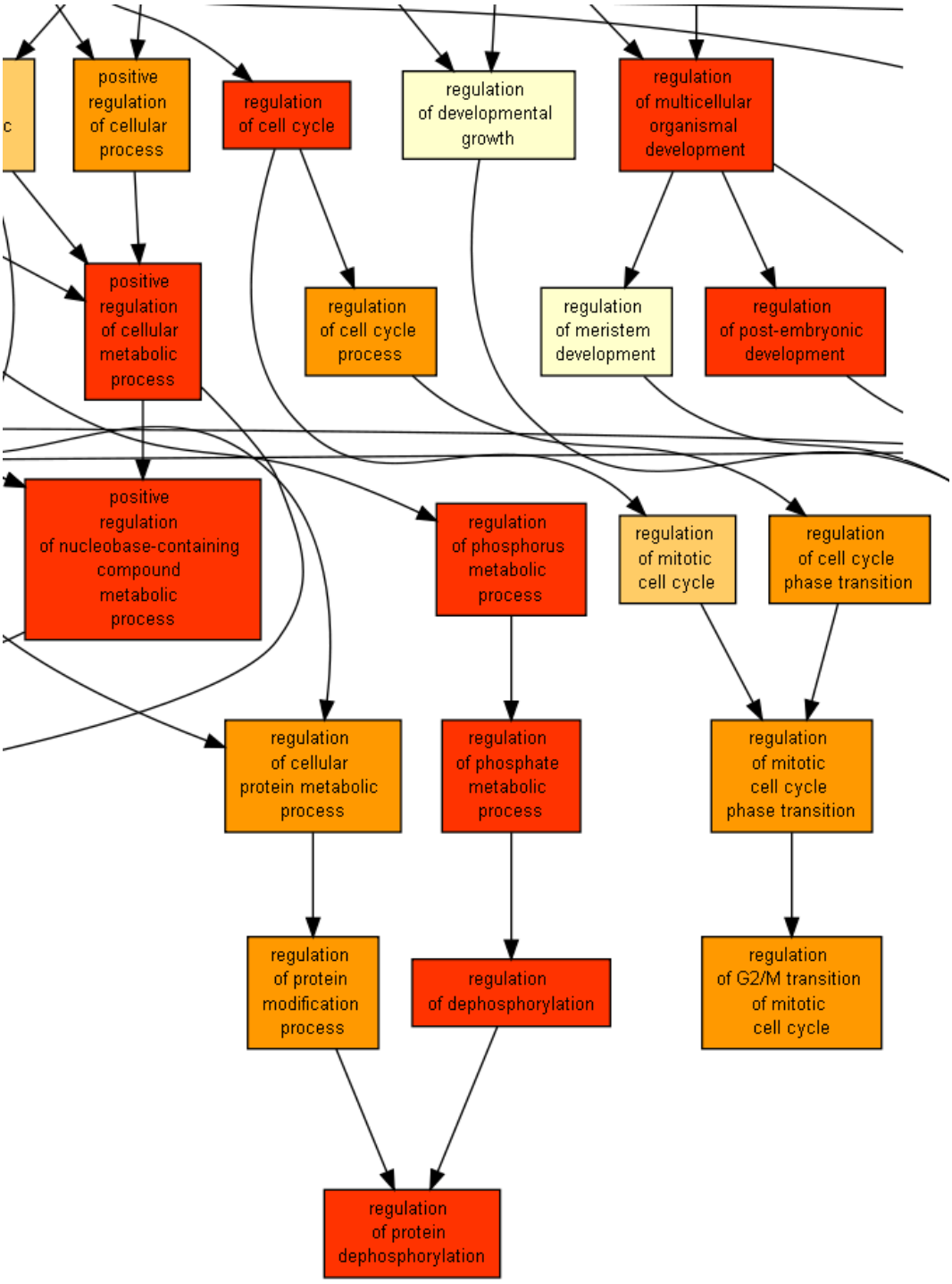


Figure 20.b GO terms of the 6th percentile. Red boxes are associated to GO terms with high significance (p-value<10^{-9}) in the enrichment. Orange boxes have a lower p-value enrichment, between 10^{-7} and 10^{-9}

3.3 Network of co-expression

We have used our gene library of 20908 elements to define gene networks (see chapter 1.2), where each node has at least one correlation with at least one member of the same network, at a specific Pearson's value. In these networks each node can belong only to one network. Using different Pearson's values as thresholds, from r≥|0.70| to |0.99|, with step of 0.1, we have obtained different classes of genes that were organized in networks. In figure 21 each cell contains for each gene the Id number specifying the network to whom each gene belongs to. While moving at predefined thresholds from the lowest Pearson's value of |0.70| to the higher one of |0.90|, the number of networks increases (Tab. 4, Fig. 22) and these latter become slimmer. As shown in tables 5a, 5b and figure 22, using a Pearson's correlation threshold of |0.70|, five small clusters are defined (yellow boxes in tables 5) and a very big one of 18125 genes is defined too. This latter decreases its size while reaching very high Pearson's correlation value. The decrease of the networks number and their size reduction after reaching a Pearson's correlation of |0.90| is

Pearson's coefficient	Nr. of networks
0.7	6
0.71	13
0.72	19
0.73	21
0.74	26
0.75	20
0.76	22
0.77	30
0.78	37
0.79	43
0.8	43
0.81	61
0.82	72
0.83	101
0.84	115
0.85	131
0.86	148
0.87	166
0.88	211
0.89	221
0.9	236
0.91	226
0.92	213
0.93	181
0.94	120
0.95	72
0.96	75
0.97	52
0.98	38
0.99	10

Table 4 Number of clusters according to the Pearson's thresholds.

explained by the fact that correlations are becoming fewer in comparison to the ones established at low Pearson’s coefficients.

Pearson's correlation (r)

Gene	0.7	0.71	0.72	0.73	0.74	0.75	0.76	0.77	0.78	0.79	0.8	0.81	0.82	0.83	0.84	0.85	0.86	0.87	0.88	0.89	0.9	0.91	0.92	0.93	0.94	0.95	0.96	0.97	0.98	0.99
AT2G29370	1	1	1	1	1	1	1	1	2	4	5	41																		
AT1G61800	1	1	1	1	1	1	1	1	2	4	5	41																		
AT3G17440	1	1	1	1	1	1	1	1	2	4	5																			
AT4G23210	1	1	1	1	1	1	1	1	1	41	40	55																		
AT4G23200	1	1	1	1	1	1	1	1	1	41	40	55																		
AT3G26590	1	1	1	1	1	1	1	1	1	38	35	48	56	79	92															
AT3G14760	1	1	1	1	1	1	1	1	1	38	35	48	56	79	92															
AT3G55560	1	1	1	1	1	1	1	1	1	37	34	47	55																	
AT1G33790	1	1	1	1	1	1	1	1	1	37	34	47	55																	
AT4G26270	1	1	1	1	1	1	1	1	1	35	31																			
AT5G64990	1	1	1	1	1	1	1	1	1	35	31																			
AT2G04750	1	1	1	1	1	1	1	1	1	32																				
AT4G08530	1	1	1	1	1	1	1	1	1	32																				
AT2G44080	1	1	1	1	1	1	1	1	1	31																				
AT3G59900	1	1	1	1	1	1	1	1	1	31																				
AT4G34590	1	1	1	1	1	1	1	1	1	30																				
AT1G12220	1	1	1	1	1	1	1	1	1	30																				
AT2G21130	1	1	1	1	1	1	1	1	1	29																				
AT3G12470	1	1	1	1	1	1	1	1	1	29																				
AT1G48300	1	1	1	1	1	1	1	1	1	25	25																			
AT5G05250	1	1	1	1	1	1	1	1	1	25	25																			
AT1G09070	1	1	1	1	1	1	1	1	1	22	22	31	34																	
AT1G02660	1	1	1	1	1	1	1	1	1	22	22	31	34																	
AT4G14790	1	1	1	1	1	1	1	1	1	19																				
AT5G05850	1	1	1	1	1	1	1	1	1	19																				
AT3G49880	1	1	1	1	1	1	1	1	1	17	20																			
AT1G72990	1	1	1	1	1	1	1	1	1	17	20																			
AT1G80400	1	1	1	1	1	1	1	1	1	14																				
AT4G02340	1	1	1	1	1	1	1	1	1	14																				
AT2G27900	1	1	1	1	1	1	1	1	1	13	14	17	14																	
AT2G11240	1	1	1	1	1	1	1	1	1	13	14	17	14																	
AT1G59610	1	1	1	1	1	1	1	1	1	12																				
AT4G25360	1	1	1	1	1	1	1	1	1	12																				
AT2G15540	1	1	1	1	1	1	1	1	1	9	10	11																		
AT3G06280	1	1	1	1	1	1	1	1	1	9	10	11																		
AT1G09360	1	1	1	1	1	1	1	1	1	9	10	11																		
AT4G37120	1	1	1	1	1	1	1	1	1	5	18	19	17	33	33	44														
AT4G18140	1	1	1	1	1	1	1	1	1	5	18	19	17	33	33	44														
AT1G15710	1	1	1	1	1	1	1	1	1	5																				
AT3G23290	1	1	1	1	1	1	1	1	1	3	4	6	9	92																
AT2G42610	1	1	1	1	1	1	1	1	1	3	4	6	9	92																
AT2G31160	1	1	1	1	1	1	1	1	1	3	4	6	9																	
AT5G28490	1	1	1	1	1	1	1	1	1	3	4	6																		
AT5G08640	1	1	1	1	1	1	1	1	1	2	2	3	2	3	2	5	29	146												
AT5G62210	1	1	1	1	1	1	1	1	1	2	2	3	2	3	2	5	29	146												
AT1G65060	1	1	1	1	1	1	1	1	1	2	2	3	2	3	2	5	29													
AT3G55120	1	1	1	1	1	1	1	1	1	2	2	3	2	3	2	5	11	36	163	173	189	174	171							
AT3G51240	1	1	1	1	1	1	1	1	1	2	2	3	2	3	2	5	11	36	163	173	189	174	171							
AT5G13930	1	1	1	1	1	1	1	1	1	2	2	3	2	3	2	5	11	36												

Figure 21 In this picture we show an example of the clusters obtained with different Pearson correlation values. In the header there are the Pearson's coefficients used as threshold for clustering, in each cell there is the Id of the cluster to whom a gene belongs. As soon as the Pearson's coefficient get close to higher values, the networks begin to split in slimmer groups.

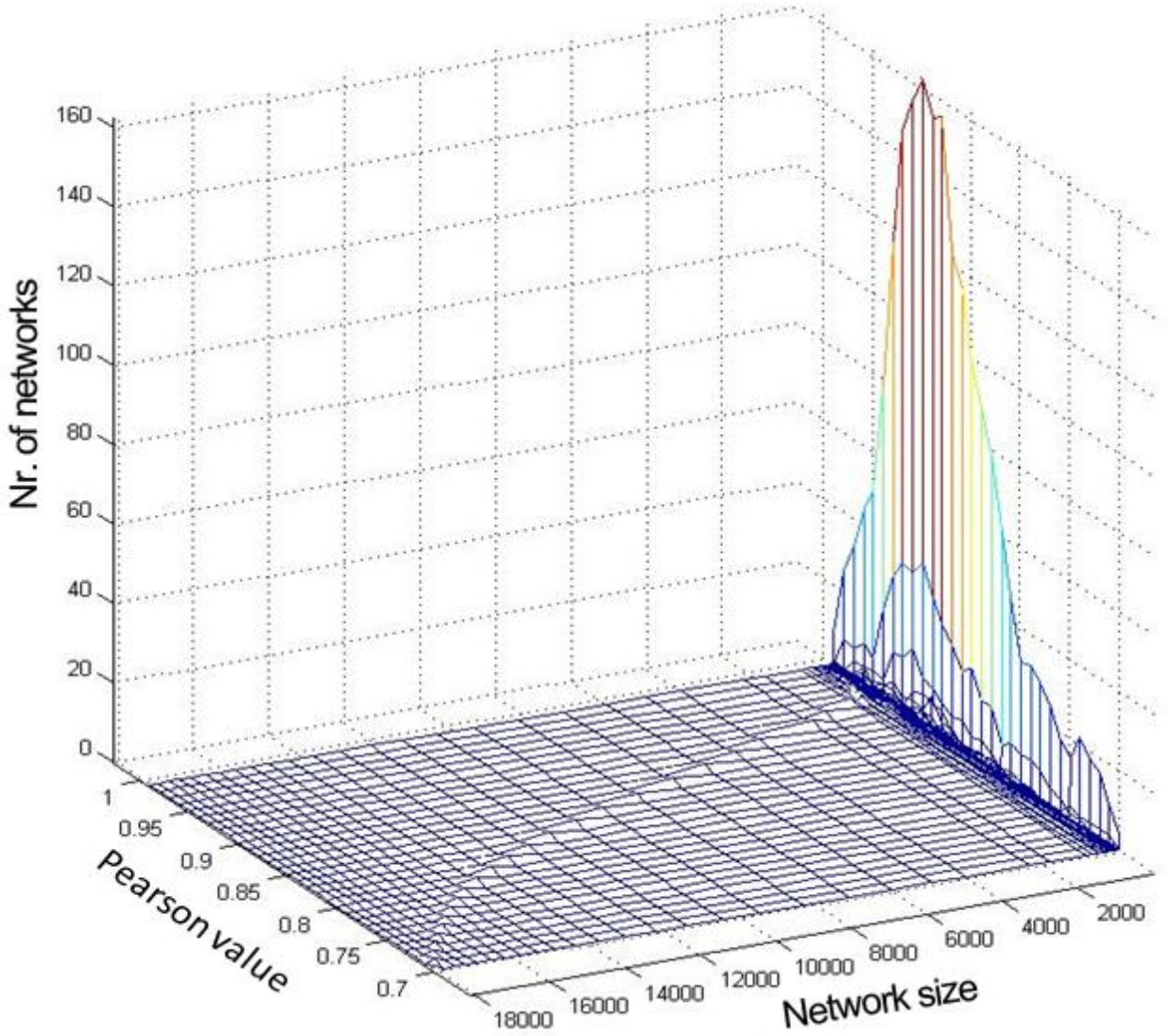


Figure 22.a The graph shows the change of the networks properties according to the Pearson's coefficient, analyzed through three different coordinates: size, r and nr. of networks. A detail of smaller networks, with the size proposed through a logarithmic scale is available in Fig. 19b.

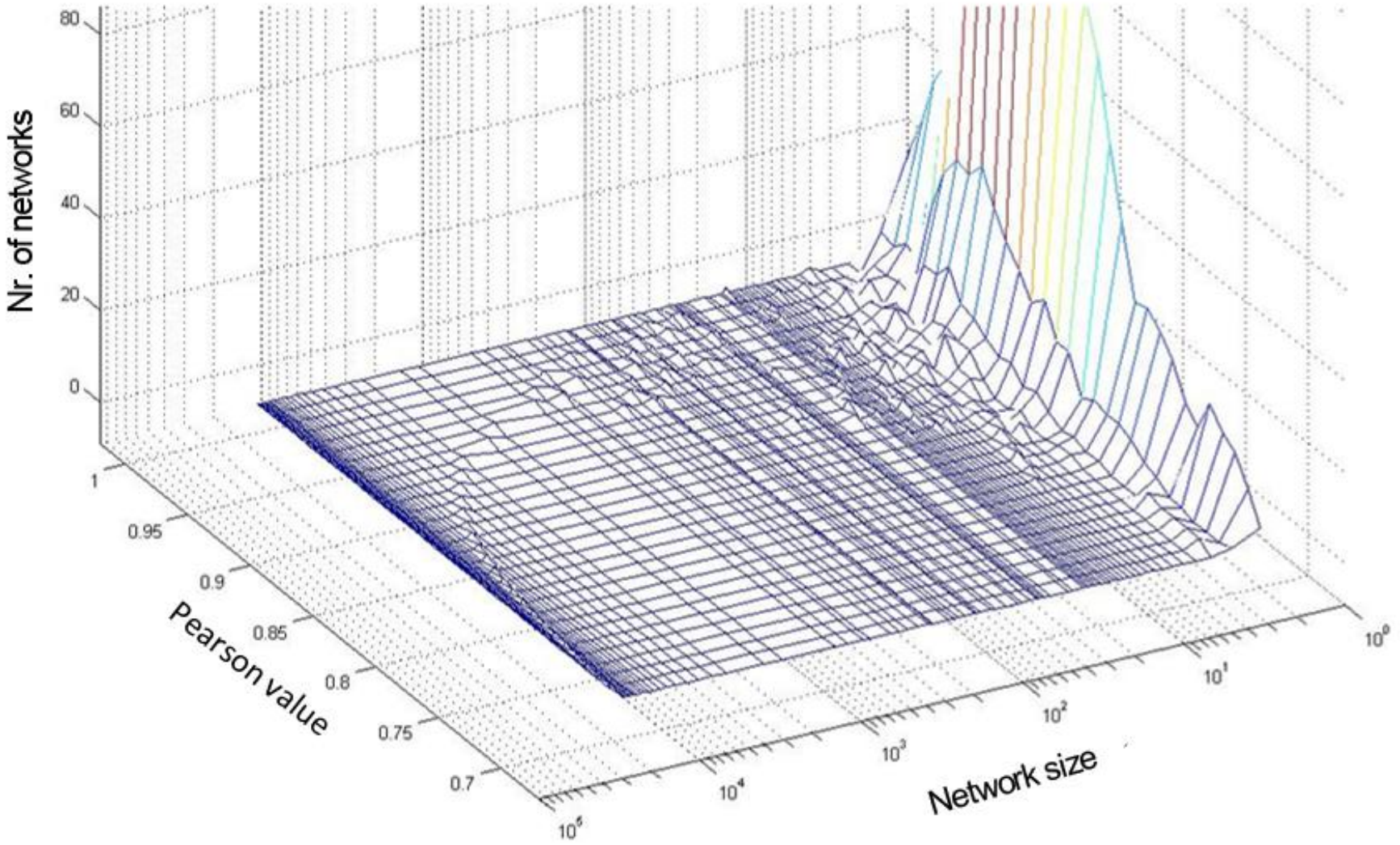


Figure 22.b A detail of smaller networks from figure 22.a, with the network size proposed through a logarithmic

Clusters Size	Pearson Correlation value																															
	0.7	0.71	0.72	0.73	0.74	0.75	0.76	0.77	0.78	0.79	0.8	0.81	0.82	0.83	0.84	0.85	0.86	0.87	0.88	0.89	0.9	0.91	0.92	0.93	0.94	0.95	0.96	0.97	0.98	0.99
2	4	10	16	18	22	15	18	25	29	32	31	46	61	78	89	99	116	123	156	154	162	155	146	116	78	50	44	33	25	8
3	1	0	0	0	1	1	3	4	7	8	8	9	7	15	15	22	19	24	28	33	40	36	37	32	22	8	10	7	8	0
4	0	2	2	2	2	3	0	0	0	1	1	2	1	3	5	4	7	6	8	10	11	17	13	13	4	5	4	4	0	0
5	0	0	0	0	0	0	0	0	0	0	1	1	0	0	2	0	0	2	6	6	7	6	1	3	3	1	4	1	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	3	3	8	3	4	3	5	3	2	1	1	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1	1	1	1	1	3	2	0	0	1	1	3	0	0	1
8	0	0	0	0	0	0	0	0	0	0	0	0	1	2	1	1	0	0	0	3	2	1	1	1	0	0	1	1	0	0
9	0	0	0	0	0	0	0	0	0	0	1	2	1	0	0	0	1	1	1	1	0	0	0	0	2	0	0	0	1	0
10	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1	1	1	2	1	2	2	1	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	2	0	0	0	1	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	3	0	0	1	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	1	1	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	0	0	0	0	2	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	1	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	1
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0
31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
40	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
43	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
45	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
47	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
48	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
53	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0
60	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
63	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
75	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0
87	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
93	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
108	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
114	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
117	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
135	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
148	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0

Table 5.a A detailed list of the clusters' properties according to the Pearson's correlation exploited to define them. In each cell there is the number of clusters with a specific size, defined with a specific Pearson's correlation.

Clusters Size	Pearson Correlation value																														
	0.7	0.71	0.72	0.73	0.74	0.75	0.76	0.77	0.78	0.79	0.8	0.81	0.82	0.83	0.84	0.85	0.86	0.87	0.88	0.89	0.9	0.91	0.92	0.93	0.94	0.95	0.96	0.97	0.98	0.99
159	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
171	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
213	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
216	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
225	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
262	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
305	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
341	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
390	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
525	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
571	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
720	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
862	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
1821	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
2386	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
4058	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
4974	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
6025	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
7146	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
8358	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
9520	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
10494	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11441	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12255	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13083	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13780	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14426	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14979	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15425	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15856	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16264	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16631	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17013	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17335	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17609	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17886	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18125	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 5.b A detailed list of the clusters' properties according to the Pearson's correlation exploited to define them. In each cell there is the number of clusters with a specific size, defined with a specific Pearson's correlation.

We have started to evaluate the standing up functions inside the networks obtained. In table 6 are shown the 10 networks of co-expressed genes calculated with the highest Pearson's correlation coefficient, r=|0.99|, together with their specific TAIR9 functional notes and the frequency of these latter among the whole dataset. As listed also in table 5.a, 8 out of 10 are pair wise networks, and excluding 2 couples due to the presence of "unknow protein", each one of this small network shows a biologically meaningful function sharing. Practically always, the genes involved in these networks concern the translation and the energetic pathways.

The bigger networks, with id 1 and 2, collecting 23 and 7 genes respectively, are also involved in translation and energetic metabolism.

Comparing the number of specific functional note of each gene in the networks at r=|0.99| to the number of each specific functional note available in the whole dataset, it is interesting to notice that in physiological condition only a few part of the co-expressed genes are sharing the same functional description.

The GO terms enrichment in the network number 2 (Tab. 7), performed by GOrilla software [Eden et al., 2009] confirms moreover the significant availability of genes involved in the translation already shown by simply looking the functional notes in table 6, while conversely the GO terms in the network 1 are not showing meaningful enrichment, although the functional notes are clearly depicting the implication of the energetic metabolism. This is due to the lack of matching between genes' name and GO terms dictionaries, and this aspect has been kept in consideration by checking manually the functional notes in the small group without any sign of GO terms enrichment.

Number of notes with the same description in the whole dataset	Network ID	Note
2995	10	unknown protein;
2995	10	unknown protein;
7	9	LHCA1; FUNCTIONS IN: chlorophyll binding;
5	9	PHOTOSYSTEM I SUBUNIT H2 (PSAH2);
5	8	PLASTID-SPECIFIC RIBOSOMAL PROTEIN 4 (PSRP4);
2	8	ribosomal protein L3 family protein;
1	7	PSAN; FUNCTIONS IN: calmodulin binding; INVOLVED IN: photosynthetic electron transport in photosystem I;
2995	7	unknown protein;
10	6	ATP synthase family; FUNCTIONS IN: hydrogen ion transmembrane transporter activity;
1	6	Encodes the P subunit of Photosystem I. About 25% of the TMP14 pool appeared to be phosphorylated, and this ratio is not affected by
2	5	ribosomal protein L1 family protein;
1	5	RIBOSOMAL PROTEIN S1 (RPS1); FUNCTIONS IN: structural constituent of ribosome;
1	4	embryo defective 2184 (emb2184)
3	4	ribosomal protein L17 family protein;
2	3	60S ribosomal protein L15 (RPL15A);
6	3	60S ribosomal protein L7A (RPL7aB);
5	2	30S ribosomal protein S10, chloroplast, putative;
1	2	encodes a plastid ribosomal protein CL15, a constituent of the large subunit of the ribosomal complex ;
1	2	PLASTID RIBOSOMAL PROTEIN L11 (PRPL11);
2	2	ribosomal protein L18 family protein;
2	2	RIBOSOMAL PROTEIN L9 (RPL9);
2	2	ribosomal protein S5 family protein;
3	2	ribosomal protein S6 family protein;
1	1	cytochrome b6f complex subunit (petM), putative;
2	1	Encodes a protein predicted by sequence similarity with spinach PsaD to be photosystem I reaction center subunit II (PsaD1)
1	1	Encodes a protein similar to photosystem II reaction center subunit W. ; PHOTOSYSTEM II REACTION CENTER W (PSBW);
1	1	Encodes one of the two subunits forming the photosynthetic glyceraldehyde-3-phosphate dehydrogenase (GAPDH) and as such a constituent of the supramolecular complex with phosphoribulokinase
1	1	Encodes subunit F of photosystem I;
1	1	FED A; FUNCTIONS IN: electron carrier activity, iron-sulfur cluster binding, 2 iron, 2 sulfur cluster binding;
1	1	LEAF FNR 1 (ATLFNR1); FUNCTIONS IN: electron transporter, transferring electrons within the cyclic electron transport pathway of photosynthesis activity, oxidoreductase activity, poly(U) binding, NADPH dehydrogenase activity, electron transporter, transferring
7	1	LHCA3; FUNCTIONS IN: chlorophyll binding; INVOLVED IN: photosynthesis, light harvesting, photosynthesis; LOCATED IN: light-harvesting complex, chloroplast thylakoid membrane, chloroplast,
7	1	Lhcb6 protein (Lhcb6), light harvesting complex of photosystem II. ; LIGHT HARVESTING COMPLEX PSII SUBUNIT 6 (LHCB6); FUNCTIONS IN: chlorophyll binding;
1	1	LIGHT HARVESTING COMPLEX OF PHOTOSYSTEM II 5 (LHCB5);
2	1	LIGHT-HARVESTING CHLOROPHYLL B-BINDING PROTEIN 3 (LHCB3);
2	1	LIGHT-HARVESTING CHLOROPHYLL-PROTEIN COMPLEX I SUBUNIT A4 (LHCA4);
5	1	photosystem I subunit E-2 (PSAE-2);
5	1	PHOTOSYSTEM I SUBUNIT G (PSAG);
5	1	photosystem I subunit L (PSAL);
5	1	photosystem I subunit O (PSAO);
4	1	PHOTOSYSTEM II SUBUNIT P-1 (PSBP-1);
4	1	photosystem II subunit X (PSBX);
1	1	PS II OXYGEN-EVOLVING COMPLEX 1 (PSBO1);
1	1	PSA E1 KNOCKOUT (PSAE-1);
3	1	PSBY; INVOLVED IN: photosynthesis; LOCATED IN: chloroplast stromal thylakoid, chloroplast thylakoid membrane, chloroplast photosystem II, photosystem II;
1	1	recombination and DNA-damage resistance protein (DRT112) One of two Arabidopsis plastocyanin genes.
9	1	thylakoid lumen 18.3 kDa protein;

Table 6 Functional notes of the genes in the networks obtained with r=|0.99|

GO term	Description	P-value	Enrichment (N, B, n, b)
GO:0006412	translation	6.96E-08	61.11 (11978,196,4,4)
GO:0034645	cellular macromolecule biosynthetic process	2.66E-06	24.70 (11978,485,4,4)
GO:0009059	macromolecule biosynthetic process	3.25E-06	23.49 (11978,510,4,4)
GO:0044267	cellular protein metabolic process	3.15E-05	13.32 (11978,899,4,4)
GO:0019538	protein metabolic process	9.44E-05	10.13 (11978,1182,4,4)
GO:0010027	thylakoid membrane organization	1.62E-04	95.06 (11978,63,4,2)
GO:0009668	plastid membrane organization	1.62E-04	95.06 (11978,63,4,2)
GO:0044802	single-organism membrane organization	1.67E-04	93.58 (11978,64,4,2)
GO:0044249	cellular biosynthetic process	1.68E-04	8.78 (11978,1365,4,4)
GO:0061024	membrane organization	1.78E-04	90.74 (11978,66,4,2)
GO:0042254	ribosome biogenesis	2.00E-04	85.56 (11978,70,4,2)
GO:1901576	organic substance biosynthetic process	2.08E-04	8.32 (11978,1439,4,4)
GO:0022613	ribonucleoprotein complex biogenesis	2.24E-04	80.93 (11978,74,4,2)
GO:0009058	biosynthetic process	2.82E-04	7.71 (11978,1553,4,4)
GO:0044260	cellular macromolecule metabolic process	4.49E-04	6.86 (11978,1745,4,4)
GO:0043170	macromolecule metabolic process	6.72E-04	6.21 (11978,1930,4,4)
GO:0044085	cellular component biogenesis	7.13E-04	45.37 (11978,132,4,2)

Table 7 GO terms in the network with id 2 at r=|0.99|

Enrichment (N, B, n, b) is defined as follows:

Enrichment = (b/n) / (B/N)

N - the total number of gene dataset

B - the total number of genes associated with a specific GO term

n - the number of genes inside the group in which we are calculating the enrichment

b - the number of genes inside the analyzed group associated with a specific GO term

We have then investigated any possible association of the 5 biggest groups of correlated genes with a Pearson's value of |0.98| and their composition changing, while reducing the Pearson's coefficient.

The first group (red box, Fig. 23) collects only 9 correlated genes, and despite its small size, the GO terms enrichment revealed quite a significant enrichment in "cell wall" GO terms, with a p-value ranging between $2.43E^{-5}$ and $3.34\ E^{-4}$ (Tab. 8). As shown in figure 23, while moving from r=|0.98| to lower values, the group of genes increases its size dramatically, reaching 75 elements at r=|0.97|, 114 elements at r=|0.96| and 171 elements with r=|0.95|.

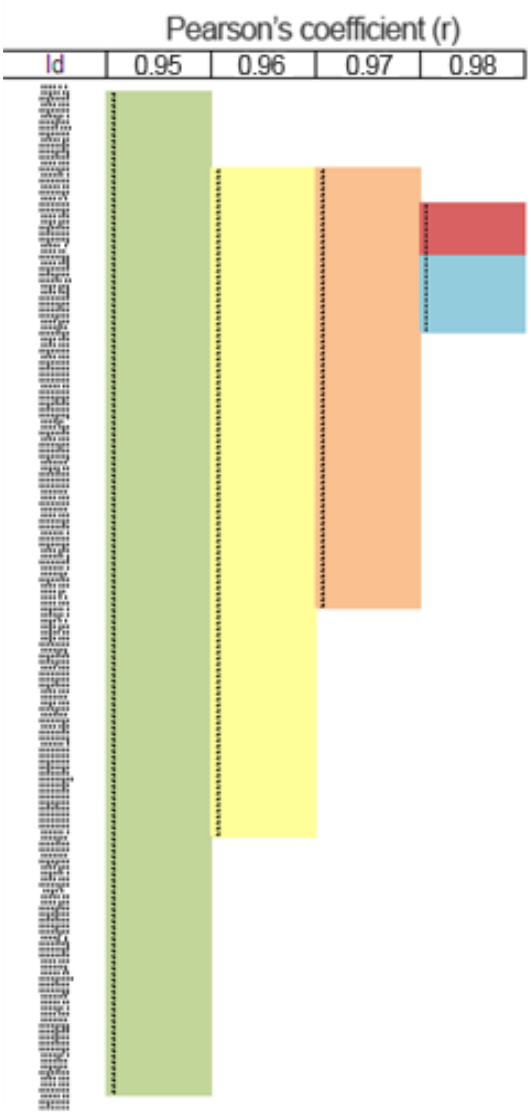


Figure 23 Evolution of co-expressed gene groups while reducing the Pearson's correlation threshold.

Pearson's coefficient

		0.98		0.97		0.96		0.95	
		p-value	enrichment	p-value	enrichment	p-value	enrichment	p-value	enrichment
GO:0009860	pollen tube growth	#N/D	#N/D	1.03E-09	24.22 (11978,92,43,8)	1.49E-13	22.32 (11978,92,70,12)	7.99E-19	22.36 (11978,92,99,17)
GO:0009932	cell tip growth	#N/D	#N/D	1.87E-09	22.51 (11978,99,43,8)	3.67E-13	20.74 (11978,99,70,12)	2.99E-18	20.78 (11978,99,99,17)
GO:0009826	unidimensional cell growth	#N/D	#N/D	2.57E-09	21.64 (11978,103,43,8)	5.97E-13	19.94 (11978,103,70,12)	6.06E-18	19.97 (11978,103,99,17)
GO:0009827	plant-type cell wall modification	2.43E-05	49.29 (11978,81,9,3)	9.82E-06	17.19 (11978,81,43,5)	3.23E-11	21.13 (11978,81,70,10)	2.92E-18	23.90 (11978,81,99,16)
GO:0042545	cell wall modification	7.50E-05	33.84 (11978,118,9,3)	6.08E-05	11.80 (11978,118,43,5)	7.01E-11	15.95 (11978,118,70,11)	6.68E-17	17.43 (11978,118,99,17)
GO:0009664	plant-type cell wall organization	9.12E-05	31.69 (11978,126,9,3)	8.31E-05	11.05 (11978,126,43,5)	2.69E-09	13.58 (11978,126,70,10)	4.61E-15	15.36 (11978,126,99,16)
GO:0071555	cell wall organization	1.36E-04	27.73 (11978,144,9,3)	1.56E-04	9.67 (11978,144,43,5)	6.09E-10	13.07 (11978,144,70,11)	2.09E-15	14.28 (11978,144,99,17)
GO:0071669	plant-type cell wall organization or biogenesis	1.50E-04	26.80 (11978,149,9,3)	1.83E-04	9.35 (11978,149,43,5)	1.37E-08	11.48 (11978,149,70,10)	6.79E-14	12.99 (11978,149,99,16)
GO:0016043	cellular component organization	1.95E-04	7.67 (11978,868,9,5)	6.58E-06	4.17 (11978,868,43,13)	1.95E-09	4.34 (11978,868,70,22)	6.41E-12	4.18 (11978,868,99,30)
GO:0045229	external encapsulating structure organization	2.03E-04	24.20 (11978,165,9,3)	2.94E-04	8.44 (11978,165,43,5)	2.60E-09	11.41 (11978,165,70,11)	2.10E-14	12.47 (11978,165,99,17)
GO:0071554	cell wall organization or biogenesis	2.98E-04	21.24 (11978,188,9,3)	5.36E-04	7.41 (11978,188,43,5)	1.03E-08	10.01 (11978,188,70,11)	1.85E-13	10.94 (11978,188,99,17)
GO:0071840	cellular component organization or biogenesis	3.34E-04	6.84 (11978,973,9,5)	2.26E-05	3.72 (11978,973,43,13)	1.61E-08	3.87 (11978,973,70,22)	1.11E-10	3.73 (11978,973,99,30)

Table 8 GO terms of the first group

Enrichment (N, B, n, b) is defined as follows:

Enrichment = (b/n) / (B/N)

N - the total number of gene dataset

B - the total number of genes associated with a specific GO term

n - the number of genes inside the group in which we are calculating the enrichment

b - the number of genes inside the analyzed group associated with a specific GO term

Grey boxes indicate the GO terms with a p-value $<10^{-7}$

As expected, the GO terms enrichment was strongly affected during this downhill, because of the addition of many other components. However the p-value of the GO terms found is getting higher, underlying that GO terms for "plant wall" are getting more significant also in these larger groups at lower Pearson's thresholds (Tab. 8)

The second group of 13 genes (blue box, Fig. 23), as shown in figure 23 is related to the first, the red one, because it merges with this latter when r=|0.97|. However the GO analysis did not revealed significant terms, although checking the Tair 9 functional notation of these genes (Tab. 9)

GeneName	Note
AT3G28830	unknown protein;
AT5G61720	unknown protein;;
AT3G57690	ARABINOGALACTAN-PROTEIN 23 (AGP23)
AT5G14380	Arabinogalactan proteins 6 (AGP6). ;
AT3G01270	pectate lyase family protein; FUNCTIONS IN: lyase activity, pectate lyase activity; I
AT3G13400	SKU5 Similar 13 (sks13); FUNCTIONS IN: oxidoreductase activity, copper ion binding;
AT3G62170	Vanguard 1 homolog 2 (VGDH2); FUNCTIONS IN: enzyme inhibitor activity, pectinesterase activity;). ;
AT1G02790	encodes a exopolygalacturonase. ; POLYGALACTURONASE 4 (PGA4); FUNCTIONS IN: polygalacturonase activity; INVOLVED IN: carbohydrate metabolic process; LOCATED IN: : endomembrane system
AT5G07430	pectinesterase family protein; FUNCTIONS IN: pectinesterase activity; INVOLVED IN: cell wall modification;
AT1G55570	SKU5 Similar 12 (sks12); FUNCTIONS IN: oxidoreductase activity, copper ion binding; LOCATED IN: endomembrane system; EXPRESSED IN
AT3G07820	polygalacturonase 3 (PGA3) / pectinase; FUNCTIONS IN: polygalacturonase activity; INVOLVED IN: carbohydrate metabolic process; LOCATED IN: endomembrane system;
AT1G28270	ralf-like 4 (RALFL4); FUNCTIONS IN: signal transducer activity;
AT5G38760	unknown protein;

Table 9 Tair 9 functional notation of the second group

we noticed that six genes out of thirteen are involved in the metabolism of the pectin, which is a compound also needed to the cell wall formation. So the identification of gene networks through Pearson's coefficient is able to distinguish in a deeper way genes involved in the same function, but co-expressed in different patterns.

The third network (red box, Fig. 24) includes 53 genes. As seen for the first network, also here, reducing r to lower values, implies at least the doubling of the group sizes in the next threshold of r= |0.97|, where it collects 117 elements, and goes bigger to 148 at r=|0.96| and much bigger at r=|0.95| with 390 elements.

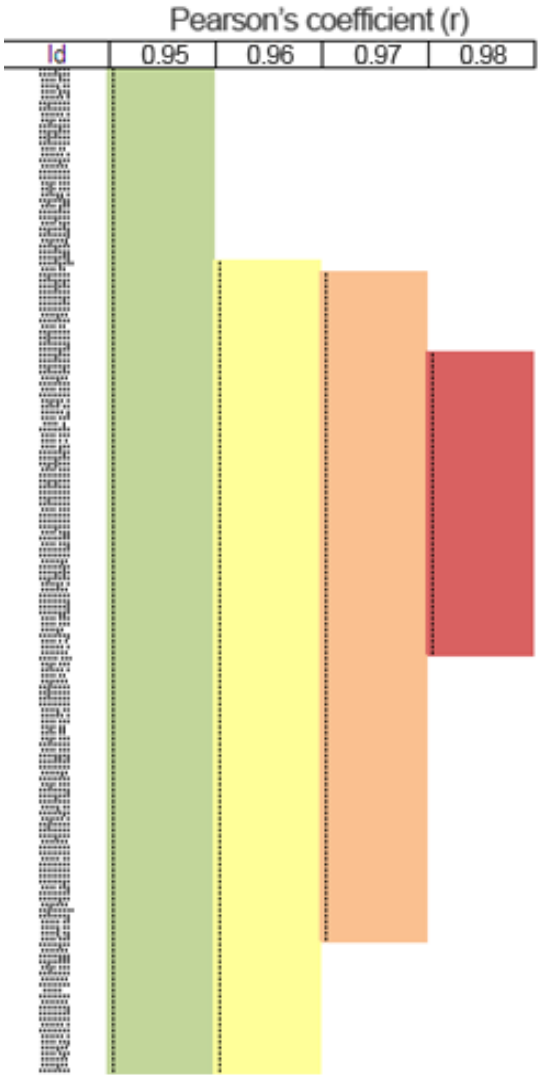


Figure 24 Evolution of co-expressed gene groups while reducing the Pearson's correlation threshold

Pearson's coefficient

		0.98		0.97		0.96		0.95	
		p-value	enrichment	p-value	enrichment	p-value	enrichment	p-value	enrichment
GO:0006412	translation	1.20E-60	53.66 (11978,196,41,36)	9.97E-130	51.71 (11978,196,91,77)	6.47E-153	49.32 (11978,196,114,92)	2.93E-136	26.69 (11978,196,245,107)
GO:0034645	cellular macromolecule biosynthetic process	2.74E-50	22.89 (11978,485,41,38)	3.38E-106	22.25 (11978,485,91,82)	1.01E-122	21.23 (11978,485,114,98)	2.59E-105	12.20 (11978,485,245,121)
GO:0009059	macromolecule biosynthetic process	1.98E-49	21.77 (11978,510,41,38)	2.98E-104	21.16 (11978,510,91,82)	2.33E-120	20.19 (11978,510,114,98)	2.09E-102	11.60 (11978,510,245,121)
GO:0001510	RNA methylation	7.26E-40	64.92 (11978,108,41,24)	3.24E-85	62.16 (11978,108,91,51)	4.05E-92	55.45 (11978,108,114,57)	1.79E-123	37.57 (11978,108,245,83)
GO:0042254	ribosome biogenesis	3.29E-16	45.91 (11978,70,41,11)	1.22E-40	50.77 (11978,70,91,27)	4.83E-54	52.54 (11978,70,114,35)	3.30E-50	27.94 (11978,70,245,40)
GO:0071840	cellular component organization or biogenesis	6.81E-10	5.40 (11978,973,41,18)	2.21E-23	5.82 (11978,973,91,43)	3.59E-30	5.94 (11978,973,114,55)	3.94E-44	4.92 (11978,973,245,98)
GO:0009165	nucleotide biosynthetic process	5.24E-06	13.28 (11978,132,41,6)	5.65E-08	9.97 (11978,132,91,10)	1.58E-12	11.94 (11978,132,114,15)	3.63E-51	18.52 (11978,132,245,50)
GO:1901293	nucleoside phosphate biosynthetic process	5.47E-06	13.18 (11978,133,41,6)	6.08E-08	9.90 (11978,133,91,10)	1.77E-12	11.85 (11978,133,114,15)	5.73E-51	18.38 (11978,133,245,50)
GO:0009220	pyrimidine ribonucleotide biosynthetic process	5.64E-06	19.22 (11978,76,41,5)	1.65E-06	12.12 (11978,76,91,7)	2.31E-09	13.83 (11978,76,114,10)	1.19E-44	24.44 (11978,76,245,38)
GO:0009218	pyrimidine ribonucleotide metabolic process	5.64E-06	19.22 (11978,76,41,5)	1.65E-06	12.12 (11978,76,91,7)	2.31E-09	13.83 (11978,76,114,10)	1.19E-44	24.44 (11978,76,245,38)
GO:0006221	pyrimidine nucleotide biosynthetic process	6.02E-06	18.97 (11978,77,41,5)	1.81E-06	11.97 (11978,77,91,7)	2.63E-09	13.65 (11978,77,114,10)	2.32E-44	24.13 (11978,77,245,38)
GO:0006220	pyrimidine nucleotide metabolic process	6.02E-06	18.97 (11978,77,41,5)	1.81E-06	11.97 (11978,77,91,7)	2.63E-09	13.65 (11978,77,114,10)	2.32E-44	24.13 (11978,77,245,38)
GO:0072528	pyrimidine-containing compound biosynthetic process	6.42E-06	18.73 (11978,78,41,5)	1.97E-06	11.81 (11978,78,91,7)	2.99E-09	13.47 (11978,78,114,10)	4.44E-44	23.82 (11978,78,245,38)
GO:0046390	ribose phosphate biosynthetic process	8.21E-06	17.81 (11978,82,41,5)	2.77E-06	11.24 (11978,82,91,7)	4.93E-09	12.81 (11978,82,114,10)	5.35E-43	22.66 (11978,82,245,38)

Table 10 GO terms of the third group

Enrichment (N, B, n, b) is defined as follows:

Enrichment = (b/n) / (B/N)

N - the total number of gene dataset

B - the total number of genes associated with a specific GO term

n - the number of genes inside the group in which we are calculating the enrichment

b - the number of genes inside the analyzed group associated with a specific GO term

Grey boxes indicate the GO terms with a p-value $<10^{-7}$

The GO terms analyses identify in this group the “translation” and the “transcription machineries” (Tab. 10). Enrichment is less responsive to the reduction of r while passing from |0.98| to |0.97|, indicating that during this last step, the new genes added to the initial main core from r= |0.98| are involved in the same functions.

The fourth group (red box, Fig. 25) collects 60 genes at r=|0.98|. Also here moving from r=|0.98| implies adding more genes, reaching a group of 305 elements at r=|0.97|, 525 at r=|0.96| and 720 elements with r=|0.95|. The GO Terms analysis has revealed a significant enrichment in the GO terms related to the energy metabolism, like photosynthesis, plastid and photosystem (Fig. 26). As shown also in figure 25 (blue box), the firth group composed by 87 genes, collapses with the fourth one (red box) when using r=|0.97|.

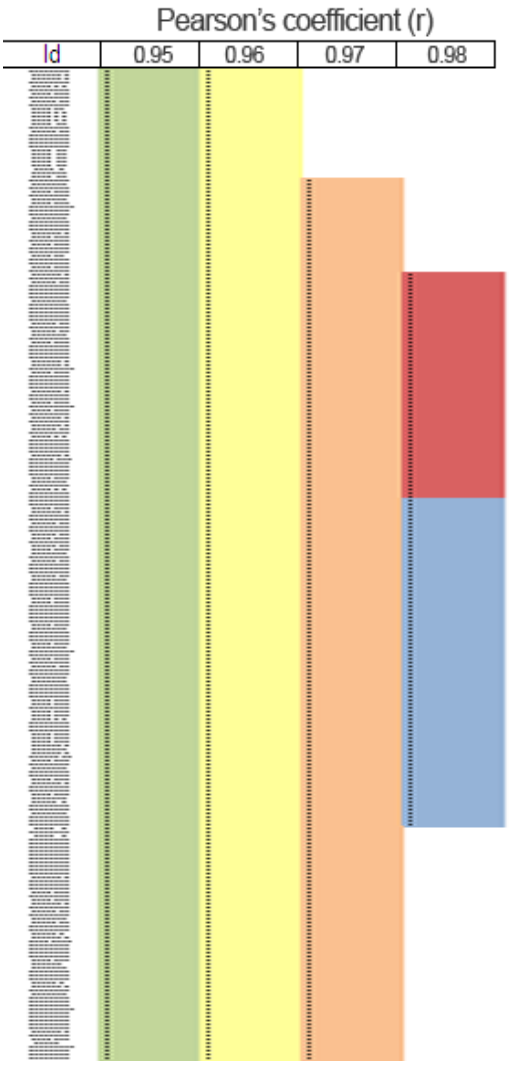


Figure 25 Evolution of co-expressed gene groups while reducing the Pearson’s correlation threshold

The firth group is also involved in the energy metabolism but in a deeper way, in fact the enrichment of GO Terms (Tab.11) at r=|0.98| is quite three times higher for the terms related to "photosynthesis" and more significant according to their p-value (Fig. 26).

		Pearson's coefficient			
		0.98		0.98	
		p-value	enrichment	p-value	enrichment
GO:0019682	glyceraldehyde-3-phosphate metabolic	2.86E-40	90.37 (11978,81,36,22)	3.59E-11	51.76 (11978,81,20,7)
GO:0019288	isopentenyl diphosphate biosynthetic	2.86E-40	90.37 (11978,81,36,22)	3.59E-11	51.76 (11978,81,20,7)
GO:0006090	pyruvate metabolic process	3.90E-40	89.27 (11978,82,36,22)	3.92E-11	51.13 (11978,82,20,7)
GO:0015979	photosynthesis	6.18E-05	38.39 (11978,26,36,3)	4.86E-10	115.17 (11978,26,20,5)
GO:0010207	photosystem II assembly	2.30E-07	36.17 (11978,46,36,5)	8.51E-11	78.12 (11978,46,20,6)
GO:0019684	photosynthesis, light reaction	6.94E-05	36.97 (11978,27,36,3)	5.96E-10	110.91 (11978,27,20,5)
GO:0055114	oxidation-reduction process	9.90E-06	11.81 (11978,169,36,6)	1.45E-10	28.35 (11978,169,20,8)
GO:1901576	organic substance biosynthetic process	1.53E-27	7.63 (11978,1439,36,33)	3.54E-08	5.41 (11978,1439,20,13)
GO:1901135	carbohydrate derivative metabolic process	1.74E-27	26.05 (11978,281,36,22)	8.20E-09	17.05 (11978,281,20,8)
GO:0009657	plastid organization	1.64E-17	44.36 (11978,90,36,12)	5.96E-17	66.54 (11978,90,20,10)
GO:0032787	monocarboxylic acid metabolic process	1.10E-25	18.99 (11978,403,36,23)	1.35E-07	11.89 (11978,403,20,8)
GO:0006629	lipid metabolic process	5.53E-22	14.73 (11978,497,36,22)	9.82E-06	8.44 (11978,497,20,7)
GO:0019637	organophosphate metabolic process	1.38E-25	18.80 (11978,407,36,23)	1.99E-13	17.66 (11978,407,20,12)
GO:0008610	lipid biosynthetic process	6.46E-25	20.00 (11978,366,36,22)	1.29E-06	11.45 (11978,366,20,7)
GO:0034645	cellular macromolecule biosynthetic process	7.92E-24	15.78 (11978,485,36,23)	8.37E-06	8.64 (11978,485,20,7)
GO:0044255	cellular lipid metabolic process	1.36E-23	17.43 (11978,420,36,22)	3.23E-06	9.98 (11978,420,20,7)
GO:0009059	macromolecule biosynthetic process	2.52E-23	15.01 (11978,510,36,23)	1.16E-05	8.22 (11978,510,20,7)

Figure 26 GO terms of fourth (red box) and fifth group (blue box).

Enrichment (N, B, n, b) is defined as follows: Enrichment = (b/n) / (B/N)

N - the total number of gene dataset; B - the total number of genes associated with a specific GO term; n - the number of genes inside the group in which we are calculating the enrichment; b - the number of genes inside the analyzed group associated with a specific GO term

Pearson's coefficient

		0.98		0.97		0.96		0.95	
		p-value	enrichment	p-value	enrichment	p-value	enrichment	p-value	enrichment
GO:0019682	glyceraldehyde-3-phosphate metabolic process	3.59E-11	51.76 (11978,81,20,7)	1.03E-81	53.87 (11978,81,140,51)	9.16E-102	41.36 (11978,81,236,66)	4.22E-111	33.77 (11978,81,324,74)
GO:0019288	isopentenyl diphosphate biosynthetic	3.59E-11	51.76 (11978,81,20,7)	1.03E-81	53.87 (11978,81,140,51)	9.16E-102	41.36 (11978,81,236,66)	4.22E-111	33.77 (11978,81,324,74)
GO:0006090	pyruvate metabolic process	3.92E-11	51.13 (11978,82,20,7)	2.71E-81	53.21 (11978,82,140,51)	4.63E-101	40.85 (11978,82,236,66)	4.24E-110	33.36 (11978,82,324,74)
GO:0015979	photosynthesis	4.86E-10	115.17 (11978,26,20,5)	3.92E-21	46.07 (11978,26,140,14)	1.52E-23	33.19 (11978,26,236,17)	4.83E-23	25.59 (11978,26,324,18)
GO:0010207	photosystem II assembly	8.51E-11	78.12 (11978,46,20,6)	1.75E-25	35.34 (11978,46,140,19)	7.14E-37	30.89 (11978,46,236,28)	2.51E-44	27.33 (11978,46,324,34)
GO:0019684	photosynthesis, light reaction	5.96E-10	110.91 (11978,27,20,5)	5.89E-25	50.70 (11978,27,140,16)	8.30E-34	41.36 (11978,27,236,22)	8.44E-38	34.23 (11978,27,324,25)
GO:0055114	oxidation-reduction process	1.45E-10	28.35 (11978,169,20,8)	4.42E-33	17.21 (11978,169,140,34)	7.28E-39	13.51 (11978,169,236,45)	1.38E-38	10.94 (11978,169,324,50)
GO:1901576	organic substance biosynthetic process	3.54E-08	5.41 (11978,1439,20,13)	1.89E-53	5.59 (11978,1439,140,94)	1.90E-73	5.08 (11978,1439,236,144	1.15E-75	4.44 (11978,1439,324,173)
GO:1901135	carbohydrate derivative metabolic process	8.20E-09	17.05 (11978,281,20,8)	2.27E-52	16.44 (11978,281,140,54)	1.48E-64	13.37 (11978,281,236,74)	1.18E-68	11.31 (11978,281,324,86)
GO:0009657	plastid organization	5.96E-17	66.54 (11978,90,20,10)	3.55E-51	36.12 (11978,90,140,38)	1.17E-68	29.89 (11978,90,236,53)	1.12E-81	26.29 (11978,90,324,64)
GO:0032787	monocarboxylic acid metabolic process	1.35E-07	11.89 (11978,403,20,8)	1.51E-43	11.46 (11978,403,140,54)	7.95E-50	9.07 (11978,403,236,72)	8.32E-50	7.52 (11978,403,324,82)

Table 11 GO terms of the firth group

Enrichment (N, B, n, b) is defined as follows:

Enrichment = (b/n) / (B/N)

N - the total number of gene dataset

B - the total number of genes associated with a specific GO term

n - the number of genes inside the group in which we are calculating the enrichment

b - the number of genes inside the analyzed group associated with a specific GO term

Grey boxes indicate the GO terms with a p-value $<10^{-7}$

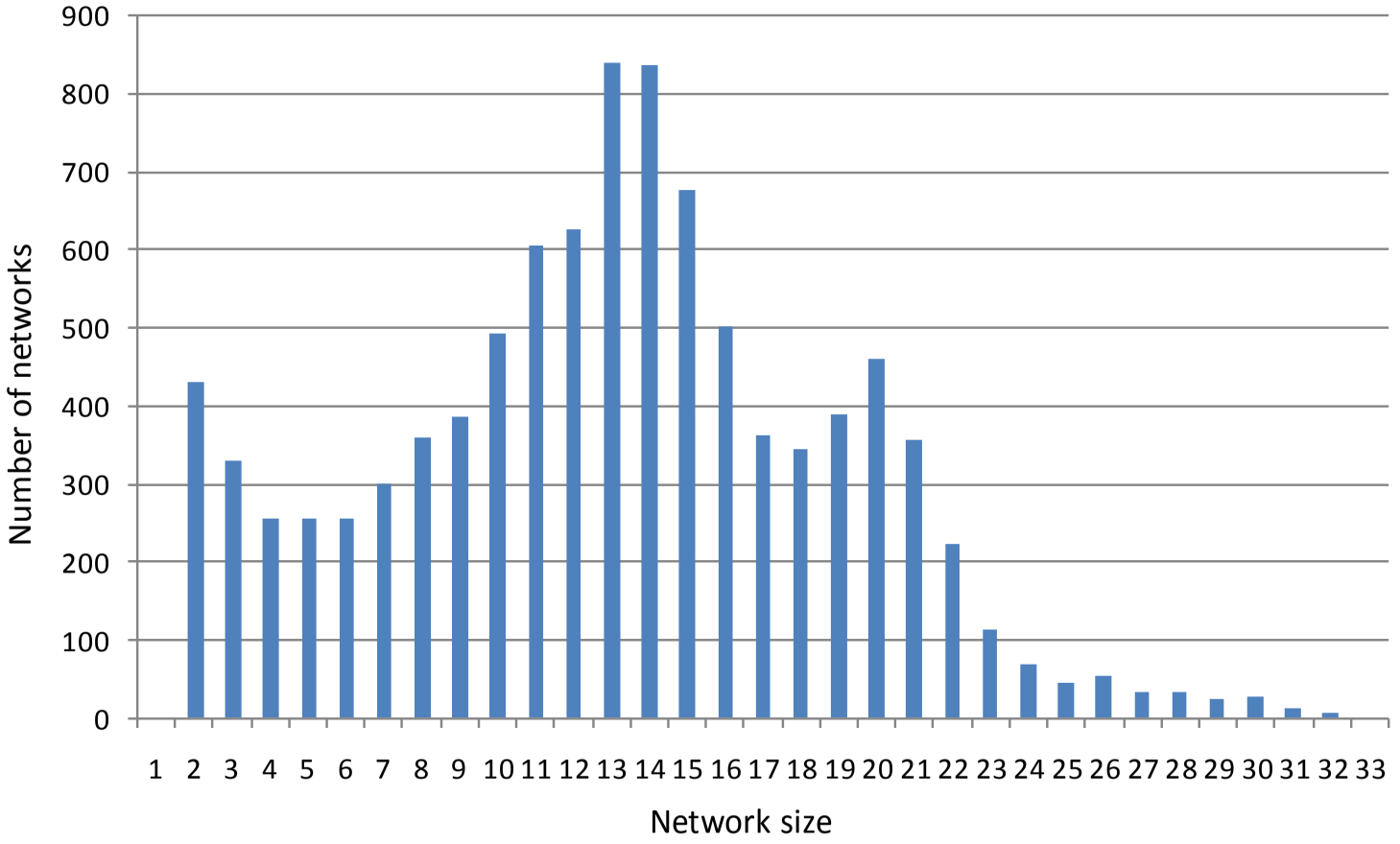
As happened for the first and second group it appears that the analysis based on Pearson's coefficients can detect not only the correlated genes, but also distinguish the sub-functionalities coordinated in specific process. Moreover with a very high Pearson's correlation value (r=0.98) we have found three different and specific functional "cores": "plant shape", for the first and second group; "transcription" and "translation" in the third one; "energy metabolism" in the fourth and fifth one. This underlines that the genes involved in the main functionalities of the cell are interacting in specific networks, and are only few ones among all possible when considering physiological conditions.

3.4 Network analysis by Clique

The Pearson test is a strong tool, but it can only suggest the presence of a biological meaning behind the network of co-expressed genes, without clarifying the interactions and so the relationships of the genes inside the genome. Chains of pair wise co-expressed genes based on their expression are hard to be understood due to the high number of low connected elements inside the cluster, probably not involved in a specific process, but collected because of methodological issues.

We have therefore further analyzed the co-expressed genes according to the number of correlations shared by them. Following this purpose, we have discovered by literature that the most efficient way to identify groups of elements in an environment is the search for the cliques [Palla G., 2005]. The research of maximal cliques has been initially exploited to detect group of co-expressed genes, through a Matlab software (http://www.mathworks.com/ matlabcentral/fileexchange/30413-bron-kerbosch-maximal-clique-finding-algorithm). In the first trial, this algorithm was applied on a dataset collecting all the genes with at least one significant correlation ($r \geq |0.7|$), with a total number of 18136 genes, as described.

However, the huge amount of time and resources required by the clique algorithm has made impossible to get the results in a convenient way. So we have decided to reduce the size of the dataset to investigate only the genes with at least one Pearson correlation value of |0.95|. By this way, the dataset was built only by 1589 genes and applying the maximal clique algorithm we were able to obtain 9729 cliques in 1,5 hour (Fig. 27).



Net Size	Nr. Of Nets
2	432
3	331
4	257
5	256
6	257
7	300
8	360
9	388
10	494
11	605
12	626
13	841
14	837
15	678
16	502
17	364
18	346
19	389
20	460
21	358
22	225
23	115
24	69
25	45
26	53
27	34
28	35
29	25
30	28
31	12
32	6
33	1

Figure 27 Number of the cliques detected with a Pearson's coefficient of |0.95| and distributed according to their size.

The number is very high and it is due to the fact that the algorithm exploited for the analyses considers as different cliques, the ones with the same size, but built with even only one member inclusion and another member exclusion, with the aim to keep the all-versus-all relationship among the clique members (Fig. 28). This behavior provides redundant cliques, i.e. cliques sharing identical genes

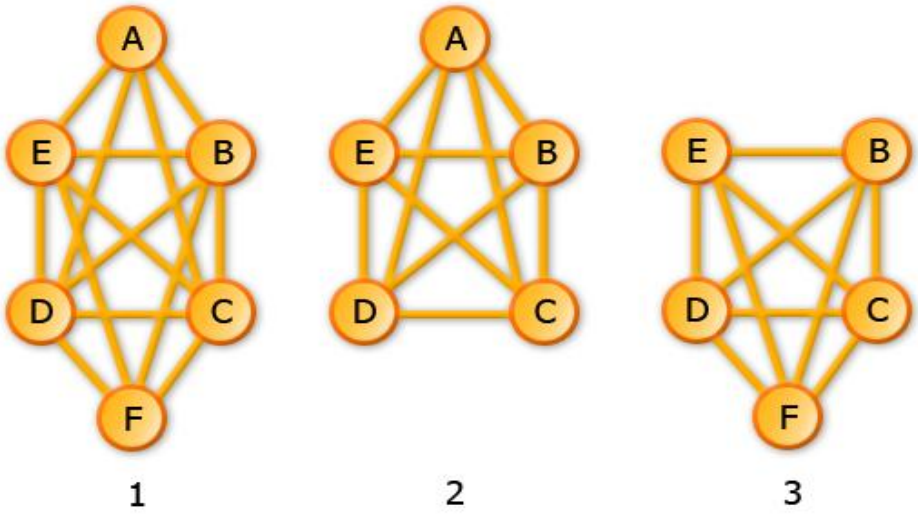


Figure 28 The network 1 may appear as a clique but it is not so, in fact the node A and F are not linked. Hence, the maximal clique algorithm will find two cliques, 2 and 3, which are differing only for one node.

We have merged the cliques sharing more than 50% of their elements and checked the results through GO terms enrichment. In table 12 we listed the genes within five of the biggest cliques obtained. The GO terms of the group number 1 (Tab. 13) reveal an enrichment of GO terms related to RNA "methylation", "translation" and "different metabolic process". However the p-value is never below $10^{-7,}$ which means that the enrichment is not significant to declare that GO terms found in this group are depicting the whole GO terms dictionary to which they belong.

1	2	3	4	5
AT4G14320	AT5G27850	AT4G34190	AT5G28500	AT5G64040
AT3G59540	AT5G03850	AT4G09650	AT5G53490	AT1G42970
AT3G53020	AT3G48930	AT1G09340	AT5G44650	AT1G67740
AT3G52590	AT3G24830	AT5G23120	AT3G55330	AT5G66190
AT1G34030	AT1G22780	AT3G54050	AT3G47650	AT5G46110
AT1G72370	AT2G27530	AT3G12780	AT3G18890	AT3G54890
AT1G67430	AT2G44120	AT1G71500	AT3G23700	AT4G28750
AT1G43170	AT4G15000	AT1G68010	AT3G01480	AT4G05180
AT2G27720	AT5G59850	AT1G74730	AT1G76450	AT4G01150
	AT5G07090	AT1G65230	AT2G43560	AT2G20260
	AT3G49010	AT2G34860	AT1G12800	AT2G26500
	AT3G02560		AT1G44920	AT1G44575
	AT2G39460		AT1G68590	AT5G66570
	AT5G52650		AT1G62780	AT5G54270
	AT3G60770			AT5G02120
	AT4G00100			AT3G56940
	AT1G18540			AT3G47470
	AT3G28900			AT4G12800
	AT3G23390			AT4G10340
	AT3G02080			AT4G02770
	AT3G09200			AT1G20340
	AT3G04920			AT3G21055
	AT2G41840			AT3G26650
	AT2G01250			AT1G15820
	AT2G19730			AT1G52230
				AT1G12900
				AT1G08380
				AT1G31330
				AT1G06680
				AT1G54780
				AT1G61520
				AT2G06520
				AT2G30570

Table 12 Genes within five of the biggest cliques obtained

GO term	Description	P-value	Enrichment (N, B, n, b)
GO:0034645	cellular macromolecule biosynthetic process	1.18E-07	24.19 (11998,496,5,5)
GO:0009059	macromolecule biosynthetic process	1.52E-07	23.03 (11998,521,5,5)
GO:0006412	translation	3.41E-07	48.97 (11998,196,5,4)
GO:0001510	RNA methylation	7.00E-06	66.66 (11998,108,5,3)
GO:0044249	cellular biosynthetic process	1.97E-05	8.72 (11998,1376,5,5)
GO:1901576	organic substance biosynthetic process	2.56E-05	8.27 (11998,1450,5,5)
GO:0009058	biosynthetic process	3.61E-05	7.73 (11998,1553,5,5)
GO:0009451	RNA modification	4.52E-05	35.81 (11998,201,5,3)
GO:0032259	methylation	6.50E-05	31.71 (11998,227,5,3)
GO:0043414	macromolecule methylation	6.50E-05	31.71 (11998,227,5,3)
GO:0044260	cellular macromolecule metabolic process	6.51E-05	6.87 (11998,1747,5,5)
GO:0043170	macromolecule metabolic process	1.08E-04	6.21 (11998,1932,5,5)
GO:0044267	cellular protein metabolic process	1.47E-04	10.68 (11998,899,5,4)
GO:0042254	ribosome biogenesis	3.32E-04	68.56 (11998,70,5,2)
GO:0022613	ribonucleoprotein complex biogenesis	3.71E-04	64.85 (11998,74,5,2)
GO:0019538	protein metabolic process	4.32E-04	8.12 (11998,1182,5,4)
GO:0044237	cellular metabolic process	7.23E-04	4.25 (11998,2826,5,5)
GO:0071704	organic substance metabolic process	9.83E-04	3.99 (11998,3005,5,5)

Table 13 The GO terms of the first group number

Enrichment (N, B, n, b) is defined as follows:

Enrichment = (b/n) / (B/N)

N - the total number of gene dataset

B - the total number of genes associated with a specific GO term

n - the number of genes inside the group in which we are calculating the enrichment

b - the number of genes inside the analyzed group associated with a specific GO term

The group number 2 instead is significantly enriched with terms related to translation and RNA methylation as reported in table 14

GO term	Description	P-value	Enrichment (N, B, n, b)
GO:0006412	translation	4.62E-35	61.21 (11998,196,19,19)
GO:0034645	cellular macromolecule biosynthetic process	3.68E-27	24.19 (11998,496,19,19)
GO:0009059	macromolecule biosynthetic process	9.52E-27	23.03 (11998,521,19,19)
GO:0044267	cellular protein metabolic process	3.48E-22	13.35 (11998,899,19,19)
GO:0019538	protein metabolic process	6.60E-20	10.15 (11998,1182,19,19)
GO:0044249	cellular biosynthetic process	1.21E-18	8.72 (11998,1376,19,19)
GO:1901576	organic substance biosynthetic process	3.29E-18	8.27 (11998,1450,19,19)
GO:0009058	biosynthetic process	1.22E-17	7.73 (11998,1553,19,19)
GO:0044260	cellular macromolecule metabolic process	1.16E-16	6.87 (11998,1747,19,19)
GO:0043170	macromolecule metabolic process	7.92E-16	6.21 (11998,1932,19,19)
GO:0001510	RNA methylation	2.37E-14	52.62 (11998,108,19,9)
GO:0044238	primary metabolic process	2.59E-13	4.58 (11998,2617,19,19)
GO:0044237	cellular metabolic process	1.12E-12	4.25 (11998,2826,19,19)

Table 14 The GO terms of the group number 2

Enrichment (N, B, n, b) is defined as follows: Enrichment = (b/n) / (B/N)

N - the total number of gene dataset; B - the total number of genes associated with a specific GO term; n - the number of genes inside the group in which we are calculating the enrichment; b - the number of genes inside the analyzed group associated with a specific GO term.

The group number 3 instead did not provide any results in the GO terms analysis, despite its functional annotations (Tab. 15) are almost referring to a chloroplast activity.

AT4G34190	Encodes a stress enhanced protein that localizes to the thylakoid membrane and whose mRNA is upregulated in response to high light intensity. It may be involved in chlorophyll binding.
AT4G09650	Encodes the chloroplast ATPase delta-subunit. ; ATP SYNTHASE DELTA-SUBUNIT GENE (ATPD); FUNCTIONS IN: hydrogen ion transporting ATP synthase activity, rotational mechanism, proton-transporting ATPase activity, rotational mechanism; INVOLVED IN: response to cold, defense response to bacterium, photosynthetic electron transport in photosystem II, photosynthetic electron transport in photosystem I, photosynthesis;
AT1G09340	CHLOROPLAST RNA BINDING (CRB); FUNCTIONS IN: coenzyme binding, binding, catalytic activity; INVOLVED IN: in 6 processes; L
AT5G23120	encodes a stability and/or assembly factor of photosystem II ; HCF136; FUNCTIONS IN: protein binding; INVOLVED IN: response to cadmium ion, plastid organization, protein complex assembly; LOCATED IN: in 8 components; EXPRESSED IN: 25 plant structures; EXPRESSED DURING: 14 growth stages;
AT3G54050	fructose-1,6-bisphosphatase, putative / D-fructose-1,6-bisphosphate 1-phosphohydrolase, putative / FBPase, putative; FUNCTIONS IN: fructose 1,6-bisphosphate 1-phosphatase activity, phosphoric ester hydrolase activity; INVOLVED IN: response to cold, fructose metabolic process;
AT3G12780	nuclear phosphoglycerate kinase (PGK1) ; PHOSPHOGLYCERATE KINASE 1 (PGK1); FUNCTIONS IN: phosphoglycerate kinase activity; INVOLVED IN: response to cadmium ion, response to cold, glycolysis, peptidyl-cysteine S-nitrosylation; LOCATED IN: in 11 components;
AT1G71500	Rieske (2Fe-2S) domain-containing protein; FUNCTIONS IN: electron carrier activity, oxidoreductase activity, 2 iron, 2 sulfur cluster binding; INVOLVED IN: oxidation reduction; LOCATED IN: chloroplast thylakoid membrane, chloroplast, chloroplast envelope;
AT1G68010	HPR; FUNCTIONS IN: glycerate dehydrogenase activity, poly(U) binding; INVOLVED IN: photorespiration; LOCATED IN: apoplast, chloroplast, peroxisome;
AT1G74730	unknown protein; FUNCTIONS IN: molecular_function unknown; INVOLVED IN: biological_process unknown; LOCATED IN: chloroplast thylakoid membrane, chloroplast;
AT1G65230	unknown protein; FUNCTIONS IN: molecular_function unknown;
AT2G34860	embryo sac development arrest 3 (EDA3); FUNCTIONS IN: unfolded protein binding, heat shock protein binding; INVOLVED IN: megagametogenesis;

Table 15 Functional annotations of third group

The group number 4, despite its size of only 14 elements, is able to detect strongly and significantly the genes belonging to the GO terms related to “energy processes” (Tab. 16)

GO term	Description	P-value	Enrichment (N, B, n, b)
GO:0019682	glyceraldehyde-3-phosphate metabolic process	4.91E-16	107.73 (11998,81,11,8)
GO:0019288	isopentenyl diphosphate biosynthetic process,	4.91E-16	107.73 (11998,81,11,8)
GO:0009240	isopentenyl diphosphate biosynthetic process	4.91E-16	107.73 (11998,81,11,8)
GO:0046490	isopentenyl diphosphate metabolic process	4.91E-16	107.73 (11998,81,11,8)
GO:0006090	pyruvate metabolic process	5.44E-16	106.41 (11998,82,11,8)
GO:0006081	cellular aldehyde metabolic process	1.90E-14	69.25 (11998,126,11,8)
GO:0010027	thylakoid membrane organization	2.53E-14	121.19 (11998,63,11,7)
GO:0009668	plastid membrane organization	2.53E-14	121.19 (11998,63,11,7)
GO:0044802	single-organism membrane organization	2.84E-14	119.30 (11998,64,11,7)
GO:0061024	membrane organization	3.56E-14	115.68 (11998,66,11,7)
GO:0008654	phospholipid biosynthetic process	1.99E-13	51.94 (11998,168,11,8)
GO:0006644	phospholipid metabolic process	2.19E-13	51.33 (11998,170,11,8)
GO:0009657	plastid organization	3.39E-13	84.83 (11998,90,11,7)
GO:0019637	organophosphate metabolic process	2.82E-12	24.12 (11998,407,11,9)
GO:0034660	ncRNA metabolic process	3.80E-12	60.60 (11998,126,11,7)
GO:0006739	NADP metabolic process	8.20E-12	102.26 (11998,64,11,6)
GO:0006740	NADPH regeneration	8.20E-12	102.26 (11998,64,11,6)
GO:0006629	lipid metabolic process	1.21E-09	17.56 (11998,497,11,8)
GO:0010207	photosystem II assembly	7.74E-04	47.42 (11998,46,11,2)

Table 16 Go terms of group number 4

Unexpected, the clique number 5 which is also the largest one does not provide any significant GO terms enrichment.

This evaluation of co-expressed groups of genes through the clique method has shown to be limited to small datasets due to its time of complexity. Our analysis in fact was able to provide results in a convenient time only when exploiting a dataset of 1589 genes. Despite its strong skill to identify genes co-expressed and cross-linked together, the networks proposed with this method are too much and very small for a supportive functional screening with GO terms enrichment. Moreover this is a time expensive approach which is general hard to manage

3.5 Social network approach

The clique approach is a very powerful way to identifies networks of co-expressed genes, but it is hard to manage due to its calculation time and redundant results. We have hence tried to produce network analysis able to be fast and efficient, inspired to the social networks.

The social network algorithm we applied (see chapter 2.5) has been used to analyze our collection of 20908 genes, exploiting the expression signals available in our dataset of 79 samples prepared as described in the materials and methods section. Hence the matrix input to the algorithm was built by 20908x79 cells. Different thresholds of θ_1 and θ_2 have been exploited and combined: θ_1=0.7; θ_1=0.8; θ_1=0.9 with steps of θ_2 equal to 0.5 ranging from 0 to 1. All the results have been collected in several tables as the one shown in figure 29, whose cells store a number identifying the genes belonging to a specific group, or better saying "community".

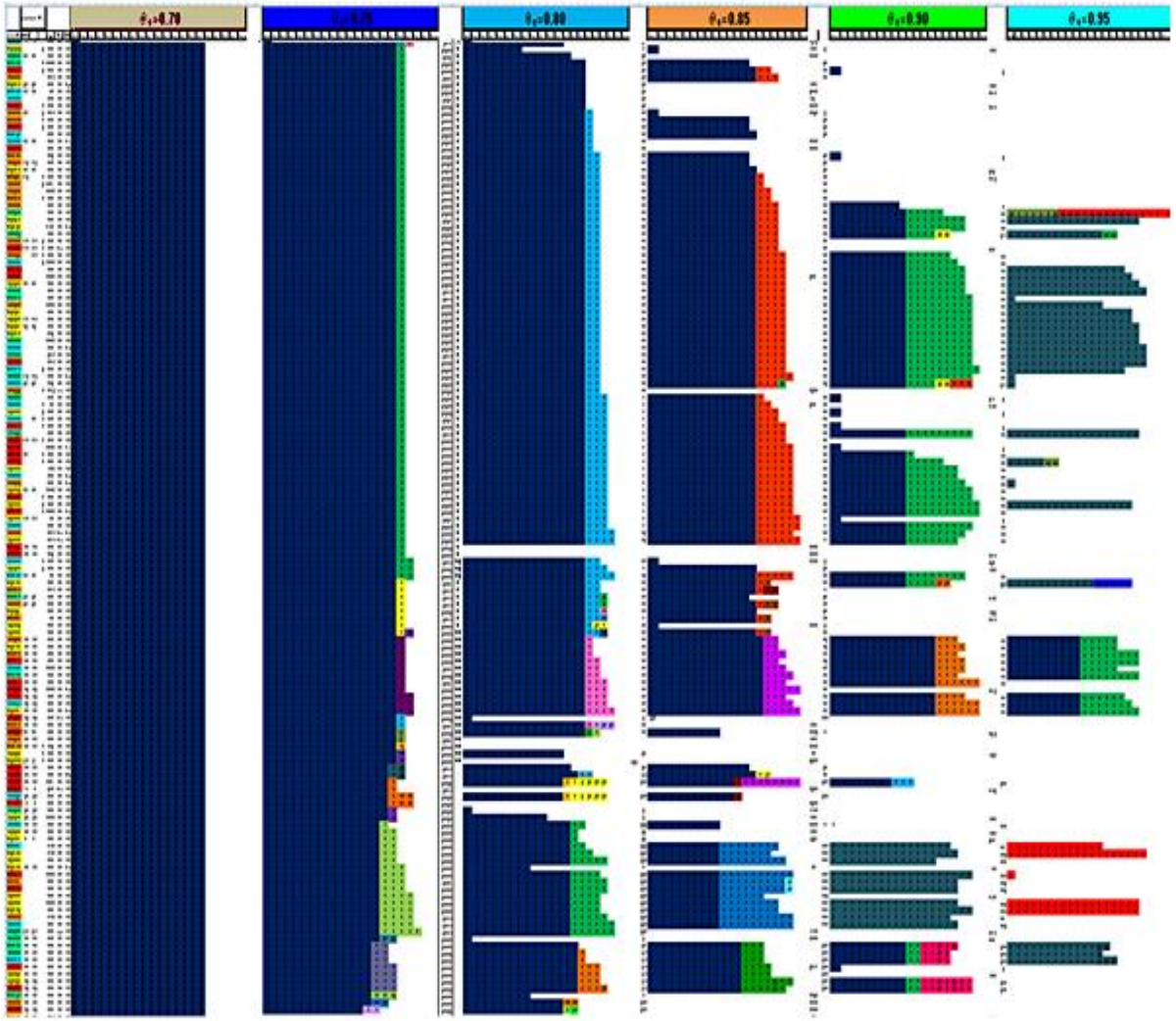


Figure 29.a A landscape of the "communities" obtained exploiting different θ_1 and θ_2. A detail of this picture is shown in figure 22b. In the first column there are the *A. thaliana* gene names according to their AGI code. In each other columns there are the numbers of the communities the genes belong to, obtained while increasing the θ_2 threshold. As expected, getting closer to θ_2=1 the communities begin to be fragmented and their sizes become thinner.

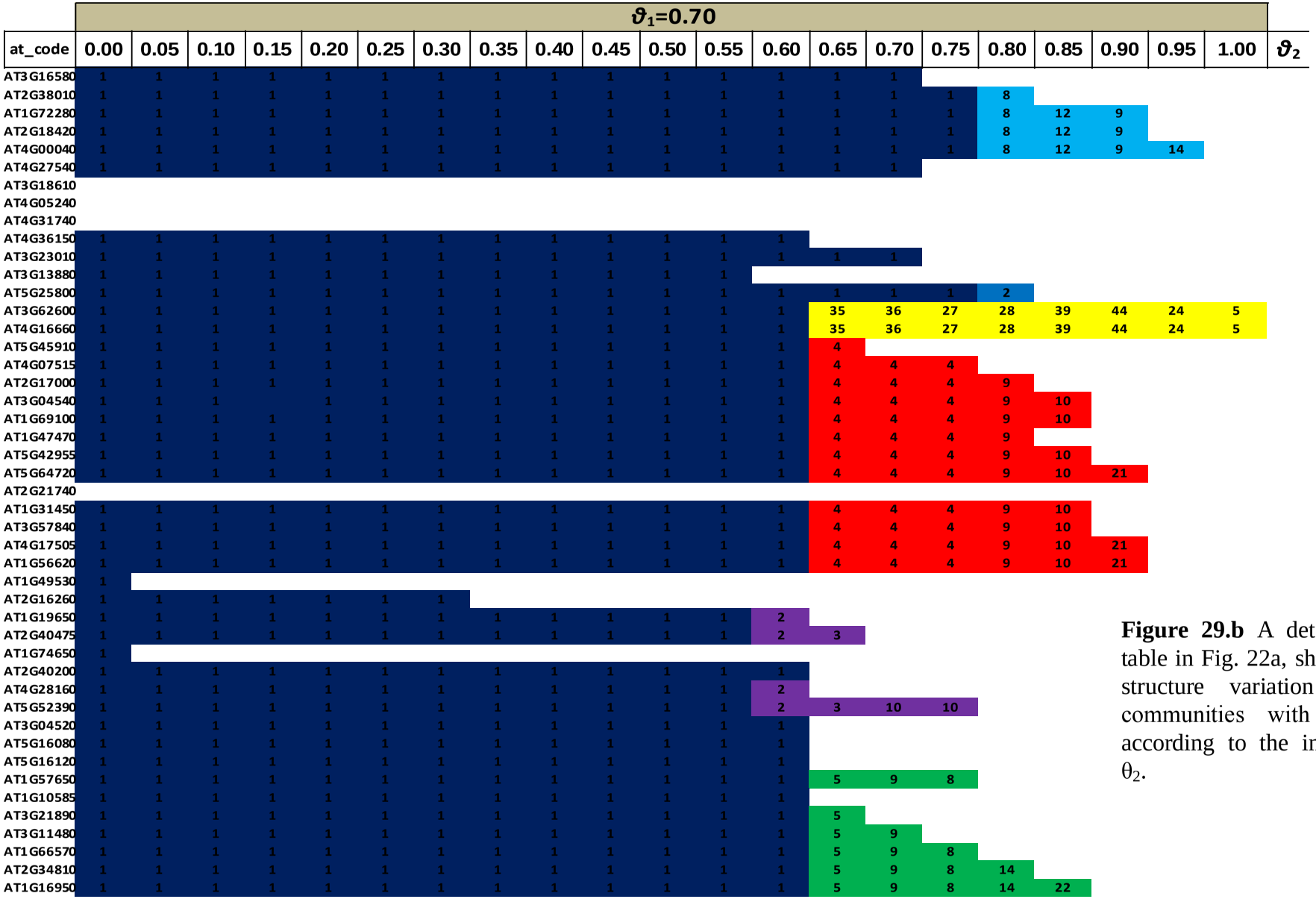


Figure 29.b A detail of the table in Fig. 22a, showing the structure variation of the communities with θ_1=0.70, according to the increase of θ_2.

In order to evaluate our approach we have compared the networks obtained to the cliques analyzed and reported in the previous section (see chapter 3.4). First of all we have considered the "communities" established with a Pearson's coefficient of r=|0.95| and θ_2=1, to be in the same settings of the data obtained by the clique approach. The results are very different, with a smaller number of networks to manage in the social approach and a shorter esecution time too (Tab. 17).

	Social network approach	Clique
Nr. of networks	26	9729
Time	0.1 h	1.5 h

Table 17 Comparison clique approach versus social network approach

We have hence evaluated the distribution of the clique elements inside the "communities" (Tab. 18). The clique number 1 has only element in common with a "community", however this latter is very specific: since formed by only 3 genes its functional annotation describes these 3 genes are involved in the same function (Tab.19)

	Size of the communities containing clique's element	Community elements shared with the Clique group	Clique size
Clique nr. 1	3	1	9
Clique nr. 2	36	1	25
Clique nr. 3	180	3	11
Clique nr. 4	180	4	14
Clique nr. 5	180	1	33

Table 18 The distribution of the clique elements inside the "communities"

ID	Social network id	Clique 1 ID	Note
AT1G67430	17	AT1G67430	60S ribosomal protein L17 (RPL17B); FUNCTIONS IN: structural constituent of ribosome;
AT1G01100	17		60S acidic ribosomal protein P1 (RPP1A); FUNCTIONS IN: structural constituent of ribosome;
AT3G61110	17		ARABIDOPSIS RIBOSOMAL PROTEIN S27 (ARS27A);

Table 19 Genes inside the community able to contain the clique nr. 1

Despite the θ_2=1, the communities were however too much large to be compared to the cliques obtained. This implies that probably the communities are incorporating genes not strictly correlated as happens in the clique approach. To test this hypothesis we have used lower θ_2 until a community was able to incorporate all the elements of a clique. The community 2, with 171 genes at r=|0.95| and θ_2=0.75 is able to collect all the elements of the Clique 1. Despite its very big size in comparison to the clique 1 size (9 elements) the GO terms analysis confirms the biological meaning of this communities (Tab. 20)

GO term	Description	P-value	Enrichment (N, B, n, b)
GO:0006412	translation	2.74E-132	41.54 (11978,196,128,87)
GO:0001510	RNA methylation	2.78E-102	54.59 (11978,108,128,63)
GO:0034645	cellular macromolecule biosynthetic process	2.63E-101	17.75 (11978,485,128,92)
GO:0009059	macromolecule biosynthetic process	4.02E-99	16.88 (11978,510,128,92)
GO:0032259	methylation	8.52E-84	27.62 (11978,227,128,67)
GO:0043414	macromolecule methylation	8.52E-84	27.62 (11978,227,128,67)
GO:0044260	cellular macromolecule metabolic process	1.30E-83	6.22 (11978,1745,128,116)
GO:0043170	macromolecule metabolic process	5.67E-82	5.72 (11978,1930,128,118)
GO:0009451	RNA modification	3.47E-80	29.33 (11978,201,128,63)
GO:0044267	cellular protein metabolic process	2.04E-78	9.78 (11978,899,128,94)
GO:0044249	cellular biosynthetic process	3.25E-74	7.06 (11978,1365,128,103)
GO:1901576	organic substance biosynthetic process	7.75E-72	6.70 (11978,1439,128,103)
GO:0019538	protein metabolic process	5.50E-70	7.60 (11978,1182,128,96)
GO:0009058	biosynthetic process	2.02E-68	6.21 (11978,1553,128,103)

GO:0044238	primary metabolic process	6.13E-68	4.26 (11978,2615,128,119)
GO:0071704	organic substance metabolic process	1.24E-67	3.83 (11978,3003,128,123)
GO:0044237	cellular metabolic process	2.61E-67	4.01 (11978,2825,128,121)
GO:0008152	metabolic process	2.04E-63	3.54 (11978,3248,128,123)
GO:0016070	RNA metabolic process	1.36E-62	11.68 (11978,585,128,73)
GO:0006139	nucleobase-containing compound metabolic process	1.93E-60	8.47 (11978,906,128,82)
GO:0034641	cellular nitrogen compound metabolic process	1.18E-56	7.61 (11978,1008,128,82)

Table 20 GO terms community 2 with 171 genes at r=|0.95| and θ_2=0.75

The community 2 with 141 genes at r=|0.95| and θ_2=0.80 collects all the elements of clique 2. The GO terms analysis (Tab. 21) for this group are however quite identical to its parent at θ_2=0.75 indicating that with lower θ_2 parameters the results are becoming less specific

GO term	Description	P-value	Enrichment (N, B, n, b)
GO:0006412	translation	7.34E-101	39.41 (11978,196,107,69)
GO:0001510	RNA methylation	5.69E-78	51.83 (11978,108,107,50)
GO:0034645	cellular macromolecule biosynthetic process	2.55E-77	16.85 (11978,485,107,73)
GO:0009059	macromolecule biosynthetic process	1.25E-75	16.02 (11978,510,107,73)
GO:0044260	cellular macromolecule metabolic process	3.03E-68	6.16 (11978,1745,107,96)
GO:0043170	macromolecule metabolic process	1.97E-67	5.68 (11978,1930,107,98)
GO:0032259	methylation	6.69E-66	26.63 (11978,227,107,54)
GO:0043414	macromolecule methylation	6.69E-66	26.63 (11978,227,107,54)
GO:0009451	RNA modification	1.20E-61	27.85 (11978,201,107,50)
GO:0044267	cellular protein metabolic process	1.45E-61	9.46 (11978,899,107,76)
GO:0044249	cellular biosynthetic process	1.79E-57	6.81 (11978,1365,107,83)
GO:0044238	primary metabolic process	4.22E-56	4.24 (11978,2615,107,99)

GO:1901576	organic substance biosynthetic process	1.39E-55	6.46 (11978,1439,107,83)
GO:0019538	protein metabolic process	2.56E-55	7.39 (11978,1182,107,78)
GO:0071704	organic substance metabolic process	3.75E-55	3.80 (11978,3003,107,102)
GO:0044237	cellular metabolic process	2.02E-54	3.96 (11978,2825,107,100)
GO:0009058	biosynthetic process	7.20E-53	5.98 (11978,1553,107,83)
GO:0008152	metabolic process	1.11E-51	3.52 (11978,3248,107,102)
GO:0016070	RNA metabolic process	2.05E-49	11.29 (11978,585,107,59)

Table 21 GO terms community 2 with 141 genes at r=|0.95| and θ_2=0.80

The community 1, with 410 genes at r=|0.95| and θ_2=0.75 can collect all the genes of the clique 3, which was related to the "Chloroplast activity" according to the functional notation, although the GO terms enrichment was void. According to the GO terms enrichment, and despite its very big size, this community can still address the function association of the collected genes to the energy metabolism (Tab. 22).

GO term	Description	P-value	Enrichment (N, B, n, b)
GO:0019682	glyceraldehyde-3-phosphate metabolic process	3.26E-81	40.87 (11978,81,199,55)
GO:0019288	isopentenyl diphosphate biosynthetic process, methylerythritol 4-phosphate pathway	3.26E-81	40.87 (11978,81,199,55)
GO:0009240	isopentenyl diphosphate biosynthetic process	3.26E-81	40.87 (11978,81,199,55)
GO:0046490	isopentenyl diphosphate metabolic process	3.26E-81	40.87 (11978,81,199,55)
GO:0006090	pyruvate metabolic process	9.79E-81	40.37 (11978,82,199,55)
GO:0006081	cellular aldehyde metabolic process	1.24E-71	27.71 (11978,126,199,58)
GO:0010027	thylakoid membrane organization	4.34E-70	43.95 (11978,63,199,46)
GO:0009668	plastid membrane organization	4.34E-70	43.95 (11978,63,199,46)

GO:0044802	single-organism membrane organization	1.52E-69	43.26 (11978,64,199,46)
GO:0061024	membrane organization	1.68E-68	41.95 (11978,66,199,46)
GO:0019637	organophosphate metabolic process	1.53E-67	11.98 (11978,407,199,81)
GO:0044237	cellular metabolic process	4.57E-63	3.37 (11978,2825,199,158)
GO:0008654	phospholipid biosynthetic process	1.26E-62	20.78 (11978,168,199,58)
GO:1901576	organic substance biosynthetic process	1.66E-62	5.10 (11978,1439,199,122)
GO:0044711	single-organism biosynthetic process	2.32E-62	5.95 (11978,1123,199,111)
GO:0006644	phospholipid metabolic process	2.84E-62	20.54 (11978,170,199,58)
GO:0009058	biosynthetic process	8.60E-61	4.81 (11978,1553,199,124)
GO:0044249	cellular biosynthetic process	9.16E-61	5.20 (11978,1365,199,118)

Table 22 GO terms community 1, with 410 genes at r=|0.95| and θ_2=0.75

The community 1, with 576 genes at r=|0.95| and θ_2=0.60 can collects all the genes of the clique 4. Since this community is the network incorporating the community 1 at θ_2=0.75 just described, the GO terms result is quite identical to the previous one, except for a lower enrichment (Tab. 23)

GO term	Description	P-value	Enrichment (N, B, n, b)
GO:0019682	glyceraldehyde-3-phosphate metabolic process	3.10E-106	39.55 (11978,81,258,69)
GO:0019288	isopentenyl diphosphate biosynthetic process, methylerythritol 4-phosphate pathway	3.10E-106	39.55 (11978,81,258,69)
GO:0009240	isopentenyl diphosphate biosynthetic process	3.10E-106	39.55 (11978,81,258,69)
GO:0046490	isopentenyl diphosphate metabolic process	3.10E-106	39.55 (11978,81,258,69)
GO:0006090	pyruvate metabolic process	1.93E-105	39.07 (11978,82,258,69)

GO:0010027	thylakoid membrane organization	1.52E-93	42.74 (11978,63,258,58)
GO:0009668	plastid membrane organization	1.52E-93	42.74 (11978,63,258,58)
GO:0044802	single-organism membrane organization	1.60E-92	42.07 (11978,64,258,58)
GO:0061024	membrane organization	1.18E-90	40.80 (11978,66,258,58)
GO:0006081	cellular aldehyde metabolic process	1.51E-89	26.53 (11978,126,258,72)
GO:0044711	single-organism biosynthetic process	1.15E-84	6.08 (11978,1123,258,147)
GO:0019637	organophosphate metabolic process	7.53E-83	11.52 (11978,407,258,101)
GO:1901576	organic substance biosynthetic process	5.02E-81	5.10 (11978,1439,258,158)
GO:0044249	cellular biosynthetic process	5.76E-79	5.20 (11978,1365,258,153)

Table 23 GO terms community 1, with 576 genes at r=|0.95| and θ_2=0.60

The last community with id 1, formed by 375 genes at r=|0.95| and θ_2=0.80 is the reduced version of the community 1 at θ_2=0.75, which is composed by 410 genes. As expected the GO terms are quite similar with a small increasing of the enrichment of the GO terms assigned to the energy pathway, due to the smaller size (Tab. 24).

GO term	Description	P-value	Enrichment (N, B, n, b)
GO:0019682	glyceraldehyde-3-phosphate metabolic process	5.44E-62	41.83 (11978,81,152,43)
GO:0019288	isopentenyl diphosphate biosynthetic process, methylerythritol 4-phosphate pathway	5.44E-62	41.83 (11978,81,152,43)
GO:0009240	isopentenyl diphosphate biosynthetic process	5.44E-62	41.83 (11978,81,152,43)
GO:0046490	isopentenyl diphosphate metabolic process	5.44E-62	41.83 (11978,81,152,43)
GO:0006090	pyruvate metabolic process	1.13E-61	41.32 (11978,82,152,43)
GO:0010027	thylakoid membrane organization	1.11E-57	47.53 (11978,63,152,38)
GO:0009668	plastid membrane organization	1.11E-57	47.53 (11978,63,152,38)

GO:0044802	single-organism membrane organization	2.71E-57	46.79 (11978,64,152,38)
GO:0006081	cellular aldehyde metabolic process	9.24E-57	28.77 (11978,126,152,46)
GO:0061024	membrane organization	1.51E-56	45.37 (11978,66,152,38)
GO:0019637	organophosphate metabolic process	1.12E-55	12.59 (11978,407,152,65)
GO:0044237	cellular metabolic process	1.79E-54	3.51 (11978,2825,152,126)
GO:0009987	cellular process	6.18E-51	2.92 (11978,3612,152,134)
GO:0006796	phosphate-containing compound metabolic process	2.22E-50	8.85 (11978,632,152,71)
GO:0008654	phospholipid biosynthetic process	4.93E-50	21.58 (11978,168,152,46)
GO:1901576	organic substance biosynthetic process	8.04E-50	5.20 (11978,1439,152,95)
GO:0006644	phospholipid metabolic process	9.12E-50	21.32 (11978,170,152,46)
GO:0006793	phosphorus metabolic process	2.24E-49	8.57 (11978,653,152,71)
GO:0044711	single-organism biosynthetic process	4.80E-49	6.03 (11978,1123,152,86)
GO:0044249	cellular biosynthetic process	1.54E-48	5.31 (11978,1365,152,92)
GO:0009058	biosynthetic process	6.42E-48	4.87 (11978,1553,152,96)
GO:0008152	metabolic process	3.31E-47	3.06 (11978,3248,152,126)
GO:0071704	organic substance metabolic process	1.40E-44	3.15 (11978,3003,152,120)
GO:0044763	single-organism cellular process	1.29E-43	3.69 (11978,2283,152,107)
GO:1901135	carbohydrate derivative metabolic process	5.37E-43	13.74 (11978,281,152,49)
GO:0090407	organophosphate biosynthetic process	8.91E-43	12.09 (11978,339,152,52)

Table 24 GO terms of the last community with id 1, formed by 375 genes at $r=|0.95|$ and $\theta_2=0.80$

After the comparison with the clique approach, the social network one hence appears able to identify the co-expressed genes in a faster way, without a limit in the dataset size but generally with less specific results.

The strategy that we propose is to merge the two approaches, exploiting the social one for preliminary analyses and exploiting the clique approaches for deeper investigations.

3.6 Databases comparison

Since several resources and detection tools are dedicated to collect and analyze the expression data of *Arabidopsis thaliana*, we have overviewed and compared the most referenced ones (see chapter 2.3), the data included and the tools offered (Tab. 25). We also have analyzed the variability in the results offered, highlighting the effect of the parameter settings (normalization methods, correlation techniques, and other), but also of the dataset considered, which may strongly affect the list of co-expressed genes according to the specific target gene chosen. In order to compare the results offered by the different web based resources, we used for the query the CESA7 gene (AT5G17420), coding for a xylem-specific cellulose synthase. We have chosen this gene because its co-expression with two other cellulose synthases, CESA4 (AT5G44030) and CESA8 (AT4G18780), has been confirmed experimentally and together they create a three units complex during the cell wall synthesis [American Society of Plant Biologists, 2003]. We used also AT5G06680, a gene implied in the gamma-tubulin complex whose Pearson correlations were detected as very variable according to the experiments dataset considered.

DB	Website	Release Data	Data Collection	Normalizzation method	Query for other species	Available
ACT	http://www.arabidopsis.leeds.ac.uk/ACT	2006	54 experiments from Nascarrays, organized in 122 AtGenome1 (8k probes) and 422 ATh1 (22k probes) arrays, covering several tissues, biotic and abiotic stress conditions, mutants and a range of developmental stages. After normalization, fixing to 20 the lowest signal unit detected in the	MAS 5.0	NO	NO

			experiments with at least one sample above this number, genes with an expression values under 20 in each experiment have been excluded.			
ATCOECIS	http://bioinformatics.psb.ugent.be/ATCOECIS	2009	322 (48x3 AtGenExpress Development and Tissue slides + 68x2 AtGenExpress Stress slides + 42 Birnbaum Root slides) Affymetrix ATH1 microarray slides in different tissues and under different experimental conditions, downloaded from Nottingham Arabidopsis Stock Centre	RMA	NO	YES
ATTED II	http://atted.jp/	2007	11171 slides in 737 experiments, including several tissues and biotic or abiotic stress conditions, from ArrayExpress. RNA-seq data, with 328 slides in 28 experiments.	RMA	NO	YES
BAR	http://bar.utoronto.ca/welcome.htm	2005	i) a self made microarrays database of about 175experimental samples; ii) the NASCarrays database collecting 392 samples and iii) the Atgen Express project microarray slides, organized in Hormone, Stress, Pathogen, Tissue, Seed, Root and Extend Tissue dataset; iv) a dataset organized according to A. thaliana ecotype	MAS 5.0	YES	YES
COP	http://webs2.kazusa.or.jp/kagiana/cop0911/	2010	5257 chips (CEL files or MAS5-processed data files from Affymetrix Gene Chip) form GEO and Array Express	MAS 5.0	YES	YES
CORNET	https://cornet.psb.ugent.be/	2009	425 experiments from all AtGenExpress ; 454 and 192 experiments from Microarray compendium 1 and 2, respectively; 256 experiments for abiotic stress series (cold, drought, genotoxic, heat, osmotic, oxidative, salt, UV-B, wounding); 69 for biotic stress series (Botrytis, Pseudomonas, Phytophthora, etc.) and, 336 ones with a combination of the abiotic and biotic stress; 235 experiments for developmental series (different tissues, developmental stages, developmental mutants); 72 experiments in flower, 212 in leaf, 258 in root, 83 in seed and 83 for the whole plant; 313 experiments with genetic modifications, in which transgenic lines are profiled (gene overexpression (knock-in), gene knock-out, transient transgene expression); 140 experiments with hormone treatment (ABA,	RMA	NO	YES

			brassinosteroids, GA, cytokinin, etc. and inhibitors)			
CRESS EXPRESS	http://cressexpress.org	2008	147 experiment with 486 slides from ATH1 and 80 from AG; 190 experiment with 1779 slides from ATH1. All data obtained from Affywatch I,II and III	RMA/MAS 5.0/GCRMA	NO	YES
CSB.DB	http://csbdb.mpimp-golm.mpg.de/csbdb/dbcor/ath.html	2004	nasc0271 (m0271) dataset, that contains 51 experiments covering several treatments and mutant characterization for 9694 genes; the atge0100 dataset collecting correlation profiles of 12200 valid measures of genes through 63 experiments about the development stages of several tissues and without mutants; the atge0200 dataset storing 13197 gene expression profiles in 60 experiments about above ground abiotic stress; atge0250 offering info about 15377 genes through 60 experiments about root abiotic stress.	GCOS	YES	YES
GENECAT	http://genecat.mpg.de/cgi-bin/Ainitiator.py	2008	Affymetrix 351 ATH1 microarrays containing 22810 probesets. Expression profiling tool uses 121 ATH1 microarrays generated during AtGenExpress project	RMA	NO	YES
GENEMANIA	http://www.genemania.org/	2008	Datasets obtained from literature and publicly databases as PFAM, Expression Omnibus (GEO), BioGRID, I2D, Pathway Commons, Memorial Sloan-Kettering Cancer Center, Human Protein Reference Database, HumanCyc, Systems Biology Center New York, IntAct, MINT, INTERPRO, NCI-Nature Pathway Interaction, iRefIndex and Reactome	NOT DEFINED	YES	YES
GENEN VESTIGATOR	https://www.genevestigator.com/gv/	2004	Data collect from Atgen Express, FGCZ, GEO, TAIR, ArrayExpress, Nasc, Gruissem Lab and other, while the chip platformsused for microarrays are the Ath1 22k, with 9848 experimental microarray slides available, the Ag 8k with 92 and the Agro1 whole genomic tilling array with 53 slides respectively	RMA/MAS 5.0	YES	YES
PlaNet	http://aranet.mpimp-golm.mpg.de/	2011	468 dual channel and 606 single channel microarrays slides obtained fromTAIR. All these data have been arranged in specific sets gathering in total 308 experiments, each one with at least 10, 5 arrays for dual channel and single channel essay respectively	NOT DEFINED	YES	YES

Table 25 Databases of co-expression compared in our analysis

For both genes, the analyses were performed using the default settings proposed by each platform and in this way we have collected the top 20 co-expressed genes resulting from each analysis on each database (Tab. 26, 27), ranked according to the specific correlation method proposed by each platform (Fig. 30a-e). Using CESA7, despite the relevant differences in the dataset size, correlation and normalization methods proposed by each database, GENEVESTIGATOR, ATTED, COP, GENECAT, BAR and CRESSEXPRESS share with all the other websites, about ~50% of their genes in the results and often, this value reaches or overcomes the 70% when considering couple comparisons, as it happens between COP, BAR and GENECAT, CSB.DB, PLANET, CORNET and GENEMANIA outputs instead, have less than 65% of elements shared with the results proposed by all the other databases. This can be explained by the fact that GENEMANIA and PLANET are not offering a specific ranking to list the co-expressed genes, but they are more focused on defining co-expressed gene modules. From a quality viewpoint, the presence of CESA 4 (AT5G44030) and CESA8 (AT4G18780) in the results of the CESA7 (AT5g17420) queries [Eckardt N. A. et al, 2003] underlines the prediction skill of each database. As shown in table 26, only CSB.DB and PLANET seem to have some issues in the query results, but we have to say that the first one does not show CESA8 because the probe of this gene was not included in the dataset exploited for this analysis, while PLANET does not show CESA4 (AT5G44030) in the top 20, despite it belongs to the cluster shown in its website result, simply because no rank has been proposed. Beyond these two particular databases, although in different rank positions, all the platforms confirm the co-expression of the CESA4-7-8 complex, and in the cases of CORNET, GENEVESTIGATOR, GENEMANIA, GENECAT, CRESSEXPRESS (RMA and gcRMA) and ATTED, where the rank positions of their co-expressed genes have been clearly defined by their p-value correlation methods, CESA 4 and

CESA 8 are listed in the first three positions, underlining the efficiency of these specific databases. Interestingly, collecting the top 20 co-expressed genes from each platform using AT5G06680 as query, there is not a database output very similar to another one as it happens for CESA7, and moreover the average of shared genes among the platform outputs does not exceed the 10% (Tab. 27). So, although using the same datasets and parameters, the similarity among the databases change totally when using CESA7 or AT5G06680, and the decreasing in the number of shared co-expressed genes can be very huge, as it happens between COP and BAR, where this value moves from 16 to 1. This underlines that the results proposed by the platforms must be compared among them since the common parameters developed to extract co-expressed gene lists can produce very different information. So, one single answer from only one platform is not enough, since the co-expression profile of some genes may be very inflected by the conditions of the experiments used for the dataset building, as seen for AT5G06680, while this not happens for gene like CESA7, where the co-expression network shown in the queries is less variable, and probably less conditions depending. In fact, despite some huge differences in the datasets size and experiments composition (i.e. passing from 11171 slides in GENEVESTIGATOR to 351 of GENECAT), CESA7 co-expression network remains confirmed among the platforms, while for AT5G06680 the co-expression profile may be harder to establish due to a high modulated by conditions expression, or simply due to some limits in the microarray signal detection. Beyond the dataset composition, normalization has a strong influence on the results too, as seen for AT5G06680 in Cressexpress database using the dataset version 3.1, normalized with GCRMA, and the dataset version 3.2, normalized MAS5.0, where despite the lacking of only 1 experiment out of 115 between the two versions during the analyses, there is only one gene shared by the two co-expression lists. Since the

query results of all these platforms are very influenced by technically aspects and probably by the expression modulation of the considered gene, this overview underlines that the data mining should include the “databases mining”, i.e. it is necessary to move from the idea of analyzing one single collection, to the one of comparing all or the majority of the available resources. Therefore, considering the high number of resources available, databases diversification must be exploited as an evaluation tool to ensure results robustness.

Our results underline that a correct data mining on gene co-expression analysis should include an appropriate comparison between the available resources, i.e. it is necessary to move from the idea of analyzing one single collection, to the one of comparing all or the majority of the available resources. So, considering the high number of available platforms, databases diversification must be exploited as a global opportunity to ensure results robustness.

	PLANET	ATTED	BAR	COP	CORNET	CRESS RMA	CRESS GCRMA	CRESS MASS	CSB	GENE CAT	GENE MANIA	GENEVE STIGATOR	NR OF SLIDES	ID PARAMETERS	AVERAGE OF SHARED GENES AMONG THE DATABASES
PLANET	20	3	3	3	3	5	5	2	3	3	1	5	1074	-	3.27
ATTED	3	20	11	11	9	14	12	14	10	15	4	14	11171	-	11.40
BAR	3	11	20	16	7	11	9	11	7	14	2	10	405	A	9.80
COP	3	11	16	20	6	11	8	11	8	16	2	10	5272	-	9.90
CORNET	3	9	7	6	20	12	11	11	7	9	2	11	??	B	8.50
CRESS RMA	5	14	11	11	12	20	13	13	10	14	4	14	1799	C	11.60
CRESS GCRMA	5	12	9	8	11	13	20	10	8	11	3	13			9.80
CRESS MASS	2	14	11	11	11	13	10	20	9	14	3	13			10.90
CSB	3	10	7	8	7	10	8	9	20	8	4	8	??	D	7.90
GENECAT	3	15	14	16	9	14	11	14	8	20	4	13	351	-	11.80
GENE MANIA	1	4	2	2	2	4	3	3	4	4	20	4	??	-	3.20
GENEVE STIGATOR	5	14	10	10	11	14	13	13	8	13	4	20	9211	E	11.00
NR OF SLIDES	1074	11171	405	5272	??	1799			??	351	??	9211	NORMALIZATION:		
ID PARAMETERS	-	-	A	-	B	C			D	-	-	E	RMA	MAS 5.0	
AVERAGE OF SHARED GENES AMONG THE DATABASES	3.27	11.40	9.80	9.90	8.50	11.60	9.80	10.90	7.90	11.80	3.20	11.00	GCRMA	GCOS	

A=default settings; dataset:atgenExpress Plus-Extended Tissue Compendium

B=default settings;Microarray compendium 2 TAIR10 (111 exp -no bias)

C=default settings; Cut-off value for Kolmogorov-Smirnov quality-control statistic 0.15; Select Tissue type "All"; Select Experiment "All"; Enter r-squared threshold for pathway-level co-expression: 0.36; Three datasets were exploited, respectively:

-CRESS RMA: Version 3.0 (1779 arrays - RMA processing)

-CRESS GCRMA: Version 3.1 (1779 arrays - GCRMA processing)

-CRESS MAS5.0 Version 3.2 (1779 arrays - MAS5 processing)

D="atge0100": developmental series (only WT); Ath1 chip; AtGen Express; 12200 genes; non parametric Spearman's Rho rank correlation positive, significant co-responding genes (Bonferroni's correction)

E=default settings; ATH1 platform; general filters "all data"; samples selected 9211

Table 26 Summarizing comparison of the databases, querying CESA7 gene and checking top 20 co-expressed genes. In the table, are specified the number of slides stored in each database and the average of shared genes among the databases, excluding the self matching value (yellow boxes). At the bottom, are shown the parameters exploited during the comparison.

	PLANET	ATTED	BAR	COP	CORNET	CRESS RMA	CRESS GCRMA	CRESS MAS5	CSB	GENE CAT	GENE MANIA	GENEVE STIGATOR	NR OF SLIDES	ID PARAMETERS	AVERAGE OF SHARED GENES AMONG THE DATABASES
PLANET	20	0	0	0	0	0	0	1	0	2	1	0	1074	-	0.36
ATTED	0	20	2	2	2	2	2	4	0	3	1	4	11171	-	2.00
BAR	0	2	20	1	0	0	1	1	0	0	1	1	405	A	0.64
COP	0	2	1	20	0	0	0	1	0	1	0	2	5272	-	0.64
CORNET	0	2	0	0	20	2	2	1	1	0	0	1	??	B	0.82
CRESS RMA	0	2	0	0	2	20	4	2	1	4	1	3	1799	C	1.73
CRESS GCRMA	0	2	1	0	2	4	20	1	1	3	0	3			1.55
CRESS MAS5	1	4	1	1	1	2	1	20	3	3	0	3			1.82
CSB	0	0	0	0	1	1	1	3	20	0	1	1	??	F	0.73
GENECAT	2	3	0	1	0	4	3	3	0	20	2	1	351	-	1.73
GENE MANIA	1	1	1	0	0	1	0	0	1	2	20	1	??	-	0.73
GENEVE STIGATOR	0	4	1	2	1	3	3	3	1	1	1	20	9211	G	1.82
NR OF SLIDES	1074	11171	405	5272	??	1799			??	351	??	9211			
ID PARAMETERS	-	-	A	-	B	C			F	-	-	G			
AVERAGE OF SHARED GENES AMONG THE DATABASES	0.36	2.00	0.64	0.64	0.82	1.73	1.55	1.82	0.73	1.73	0.73	1.82			

NORMALIZATION:

RMA MAS 5.0

GCRMA GCOS

A=default settings; dataset:atgenExpress Plus-Extended Tissue Compendium

B=default settings;Microarray compendium 2 TAIR10 (111 exp -no bias)

C=default settings; Cut-off value for Kolmogorov-Smirnov quality-control statistic 0.15; Select Tissue type "All"; Select Experiment "All"; Enter r-squared threshold for pathway-level co-expression: 0.36; Three datasets were exploited, respectively:

-CRESS RMA: Version 3.0 (1779 arrays - RMA processing)

-CRESS GCRMA: Version 3.1 (1779 arrays - GCRMA processing)

-CRESS MAS5.0 Version 3.2 (1779 arrays - MAS5 processing)

D="atge0100": developmental series (only WT); Ath1 chip; AtGen Express; 12200 genes; non parametric Spearman's Rho rank correlation positive, significant co-responding genes (Bonferroni's correction)

E=default settings; ATH1 platform; general filters "all data"; samples selected 9211

Table 27 Summarizing comparison of the databases, querying AT5G06680 gene and checking top 20 co-expressed genes. In the table, are specified the number of slides stored in each database and the average of shared genes among the databases, excluding the self matching value (yellow boxes). At the bottom, are shown the parameters exploited during the comparison

PLANET		ATTED		BAR	
AGI	C.V.	AGI	MR	AGI	r-value
AT4G18640	Not available	AT5G15630	1	AT5G15630	0.994
AT4G18780		AT5G44030	1.04	AT5G03170	0.989
AT1G63520		AT4G18780	2.05	AT3G50220	0.988
AT3G08490		AT5G54690	2.08	AT5G54690	0.988
AT3G27200		AT3G16920	3.02	AT3G18660	0.985
AT3G45870		AT3G18660	3.07	AT3G16920	0.980
AT1G12260		AT2G37090	4	AT1G27380	0.979
AT1G05310		AT2G38080	4.02	AT4G18780	0.977
AT1G24030		AT5G03170	4.05	AT5G44030	0.977
AT1G58070		AT1G27440	5.05	AT3G15050	0.973
AT3G52900		AT5G60020	5.07	AT1G07120	0.970
AT2G38080		AT5G60720	6	AT4G27435	0.969
AT5G45970		AT5G01360	6	AT2G38080	0.968
AT3G59690		AT1G79620	6.02	AT2G29130	0.966
AT1G33800		AT4G18640	6.08	AT1G63910	0.966
AT1G09440		AT3G62020	8.01	AT5G67210	0.965
AT5G03260		AT5G60490	8.02	AT1G22480	0.964
AT3G16920		AT4G28500	8.05	AT4G28500	0.962
AT5G51890		AT3G59690	9.07	AT1G08340	0.962
AT5G40020		AT4G27435	9.08	AT3G62020	0.961

Figure 30.a Top 20 co-expressed genes querying for CESA7 in each database, ranked by the specific statistical robustness and correlation method proposed in each platforms. CESA4 (yellow) and CESA8 (orange) are colored to highlight the precision of the query results of each database.

AGI=Arabidopsis genome initiative code; C.V=Correlation value; MR=Mutual Ranking; r-value=Pearson coefficient; r2=Pearson coefficient squared; VF=Vertex F-measure; %ile=percentile; Spearman=Spearman coefficient; Weight= *weight of the relation*.

COP			CORNET		
AGI	VF	%ile	AGI	r-value	p-value
AT1G27380	0.98	97.80	AT4G18780	0.920	2:1.11E-36
AT4G27435	0.98	97.80	AT5G44030	0.920	2:2.05E-36
AT2G41610	0.95	97.00	AT5G60020	0.910	2:7.17E-35
AT3G16920	0.95	97.00	AT5G54690	0.890	2:8.19E-30
AT3G50220	0.95	97.00	AT5G60720	0.860	2:4.57E-26
AT4G28500	0.95	97.00	AT5G03170	0.860	2:2.01E-25
AT5G15630	0.95	97.00	AT5G01360	0.850	2:1.82E-23
AT5G03170	0.93	96.40	AT3G62020	0.830	2:3.62E-21
AT5G44030	0.93	96.40	AT5G15630	0.820	2:4.72E-20
AT1G22480	0.91	95.60	AT1G62990	0.810	2:3.16E-19
AT5G67210	0.91	95.60	AT1G54790	0.790	2:1.21E-16
AT3G15050	0.89	94.60	AT1G73640	0.790	2:1.92E-16
AT2G29130	0.88	94.00	AT3G50220	0.790	2:1.92E-16
AT1G32770	0.88	94.00	AT5G03260	0.780	2:1.44E-15
AT2G38080	0.88	94.00	AT5G47530	0.770	2:9.83E-15
AT5G54690	0.87	93.50	AT1G32100	0.770	2:1.21E-14
AT3G18660	0.86	93.10	AT2G03200	0.760	2:9.15E-14
AT1G09610	0.84	91.90	AT4G08160	0.760	2:2.43E-13
AT1G27440	0.78	88.60	AT1G58070	0.740	2:3.3E-12
AT4G18780	0.73	85.5	AT2G27740	0.740	2:4.73E-12

Figure 30.b Top 20 co-expressed genes querying for CESA7 in each database, ranked by the specific statistical robustness and correlation method proposed in each platforms. CESA4 (yellow) and CESA8 (orange) are colored to highlight the precision of the query results of each database.

AGI=Arabidopsis genome initiative code; C.V=Correlation value; MR=Mutual Ranking; r-value=Pearson coefficient; r2=Pearson coefficient squared; VF=Vertex F-measure; %ile=percentile; Spearman=Spearman coefficient; Weight= *weight of the relation*.

CRESS RMA			CRESS GCRMA		
AGI	p-value	r2	AGI	p-value	r2
AT4G18780	2.15E-290	0.538	AT4G18780	0	0.758
AT5G44030	2.74E-284	0.531	AT3G16920	0	0.732
AT5G54690	2.99E-281	0.527	AT5G44030	6.66E-306	0.680
AT5G15630	3.75E-277	0.522	AT2G38080	2.00E-285	0.654
AT3G16920	3.93E-275	0.519	AT5G60020	2.24E-262	0.623
AT5G60020	1.25E-274	0.518	AT3G62020	2.14E-242	0.594
AT5G60720	5.56E-274	0.518	AT2G37090	8.93E-229	0.573
AT5G03170	7.80E-269	0.511	AT5G03260	1.28E-224	0.566
AT1G27440	2.76E-265	0.506	AT2G28760	6.49E-212	0.545
AT2G38080	4.33E-262	0.502	AT5G54690	1.52E-205	0.534
AT4G27435	1.63E-255	0.493	AT5G40020	9.58E-200	0.523
AT5G01360	3.99E-254	0.491	AT4G08160	5.59E-192	0.509
AT1G32100	4.34E-254	0.491	AT5G01360	1.33E-190	0.507
AT5G16600	1.28E-251	0.488	AT1G32100	5.78E-184	0.494
AT2G37090	7.86E-251	0.487	AT2G27740	5.26E-183	0.493
AT5G03260	1.26E-249	0.485	AT5G15630	1.10E-178	0.484
AT2G29130	2.51E-248	0.483	AT5G18970	4.33E-178	0.483
AT3G50220	6.21E-248	0.483	AT3G18660	3.30E-177	0.481
AT3G62020	2.68E-246	0.480	AT4G35350	2.10E-170	0.468
AT1G24030	4.63E-245	0.479	AT4G27435	1.85E-168	0.464

Figure 30.c Top 20 co-expressed genes querying for CESA7 in each database, ranked by the specific statistical robustness and correlation method proposed in each platforms. CESA4 (yellow) and CESA8 (orange) are colored to highlight the precision of the query results of each database.

AGI=Arabidopsis genome initiative code; C.V=Correlation value; MR=Mutual Ranking; r-value=Pearson coefficient; r2=Pearson coefficient squared; VF=Vertex F-measure; %ile=percentile; Spearman=Spearman coefficient; Weight= *weight of the relation*.

CRESS MAS			CSB		
AGI	p-value	r2	AGI	spearman	p-value
AT5G15630	0	0.777	AT5G44030	0.908	0
AT5G54690	0	0.776	AT2G38080	0.901	0
AT5G44030	0	0.751	AT5G15630	0.897	0
AT1G27440	0	0.652	AT2G28760	0.876	0
AT2G37090	0	0.627	AT5G03170	0.872	0
AT5G60720	0	0.625	AT5G60720	0.837	0
AT4G27435	0	0.614	AT5G54690	0.836	0
AT4G18780	1.25E-298	0.571	AT1G47410	0.827	2.22E-16
AT5G03170	7.21E-285	0.554	AT4G18640	0.817	4.44E-16
AT3G62020	1.11E-258	0.519	AT1G32100	0.811	6.66E-16
AT5G60490	4.20E-258	0.518	AT1G33800	0.801	3.11E-15
AT3G50220	8.85E-256	0.515	AT5G59290	0.786	2.29E-14
AT5G01360	6.57E-254	0.513	AT5G03260	0.778	6.68E-14
AT2G38080	1.02E-242	0.497	AT1G27440	0.775	8.73E-14
AT5G47530	2.70E-231	0.480	AT5G60490	0.762	4.11E-13
AT2G41610	6.66E-227	0.474	AT5G67210	0.755	8.61E-13
AT5G16490	3.63E-224	0.469	AT4G27435	0.742	3.42E-12
AT3G18660	1.20E-222	0.467	AT5G14510	0.727	1.53E-11
AT1G73640	4.26E-218	0.460	AT2G38320	0.668	2.30E-09
AT1G08340	1.82E-211	0.450	AT1G20850	0.666	2.65E-09

Figure 30.d Top 20 co-expressed genes querying for CESA7 in each database, ranked by the specific statistical robustness and correlation method proposed in each platforms. CESA4 (yellow) and CESA8 (orange) are colored to highlight the precision of the query results of each database.

AGI=Arabidopsis genome initiative code; C.V=Correlation value; MR=Mutual Ranking; r-value=Pearson coefficient; r2=Pearson coefficient squared; VF=Vertex F-measure; %ile=percentile; Spearman=Spearman coefficient; Weight= *weight of the relation*.

GENEMANIA		GENEVESTIGATOR		GENE CAT	
AGI	weight	AGI	r-value	AGI	r-value
AT4G18780	0.102	AT5G44030	0.900	AT5G15630	0.950
AT5G44030	0.100	AT5G15630	0.880	AT5G44030	0.934
AT5G03170	0.074	AT4G18780	0.880	AT4G18780	0.933
AT2G25540	0.073	AT5G54690	0.870	AT5G54690	0.922
AT3G16920	0.071	AT5G60020	0.830	AT5G03170	0.918
AT2G32540	0.069	AT2G37090	0.830	AT3G16920	0.899
AT2G32530	0.069	AT5G01360	0.820	AT1G27440	0.895
AT4G24010	0.069	AT5G60720	0.820	AT5G60720	0.894
AT2G32610	0.069	AT2G38080	0.810	AT3G18660	0.891
AT2G33100	0.069	AT3G16920	0.810	AT2G38080	0.890
AT1G32180	0.069	AT5G03170	0.790	AT3G62020	0.877
AT4G15290	0.069	AT1G27440	0.780	AT4G28500	0.875
AT4G15320	0.069	AT5G03260	0.770	AT2G37090	0.872
AT4G38190	0.069	AT3G50220	0.760	AT2G41610	0.865
AT4G23990	0.069	AT3G18660	0.750	AT4G27435	0.863
AT4G24000	0.069	AT4G08160	0.730	AT5G60020	0.863
AT5G60720	0.068	AT1G79620	0.720	AT3G50220	0.858
AT5G54690	0.062	AT5G40020	0.720	AT1G09610	0.836
AT2G37090	0.060	AT1G132100	0.710	AT3G15050	0.831
AT2G32620	0.005	AT1G08340	0.710	AT1G27380	0.823

Figure 30.e Top 20 co-expressed genes querying for CESA7 in each database, ranked by the specific statiscal robusteness and correlation method proposed in each platforms. CESA4 (yellow) and CESA8 (orange) are colored to highlight the precision of the query results of each database.

AGI=Arabidopsis genome initiative code; C.V=Correlation value; MR=Mutual Ranking; r-value=Pearson coefficient; r2=Pearson coefficient squared; VF=Vertex F-measure; %ile=percentile; Spearman=Spearman coefficient; Weight= *weight of the relation.*

PLANET		ATTED		BAR		COP			
AGI	Corr. Value	AGI	MR	AGI	r-value	AGI	VF	%ile	CC
AT4G18640	Not available	AT5G15630	1	AT5G15630	0.994	AT1G27380	0.98	97.80	0.860
AT4G18780		AT5G44030	1.04	AT5G03170	0.989	AT4G27435	0.98	97.80	0.940
AT1G63520		AT4G18780	2.05	AT3G50220	0.988	AT2G41610	0.95	97.00	0.900
AT3G08490		AT5G54690	2.08	AT5G54690	0.988	AT3G16920	0.95	97.00	0.930
AT3G27200		AT3G16920	3.02	AT3G18660	0.985	AT3G50220	0.95	97.00	0.910
AT3G45870		AT3G18660	3.07	AT3G16920	0.980	AT4G28500	0.95	97.00	0.840
AT1G12260		AT2G37090	4	AT1G27380	0.979	AT5G15630	0.95	97.00	0.970
AT1G05310		AT2G38080	4.02	AT4G18780	0.977	AT5G03170	0.93	96.40	0.940
AT1G24030		AT5G03170	4.05	AT5G44030	0.977	AT5G44030	0.93	96.40	0.940
AT1G58070		AT1G27440	5.05	AT3G15050	0.973	AT1G22480	0.91	95.60	0.850
AT3G52900		AT5G60020	5.07	AT1G07120	0.970	AT5G67210	0.91	95.60	0.850
AT2G38080		AT5G60720	6	AT4G27435	0.969	AT3G15050	0.89	94.60	0.890
AT5G45970		AT5G01360	6	AT2G38080	0.968	AT2G29130	0.88	94.00	0.840
AT3G59690		AT1G79620	6.02	AT2G29130	0.966	AT1G32770	0.88	94.00	0.830
AT1G33800		AT4G18640	6.08	AT1G63910	0.966	AT2G38080	0.88	94.00	0.930
AT1G09440		AT3G62020	8.01	AT5G67210	0.965	AT5G54690	0.87	93.50	0.920
AT5G03260		AT5G60490	8.02	AT1G22480	0.964	AT3G18660	0.86	93.10	0.880
AT3G16920		AT4G28500	8.05	AT4G28500	0.962	AT1G09610	0.84	91.90	0.860
AT5G51890		AT3G59690	9.07	AT1G08340	0.962	AT1G27440	0.78	88.60	0.890
AT5G40020		AT4G27435	9.08	AT3G62020	0.961	AT4G18780	0.73	85.5	0.870

Figure 31.a complete list of the CESA7 query results as offered by each database.

The table shows also the statistical parameters exploited to describe the results.

AGI=Arabidopsis genome initiative code; C.V=Correlation value; MR=Mutual Ranking; r-value=Pearson coefficient; r2=Pearson coefficient squared; VF=Vertex F-measure; %ile=percentile; Spearman=Spearman coefficient; Weight= *weight of the relation*.

CORNET			CRESS RMA					
AGI	r-value	p-value	AGI	p-value	slope	T	DOF	r2
AT4G18780	0.920	2:1.11E-36	AT4G18780	2.15E-290	0.45	44.74	1717	0.538
AT5G44030	0.920	2:2.05E-36	AT5G44030	2.74E-284	0.53	44.06	1717	0.531
AT5G60020	0.910	2:7.17E-35	AT5G54690	2.99E-281	0.51	43.72	1717	0.527
AT5G54690	0.890	2:8.19E-30	AT5G15630	3.75E-277	0.56	43.26	1717	0.522
AT5G60720	0.860	2:4.57E-26	AT3G16920	3.93E-275	0.39	43.04	1717	0.519
AT5G03170	0.860	2:2.01E-25	AT5G60020	1.25E-274	0.50	42.98	1717	0.518
AT5G01360	0.850	2:1.82E-23	AT5G60720	5.56E-274	0.77	42.91	1717	0.518
AT3G62020	0.830	2:3.62E-21	AT5G03170	7.80E-269	0.45	42.34	1717	0.511
AT5G15630	0.820	2:4.72E-20	AT1G27440	2.76E-265	0.76	41.94	1717	0.506
AT1G62990	0.810	2:3.16E-19	AT2G38080	4.33E-262	0.33	41.59	1717	0.502
AT1G54790	0.790	2:1.21E-16	AT4G27435	1.63E-255	0.54	40.86	1717	0.493
AT1G73640	0.790	2:1.92E-16	AT5G01360	3.99E-254	0.48	40.70	1717	0.491
AT3G50220	0.790	2:1.92E-16	AT1G32100	4.34E-254	0.52	40.70	1717	0.491
AT5G03260	0.780	2:1.44E-15	AT5G16600	1.28E-251	0.74	40.42	1717	0.488
AT5G47530	0.770	2:9.83E-15	AT2G37090	7.86E-251	0.48	40.34	1717	0.487
AT1G32100	0.770	2:1.21E-14	AT5G03260	1.26E-249	0.66	40.20	1717	0.485
AT2G03200	0.760	2:9.15E-14	AT2G29130	2.51E-248	0.63	40.06	1717	0.483
AT4G08160	0.760	2:2.43E-13	AT3G50220	6.21E-248	0.50	40.02	1717	0.483
AT1G58070	0.740	2:3.3E-12	AT3G62020	2.68E-246	0.45	39.83	1717	0.480
AT2G27740	0.740	2:4.73E-12	AT1G24030	4.63E-245	0.27	39.70	1717	0.479

Figure 31.b complete list of the CESA7 query results as offered by each database.

The table shows also the statistical parameters exploited to describe the results.

AGI=Arabidopsis genome initiative code; C.V=Correlation value; CC=Cosine correlation value; MR=Mutual Ranking; r-value=Pearson coefficient; VF=Vertex F-measure; %ile=percentile; T=t-test value; DOF=degree of freedom; r2=Pearson coefficient squared; Spearman=Spearman coefficient; Weight=*weight of the relation*

CRESS GCRMA						CRESS MAS					
AGI	p-value	slope	T	DOF	r2	AGI	p-value	slope	T	DOF	r2
AT4G18780	0	0.95	62.04	1228	0.758	AT5G15630	0	0.82	74.95	1613	0.777
AT3G16920	0	0.74	57.84	1228	0.732	AT5G54690	0	0.85	74.69	1613	0.776
AT5G44030	6.66E-306	0.84	51.05	1228	0.680	AT5G44030	0	0.78	69.80	1613	0.751
AT2G38080	2.00E-285	0.63	48.20	1228	0.654	AT1G27440	0	1.18	54.91	1613	0.652
AT5G60020	2.24E-262	0.70	45.05	1228	0.623	AT2G37090	0	0.69	52.10	1613	0.627
AT3G62020	2.14E-242	0.94	42.36	1228	0.594	AT5G60720	0	0.80	51.87	1613	0.625
AT2G37090	8.93E-229	0.92	40.54	1228	0.573	AT4G27435	0	0.77	50.68	1613	0.614
AT5G03260	1.28E-224	0.80	39.99	1228	0.566	AT4G18780	1.25E-298	0.48	46.32	1613	0.571
AT2G28760	6.49E-212	0.78	38.31	1228	0.545	AT5G03170	7.21E-285	0.66	44.73	1613	0.554
AT5G54690	1.52E-205	0.88	37.47	1228	0.534	AT3G62020	1.11E-258	0.61	41.72	1613	0.519
AT5G40020	9.58E-200	0.66	36.71	1228	0.523	AT5G60490	4.20E-258	0.70	41.66	1613	0.518
AT4G08160	5.59E-192	0.76	35.69	1228	0.509	AT3G50220	8.85E-256	0.41	41.39	1613	0.515
AT5G01360	1.33E-190	0.64	35.51	1228	0.507	AT5G01360	6.57E-254	0.63	41.18	1613	0.513
AT1G32100	5.78E-184	0.55	34.64	1228	0.494	AT2G38080	1.02E-242	0.44	39.89	1613	0.497
AT2G27740	5.26E-183	0.79	34.51	1228	0.493	AT5G47530	2.70E-231	0.65	38.58	1613	0.480
AT5G15630	1.10E-178	1.10	33.94	1228	0.484	AT2G41610	6.66E-227	0.72	38.08	1613	0.474
AT5G18970	4.33E-178	0.88	33.87	1228	0.483	AT5G16490	3.63E-224	0.76	37.77	1613	0.469
AT3G18660	3.30E-177	1.19	33.75	1228	0.481	AT3G18660	1.20E-222	0.54	37.59	1613	0.467
AT4G35350	2.10E-170	0.56	32.86	1228	0.468	AT1G73640	4.26E-218	0.64	37.07	1613	0.460
AT4G27435	1.85E-168	0.73	32.60	1228	0.464	AT1G08340	1.82E-211	0.63	36.31	1613	0.450

Figure 31.c A complete list of the CESA7 query results as offered by each database. The table shows also the statistical parameters exploited to describe the results.

AGI=Arabidopsis genome initiative code; C.V=Correlation value; CC=Cosine correlation value; MR=Mutual Ranking; r-value=Pearson coefficient; VF=Vertex F-measure; %ile=percentile; T=t-test value; DOF=degree of freedom; r2=Pearson coefficient squared; Spearman=Spearman coefficient; Weight=*weight of the relation*

CSB			GENE CAT		GENE MANIA		GENEVE STIGATOR	
AGI	spearman	p-value	AGI	r-value	AGI	weight	AGI	r-value
AT5G44030	0.908	0	AT5G15630	0.950	AT4G18780	0.102	AT5G44030	0.900
AT2G38080	0.901	0	AT5G44030	0.934	AT5G44030	0.100	AT5G15630	0.880
AT5G15630	0.897	0	AT4G18780	0.933	AT5G03170	0.074	AT4G18780	0.880
AT2G28760	0.876	0	AT5G54690	0.922	AT2G25540	0.073	AT5G54690	0.870
AT5G03170	0.872	0	AT5G03170	0.918	AT3G16920	0.071	AT5G60020	0.830
AT5G60720	0.837	0	AT3G16920	0.899	AT2G32540	0.069	AT2G37090	0.830
AT5G54690	0.836	0	AT1G27440	0.895	AT2G32530	0.069	AT5G01360	0.820
AT1G47410	0.827	2.22E-16	AT5G60720	0.894	AT4G24010	0.069	AT5G60720	0.820
AT4G18640	0.817	4.44E-16	AT3G18660	0.891	AT2G32610	0.069	AT2G38080	0.810
AT1G32100	0.811	6.66E-16	AT2G38080	0.890	AT2G33100	0.069	AT3G16920	0.810
AT1G33800	0.801	3.11E-15	AT3G62020	0.877	AT1G32180	0.069	AT5G03170	0.790
AT5G59290	0.786	2.29E-14	AT4G28500	0.875	AT4G15290	0.069	AT1G27440	0.780
AT5G03260	0.778	6.68E-14	AT2G37090	0.872	AT4G15320	0.069	AT5G03260	0.770
AT1G27440	0.775	8.73E-14	AT2G41610	0.865	AT4G38190	0.069	AT3G50220	0.760
AT5G60490	0.762	4.11E-13	AT4G27435	0.863	AT4G23990	0.069	AT3G18660	0.750
AT5G67210	0.755	8.61E-13	AT5G60020	0.863	AT4G24000	0.069	AT4G08160	0.730
AT4G27435	0.742	3.42E-12	AT3G50220	0.858	AT5G60720	0.068	AT1G79620	0.720
AT5G14510	0.727	1.53E-11	AT1G09610	0.836	AT5G54690	0.062	AT5G40020	0.720
AT2G38320	0.668	2.30E-09	AT3G15050	0.831	AT2G37090	0.060	AT1G132100	0.710
AT1G20850	0.666	2.65E-09	AT1G27380	0.823	AT2G32620	0.005	AT1G08340	0.710

Figure 31.d A complete list of the CESA7 query results as offered by each database. The table shows also the statistical parameters exploited to describe the results.

AGI=Arabidopsis genome initiative code; C.V=Correlation value; CC=Cosine correlation value; MR=Mutual Ranking; r-value=Pearson coefficient; VF=Vertex F-measure; %ile=percentile; T=t-test value; DOF=degree of freedom; r2=Pearson coefficient squared; Spearman=Spearman coefficient; Weight=*weight of the relation*

PLANET		ATTED		BAR	
AGI		AGI	MR	AGI	r-value
AT3G43690	Not Available	AT1G62020	3	AT1G09820	0.801
AT5G13300		AT2G21390	3	AT1G09290	0.8
AT3G18000		AT4G20740	8.9	AT2G29190	0.788
AT3G61850		AT4G09980	15.4	AT1G73820	0.784
AT5G05640		AT2G38770	16.9	AT2G38770	0.784
AT5G24600		AT1G65380	21.6	AT5G45790	0.78
AT3G07180		AT1G55325	31	AT5G02850	0.779
AT1G34065		AT5G18960	35.7	AT5G55040	0.777
AT3G49310		AT4G02070	44.8	AT5G55660	0.773
AT4G35220		AT1G26370	45.3	AT2G33500	0.772
AT4G39120		AT3G06340	48.7	AT3G19120	0.768
AT5G64340		AT4G24490	51.8	AT3G27520	0.765
AT5G09810		AT3G63290	53.4	AT3G06340	0.752
AT5G18960		AT5G66770	55.1	AT3G55320	0.746
AT3G17310		AT1G55350	60.9	AT2G33610	0.746
AT2G46590		AT3G18524	61.4	AT5G17410	0.744
AT5G13050		AT3G20010	68.3	AT4G22140	0.742
AT1G23890		AT5G10020	70.7	AT1G08610	0.741
AT5G04110		AT4G20910	74.8	AT1G30460	0.741
AT1G67570		AT3G01380	79.1	AT1G28420	0.74

Figure 32.a A complete list of the AT5G06680 query results as offered by each database.

The table shows also the statistical parameters exploited to describe the results.

AGI=Arabidopsis genome initiative code; C.V=Correlation value; CC=Cosine correlation value; MR=Mutual Ranking; r-value=Pearson coefficient; VF=Vertex F-measure; %ile=percentile; T=t-test value; DOF=degree of freedom; r2=Pearson coefficient squared; Spearman=Spearman coefficient; Weight=*weight of the relation*

COP				CORNET		
AGI	VF	%ile	CC	AGI	r-value	p-value
AT1G55325	0.61	76.70	0.95	AT2G01210	0.75	2:2.94E-13
AT1G12930	0.56	73.00	0.92	AT1G64450	0.73	2:3.82E-11
AT2G25760	0.55	70.60	0.92	AT3G57830	0.7	2:2.46E-9
AT5G58100	0.54	69.50	0.92	AT3G57860	0.7	2:2.86E-9
AT2G35110	0.54	69.50	0.92	AT2G33560	0.7	2:2.86E-9
AT3G06340	0.54	69.50	0.92	AT3G54080	0.7	2:4.49E-9
AT1G27595	0.53	68.60	0.92	AT5G43020	0.69	2:8.11E-9
AT5G51340	0.53	68.60	0.93	AT5G67200	0.69	2:1.45E-8
AT5G38880	0.52	67.40	0.92	AT3G63290	0.69	2:2.23E-8
AT3G45190	0.52	67.40	0.92	AT5G26850	0.69	2:2.23E-8
AT5G15680	0.51	66.30	0.92	AT1G68640	0.69	2:2.58E-8
AT1G63700	0.51	66.30	0.92	AT5G67270	0.68	2:4.53E-8
AT3G43700	0.49	63.50	0.93	AT4G38660	0.68	2:4.53E-8
AT1G26170	0.49	63.50	0.93	AT5G10020	0.68	2:5.2E-8
AT1G04950	0.48	62.50	0.92	AT5G57590	0.68	2:7.88E-8
AT1G27850	0.47	61.20	0.91	AT3G61250	0.67	2:1.56E-7
AT1G34320	0.47	61.20	0.92	AT1G54180	0.67	2:1.78E-7
AT4G32620	0.47	61.20	0.92	AT5G63920	0.67	2:2.66E-7
AT5G27970	0.47	61.20	0.92	AT5G63960	0.67	2:3.47E-7
AT3G15120	0.46	59.8	0.92	AT5G63950	0.66	2:5.86E-7

Figure 32.b A complete list of the AT5G06680 query results as offered by each database.

The table shows also the statistical parameters exploited to describe the results.

AGI=Arabidopsis genome initiative code; C.V=Correlation value; CC=Cosine correlation value; MR=Mutual Ranking; r-value=Pearson coefficient; VF=Vertex F-measure; %ile=percentile; T=t-test value; DOF=degree of freedom; r2=Pearson coefficient squared; Spearman=Spearman coefficient; Weight=*weight of the relation*

CRESS RMA						CRESS GCRMA					
AGI	p-value	slope	T	DOF	r2	AGI	p-value	slope	T	DOF	r2
AT3G18524	2.15E-290	0.45	44.74	1717	0.538	AT4G11450	2.351E-261	0.88	44.91	1228	0.622
AT4G14970	2.74E-284	0.53	44.06	1717	0.531	AT2G35530	2.318E-252	0.91	43.69	1228	0.609
AT1G04050	2.99E-281	0.51	43.72	1717	0.527	AT3G21100	1.93E-233	0.84	41.16	1228	0.580
AT4G11450	3.75E-277	0.56	43.26	1717	0.522	AT1G55540	8.93E-218	0.99	39.09	1228	0.555
AT5G63960	3.93E-275	0.39	43.04	1717	0.519	AT1G14850	8.24E-209	1.00	37.90	1228	0.539
AT3G09730	1.25E-274	0.50	42.98	1717	0.518	AT3G23780	3.60E-205	0.83	37.42	1228	0.533
AT1G26370	5.56E-274	0.77	42.91	1717	0.518	AT5G12440	2.01E-203	0.69	37.19	1228	0.530
AT5G63950	7.80E-269	0.45	42.34	1717	0.511	AT3G20010	2.06E-200	0.68	36.80	1228	0.525
AT3G10390	2.76E-265	0.76	41.94	1717	0.506	AT2G23700	7.89E-197	0.76	36.33	1228	0.518
AT1G23380	4.33E-262	0.33	41.59	1717	0.502	AT3G19120	3.64E-196	0.82	36.24	1228	0.517
AT2G21800	1.63E-255	0.54	40.86	1717	0.493	AT5G40740	1.27E-193	0.82	35.91	1228	0.512
AT1G14850	3.99E-254	0.48	40.70	1717	0.491	AT1G73590	1.74E-192	0.49	35.76	1228	0.510
AT2G20300	4.34E-254	0.52	40.70	1717	0.491	AT2G39090	4.90E-192	0.83	35.70	1228	0.509
AT2G43990	1.28E-251	0.74	40.42	1717	0.488	AT1G06590	3.52E-190	0.75	35.45	1228	0.506
AT1G77720	7.86E-251	0.48	40.34	1717	0.487	AT3G63290	1.68E-189	1.03	35.36	1228	0.505
AT2G40070	1.26E-249	0.66	40.20	1717	0.485	AT5G63950	2.23E-187	0.67	35.09	1228	0.501
AT3G20020	2.51E-248	0.63	40.06	1717	0.483	AT1G48270	8.94E-187	1.24	35.01	1228	0.500
AT4G14290	6.21E-248	0.50	40.02	1717	0.483	AT2G40640	9.93E-185	0.83	34.74	1228	0.496
AT1G21740	2.68E-246	0.45	39.83	1717	0.480	AT2G47020	4.43E-184	1.01	34.65	1228	0.495
AT1G73590	4.63E-245	0.27	39.70	1717	0.479	AT2G25420	5.40E-183	0.78	34.51	1228	0.493

Figure 32.c A complete list of the AT5G06680 query results as offered by each database.

The table shows also the statistical parameters exploited to describe the results.

AGI=Arabidopsis genome initiative code; C.V=Correlation value; CC=Cosine correlation value; MR=Mutual Ranking; r-value=Pearson coefficient; VF=Vertex F-measure; %ile=percentile; T=t-test value; DOF=degree of freedom; r2=Pearson coefficient squared; Spearman=Spearman coefficient; Weight=*weight of the relation*

CRESS MAS						CSB	
AGI	p-value	slope	T	DOF	r2	AGI	spearman
AT1G55325	3.78E-141	0.68	28.03	1613	0.328	AT1G47670	0.647
AT4G33200	5.965E-141	0.61	28.00	1613	0.327	AT3G51050	0.628
AT5G10020	2.565E-138	0.43	27.68	1613	0.322	AT1G30970	0.622
AT2G40070	1.678E-137	0.71	27.58	1613	0.321	AT1G68550	0.617
AT3G61240	4.4E-136	0.55	27.41	1613	0.318	AT1G69295	0.612
AT1G73590	3.746E-129	0.29	26.55	1613	0.304	AT1G52150	0.611
AT5G13300	9.417E-129	0.49	26.50	1613	0.303	AT1G80530	0.606
AT3G12590	1.22E-128	0.70	26.48	1613	0.303	AT1G73590	0.581
AT5G65700	1.02E-126	0.46	26.24	1613	0.299	AT5G22740	0.581
AT3G58580	1.66E-117	0.71	25.08	1613	0.281	AT4G33210	0.577
AT5G23550	2.49E-115	0.70	24.80	1613	0.276	AT2G25970	0.574
AT1G65380	5.84E-115	0.59	24.76	1613	0.275	AT1G09960	0.570
AT4G31430	3.80E-114	0.42	24.65	1613	0.274	AT3G54080	0.570
AT1G53380	4.27E-114	0.51	24.65	1613	0.274	AT5G65700	0.570
AT2G38770	1.93E-113	0.50	24.56	1613	0.272	AT1G52310	0.565
AT2G47900	2.49E-113	0.56	24.55	1613	0.272	AT5G64390	0.565
AT3G19540	6.34E-110	0.52	24.11	1613	0.265	AT5G44670	0.565
AT1G07705	1.23E-108	0.74	23.94	1613	0.262	AT5G67630	0.565
AT5G22740	1.68E-107	0.34	23.79	1613	0.260	AT3G58040	0.563
AT1G79650	3.66E-106	0.50	23.62	1613	0.257	AT4G15900	0.563

Figure 32.d A complete list of the AT5G06680 query results as offered by each database.

The table shows also the statistical parameters exploited to describe the results.

AGI=Arabidopsis genome initiative code; C.V=Correlation value; CC=Cosine correlation value; MR=Mutual Ranking; r-value=Pearson coefficient; VF=Vertex F-measure; %ile=percentile; T=t-test value; DOF=degree of freedom; r2=Pearson coefficient squared; Spearman=Spearman coefficient; Weight=*weight of the relation*

GENE CAT		GENE MANIA		GENEVE STIGATOR	
AGI	r-value	AGI	weight	AGI	r-value
AT1G26170	NoT AvAilAbe	AT5G17410	2.752	AT3G22780	0.55
AT5G05560		AT3G61650	1.297	AT1G73590	0.55
AT3G21100		AT5G05620	1.255	AT2G02560	0.55
AT5G13300		AT5G37830	0.530	AT5G17410	0.55
AT3G16620		AT1G20570	0.400	AT1G14850	0.53
AT2G16880		AT1G80260	0.400	AT5G10020	0.53
AT3G20020		AT3G43610	0.400	AT5G60690	0.53
AT5G18960		AT3G11520	0.378	AT5G15680	0.52
AT3G20010		AT2G13650	0.120	AT1G55350	0.52
AT1G14850		AT2G22425	0.096	AT3G18524	0.52
AT3G10390		AT1G79280	0.092	AT4G36180	0.52
AT4G33200		AT4G40042	0.081	AT1G55325	0.52
AT3G15970		AT1G69295	0.081	AT2G05120	0.51
AT1G77720		AT3G22590	0.071	AT5G67100	0.51
AT1G72560		AT5G35430	0.052	AT2G27040	0.51
AT2G18850		AT5G14720	0.052	AT1G61010	0.51
AT1G47230		AT3G27325	0.048	AT3G63130	0.51
AT1G10490		AT1G77720	0.047	AT3G23780	0.51
AT1G65380		AT5G18960	0.040	AT2G28380	0.51
AT1G16190		AT3G53760	0.016	AT2G44830	0.51

Figure 32.e A complete list of the AT5G06680 query results as offered by each database. The table shows also the statistical parameters exploited to describe the results. AGI=Arabidopsis genome initiative code; C.V=Correlation value; CC=Cosine correlation value; MR=Mutual Ranking; r-value=Pearson coefficient; VF=Vertex F-measure; %ile=percentile; T=t-test value; DOF=degree of freedom; r2=Pearson coefficient squared; Spearman=Spearman coefficient; Weight=*weight of the relation*

Conclusion

The co-expression of genes is one of the hot topics in Biology. The need to identify the relationships between the genes is a mandatory step to understand the cell functions and its organization. In this work we have investigated tools that could help these analyses exploiting a reference genome of plant science *Arabidopsis thaliana*.

During our analyses however we have understood that this plant is still far to be considered a perfect genome model. Beyond the lacking of a complete exhaustive gene function annotation, the microarray probe distribution checked during our dataset building has underlined the lacking of probes for 20% of the coding genes. This issue goes worse since some probes are not specific too. Probably, the huge complexity of the Arabidopsis thaliana genome in terms of gene duplications is still representing an important obstacle to the synthesis of specific probes. The impossibility to essay the expression levels of a whole genome is to consider since many 'omics' analyses cannot find the proper validation and above all can lead to misleading conclusions.

About the co-expression, our results can confirm that there is an evident relationship between co-expressed genes and their function. Exploiting the Pearson's correlations to define the gene as co-expressed, the genes with a high number of correlations are involved in activities as the translation and transcription, while the gene with a low frequency of correlation are related to vegetative functions. This is reasonable, since a large number of elements involved in the same function offer the possibility to establish multiple checkpoints in important processes. Moreover, different subunits joining the same process offer to the cell the possibility to quickly change the result of an activity, by changing only one subunit instead of the whole mechanism. This approach let the cell to answer quickly to environment changes and stimuli. A clear example of this activity is also evident in the transcription in which the

RNA polymerase can transcript different genes according to the transcription factor bound. Never the less, we have seen that 87% of all the genes analyzed are doing at least one significant correlation: this underline the huge exploitation of interactions by the cell to organize its tasks, and the presence of numerous networks of expression.

Through our Pearson's tests, with a very high stringent threshold, we have been able also to identify the most probable co-expressed genes. In fact our approach has been able to identify three co-expressed groups each one involved in a very specific function: plant cell wall organization, transcription and energy production respectively. As well know in literature, all these functions need the interactions and coordination of several genes, since they are involving a large number of pathways. Our analysis has been hence able to identify the most probable genes involved in these functions.

The identification of expression networks has been studied since it represents an important bioinformatics challenge. The investigation on co-expressed genes has in fact underlined the request of an efficient tool when working on 'omics' data, as happened for our whole genome expression profiles analysis. The networks analysis through the clique approach has revealed to be a very deep essay, with the possibility to identify only the genes always in relationship, beyond the experiment considered. However the redundancy in the results due to the clique skill to consider group of genes as different, just if one gene collected is without a relationship with another gene sharing instead all the relationships with the other partners considered, provides often a result hard to manage, especially when working on huge data.

Moreover the great differentiating skill of clique approach has a very huge cost in terms of computational resources, which limit the exploitation of this method according to the size of its dataset and the computational resource available. The social network approach, based

on the graph theory, in our work has been proposed initially as the best candidate to substitute the clique approach. Its faster time of calculation and its low computational requirements have been well fitting properties to our analysis of co-expression. The results provided with this method were however not able to be specific as seen with the clique approach, although some group of collected genes have shown a very specific functional profile. Hence, we propose the social network approach as a valid solution for preliminary analyses or when there is the need of a fast answer. Then, if needed, after the identification of a specific, smaller, group of co-expressed genes, it is possible to perform deeper analyses exploiting the clique method.

The need to compare our networks analyses methods has lead to check the solutions proposed by the public available platforms. Using CESA7 as query, one of the genes experimentally confirmed as co-expressed in the plant cell wall composition, we have been able to establish the reliability of the platforms analyzed, although with not a completely overlapping in the results. However the exploitation of the gene AT5G06680 as query, which codes for tubulin and has a less specific expression profile in comparison to CESA7, has underlined the lacking of a uniform answer when searching for its co-expressed partners. This depicts that the investigation methods available are not working evenly, and the result proposed may change according to the kind of genes considered. In fact using CESA7 has brought to almost the same co-expressed group of gene on each platform, despite the exploitation of different datasets and statistical parameters, while the query for AT5G06680 has instead lead to co-expressed gene groups totally different. We hence propose that a correct data mining on gene co-expression analysis should include an appropriate comparison between the available resources, i.e. it is necessary to move from the idea of analyzing one single collection, to the one of comparing all or the majority of the available resources. So, considering the high number of

available platforms, databases diversification must be exploited as a global opportunity to ensure results robustness.

Appendix

GO terms for the six percentile in the chapter 3.2

GO term	Description	P-value	Enrichment (N, B, n, b)
GO:0006139	nucleobase-containing compound metabolic process	3.64E-175	3.64 (11998,906,1697,467)
GO:0034641	cellular nitrogen compound metabolic process	3.28E-165	3.39 (11998,1008,1697,484)
GO:0046483	heterocycle metabolic process	7.83E-164	3.35 (11998,1029,1697,488)
GO:0006807	nitrogen compound metabolic process	4.89E-161	3.17 (11998,1150,1697,516)
GO:0006725	cellular aromatic compound metabolic process	2.72E-150	3.09 (11998,1151,1697,503)
GO:1901360	organic cyclic compound metabolic process	1.87E-145	2.97 (11998,1225,1697,515)
GO:0044260	cellular macromolecule metabolic process	1.88E-145	2.56 (11998,1745,1697,633)
GO:0043170	macromolecule metabolic process	7.59E-139	2.43 (11998,1930,1697,662)
GO:0016070	RNA metabolic process	2.62E-136	4.00 (11998,585,1697,331)
GO:0090304	nucleic acid metabolic process	1.24E-135	3.56 (11998,756,1697,381)
GO:0044237	cellular metabolic process	1.82E-135	2.07 (11998,2825,1697,828)
GO:0044238	primary metabolic process	5.44E-132	2.12 (11998,2615,1697,784)
GO:0071704	organic substance metabolic process	8.49E-126	1.98 (11998,3003,1697,843)
GO:0009987	cellular process	9.88E-126	1.85 (11998,3613,1697,946)
GO:0008152	metabolic process	3.47E-121	1.91 (11998,3248,1697,877)
GO:1901576	organic substance biosynthetic process	5.44E-116	2.57 (11998,1439,1697,523)
GO:0009059	macromolecule biosynthetic process	1.45E-111	3.91 (11998,510,1697,282)
GO:0034645	cellular macromolecule biosynthetic process	2.04E-110	3.98 (11998,485,1697,273)
GO:0009451	RNA modification	3.15E-109	5.87 (11998,201,1697,167)
GO:0009058	biosynthetic process	3.75E-109	2.44 (11998,1553,1697,537)

GO:0044249	cellular biosynthetic process	2.19E-106	2.55 (11998,1365,1697,492)
GO:0071840	cellular component organization or biogenesis	1.17E-97	2.83 (11998,973,1697,389)
GO:0006412	translation	1.77E-96	5.63 (11998,196,1697,156)
GO:0032259	methylation	6.69E-95	5.23 (11998,227,1697,168)
GO:0043414	macromolecule methylation	6.69E-95	5.23 (11998,227,1697,168)
GO:0019637	organophosphate metabolic process	3.23E-87	3.89 (11998,407,1697,224)
GO:0043412	macromolecule modification	3.44E-85	2.97 (11998,760,1697,319)
GO:0001510	RNA methylation	1.16E-81	6.74 (11998,108,1697,103)
GO:0009117	nucleotide metabolic process	1.02E-79	5.07 (11998,205,1697,147)
GO:0006753	nucleoside phosphate metabolic process	3.09E-79	5.05 (11998,206,1697,147)
GO:0006796	phosphate-containing compound metabolic process	9.70E-78	3.09 (11998,632,1697,276)
GO:0090407	organophosphate biosynthetic process	2.37E-77	4.00 (11998,339,1697,192)
GO:0055086	nucleobase-containing small molecule metabolic process	2.85E-77	4.86 (11998,218,1697,150)
GO:0006793	phosphorus metabolic process	1.06E-76	3.03 (11998,653,1697,280)
GO:0016043	cellular component organization	3.64E-76	2.70 (11998,868,1697,332)
GO:0044267	cellular protein metabolic process	7.90E-73	2.63 (11998,899,1697,334)
GO:1901135	carbohydrate derivative metabolic process	1.71E-72	4.23 (11998,281,1697,168)
GO:0044271	cellular nitrogen compound biosynthetic process	3.03E-68	4.11 (11998,282,1697,164)
GO:0044763	single-organism cellular process	8.76E-67	1.86 (11998,2283,1697,599)
GO:0018130	heterocycle biosynthetic process	1.10E-66	4.04 (11998,287,1697,164)
GO:0034654	nucleobase-containing compound biosynthetic process	1.05E-63	4.62 (11998,202,1697,132)
GO:0019538	protein metabolic process	1.84E-63	2.27 (11998,1182,1697,380)
GO:0034660	ncRNA metabolic process	5.42E-63	5.67 (11998,126,1697,101)

GO:1901564	organonitrogen compound metabolic process	1.40E-62	3.11 (11998,505,1697,222)
GO:0006996	organelle organization	6.07E-61	3.07 (11998,506,1697,220)
GO:0019682	glyceraldehyde-3-phosphate metabolic process	8.14E-61	6.72 (11998,81,1697,77)
GO:0019288	isopentenyl diphosphate biosynthetic process, methylerythritol 4-phosphate pathway	8.14E-61	6.72 (11998,81,1697,77)
GO:0009240	isopentenyl diphosphate biosynthetic process	8.14E-61	6.72 (11998,81,1697,77)
GO:0046490	isopentenyl diphosphate metabolic process	8.14E-61	6.72 (11998,81,1697,77)
GO:0044281	small molecule metabolic process	2.84E-60	2.45 (11998,912,1697,316)
GO:0006090	pyruvate metabolic process	1.16E-59	6.64 (11998,82,1697,77)
GO:0044699	single-organism process	1.24E-58	1.60 (11998,3365,1697,763)
GO:0044711	single-organism biosynthetic process	7.61E-58	2.25 (11998,1123,1697,357)
GO:0019438	aromatic compound biosynthetic process	2.28E-56	3.29 (11998,395,1697,184)
GO:1901566	organonitrogen compound biosynthetic process	1.75E-55	3.55 (11998,325,1697,163)
GO:0044710	single-organism metabolic process	4.46E-55	1.86 (11998,1940,1697,510)
GO:0034470	ncRNA processing	1.07E-51	5.93 (11998,93,1697,78)
GO:1901362	organic cyclic compound biosynthetic process	1.72E-51	2.95 (11998,477,1697,199)
GO:0006364	rRNA processing	8.64E-51	6.27 (11998,80,1697,71)
GO:0016072	rRNA metabolic process	8.64E-51	6.27 (11998,80,1697,71)
GO:0009165	nucleotide biosynthetic process	1.27E-50	5.03 (11998,132,1697,94)
GO:1901293	nucleoside phosphate biosynthetic process	3.75E-50	5.00 (11998,133,1697,94)
GO:0010027	thylakoid membrane organization	8.75E-50	6.85 (11998,63,1697,61)
GO:0009668	plastid membrane organization	8.75E-50	6.85 (11998,63,1697,61)
GO:0006081	cellular aldehyde metabolic process	1.11E-48	5.05 (11998,126,1697,90)
GO:0044802	single-organism membrane organization	1.62E-48	6.74 (11998,64,1697,61)
GO:0061024	membrane organization	2.59E-46	6.53 (11998,66,1697,61)

GO:0006221	pyrimidine nucleotide biosynthetic process	3.92E-45	6.06 (11998,77,1697,66)
GO:0006220	pyrimidine nucleotide metabolic process	3.92E-45	6.06 (11998,77,1697,66)
GO:0009657	plastid organization	5.25E-45	5.66 (11998,90,1697,72)
GO:0072528	pyrimidine-containing compound biosynthetic process	2.20E-44	5.98 (11998,78,1697,66)
GO:0009220	pyrimidine ribonucleotide biosynthetic process	2.46E-44	6.05 (11998,76,1697,65)
GO:0009218	pyrimidine ribonucleotide metabolic process	2.46E-44	6.05 (11998,76,1697,65)
GO:0022613	ribonucleoprotein complex biogenesis	2.60E-44	6.11 (11998,74,1697,64)
GO:0042254	ribosome biogenesis	1.06E-42	6.16 (11998,70,1697,61)
GO:0046390	ribose phosphate biosynthetic process	1.19E-41	5.69 (11998,82,1697,66)
GO:0009260	ribonucleotide biosynthetic process	1.19E-41	5.69 (11998,82,1697,66)
GO:0006396	RNA processing	1.95E-40	3.25 (11998,294,1697,135)
GO:0009259	ribonucleotide metabolic process	2.03E-40	5.56 (11998,84,1697,66)
GO:0019693	ribose phosphate metabolic process	2.03E-40	5.56 (11998,84,1697,66)
GO:0072527	pyrimidine-containing compound metabolic process	7.84E-40	5.49 (11998,85,1697,66)
GO:1902589	single-organism organelle organization	8.82E-39	3.14 (11998,306,1697,136)
GO:0005982	starch metabolic process	1.05E-38	5.36 (11998,87,1697,66)
GO:0019252	starch biosynthetic process	5.59E-38	5.55 (11998,79,1697,62)
GO:0006325	chromatin organization	8.84E-37	3.88 (11998,173,1697,95)
GO:0016570	histone modification	2.58E-36	4.29 (11998,135,1697,82)
GO:0034968	histone lysine methylation	3.79E-36	5.07 (11998,92,1697,66)
GO:0022607	cellular component assembly	1.80E-35	3.75 (11998,181,1697,96)
GO:0000023	maltose metabolic process	2.22E-35	5.60 (11998,72,1697,57)
GO:0016568	chromatin modification	5.08E-35	4.03 (11998,151,1697,86)
GO:0072594	establishment of protein localization to organelle	7.35E-35	3.37 (11998,231,1697,110)
GO:0016556	mRNA modification	1.81E-34	5.64 (11998,69,1697,55)
GO:0016569	covalent chromatin modification	2.13E-34	4.01 (11998,150,1697,85)

GO:1901137	carbohydrate derivative biosynthetic process	8.00E-34	3.76 (11998,171,1697,91)
GO:0006479	protein methylation	5.90E-33	4.44 (11998,113,1697,71)
GO:0008213	protein alkylation	5.90E-33	4.44 (11998,113,1697,71)
GO:0016571	histone methylation	2.71E-32	4.42 (11998,112,1697,70)
GO:0072524	pyridine-containing compound metabolic process	1.99E-31	5.55 (11998,65,1697,51)
GO:0009658	chloroplast organization	4.94E-31	5.41 (11998,68,1697,52)
GO:0006739	NADP metabolic process	1.13E-30	5.52 (11998,64,1697,50)
GO:0006740	NADPH regeneration	1.13E-30	5.52 (11998,64,1697,50)
GO:0006098	pentose-phosphate shunt	1.13E-30	5.52 (11998,64,1697,50)
GO:0019362	pyridine nucleotide metabolic process	1.13E-30	5.52 (11998,64,1697,50)
GO:0046496	nicotinamide nucleotide metabolic process	1.13E-30	5.52 (11998,64,1697,50)
GO:0010207	photosystem II assembly	6.50E-30	6.30 (11998,46,1697,41)
GO:0016482	cytoplasmic transport	6.76E-30	2.94 (11998,279,1697,116)
GO:0009250	glucan biosynthetic process	2.31E-29	4.34 (11998,106,1697,65)
GO:0006644	phospholipid metabolic process	4.43E-29	3.54 (11998,170,1697,85)
GO:0006733	oxidoreduction coenzyme metabolic process	5.61E-29	4.96 (11998,77,1697,54)
GO:0046148	pigment biosynthetic process	5.86E-29	5.11 (11998,72,1697,52)
GO:0005984	disaccharide metabolic process	6.63E-29	4.74 (11998,85,1697,57)
GO:0006073	cellular glucan metabolic process	9.09E-29	3.91 (11998,132,1697,73)
GO:0044042	glucan metabolic process	9.09E-29	3.91 (11998,132,1697,73)
GO:0008654	phospholipid biosynthetic process	9.45E-29	3.54 (11998,168,1697,84)
GO:0051276	chromosome organization	3.52E-28	3.05 (11998,239,1697,103)
GO:0044085	cellular component biogenesis	7.00E-28	3.86 (11998,132,1697,72)
GO:0009311	oligosaccharide metabolic process	1.28E-27	4.51 (11998,91,1697,58)
GO:0044767	single-organism developmental process	4.33E-27	1.93 (11998,898,1697,245)
GO:0009909	regulation of flower development	1.17E-26	4.50 (11998,88,1697,56)

GO:0032502	developmental process	6.15E-26	1.86 (11998,986,1697,259)
GO:0051186	cofactor metabolic process	7.71E-26	3.39 (11998,167,1697,80)
GO:0016226	iron-sulfur cluster assembly	1.02E-25	6.85 (11998,32,1697,31)
GO:0031163	metallo-sulfur cluster assembly	1.02E-25	6.85 (11998,32,1697,31)
GO:0048831	regulation of shoot system development	1.53E-25	4.29 (11998,94,1697,57)
GO:2000026	regulation of multicellular organismal development	4.45E-25	3.34 (11998,167,1697,79)
GO:0051649	establishment of localization in cell	6.82E-25	2.20 (11998,534,1697,166)
GO:2000241	regulation of reproductive process	1.45E-24	4.21 (11998,94,1697,56)
GO:0051239	regulation of multicellular organismal process	4.91E-24	3.25 (11998,172,1697,79)
GO:0048580	regulation of post-embryonic development	6.33E-24	4.12 (11998,96,1697,56)
GO:0016109	tetraterpenoid biosynthetic process	9.72E-24	7.07 (11998,27,1697,27)
GO:0016108	tetraterpenoid metabolic process	9.72E-24	7.07 (11998,27,1697,27)
GO:0016116	carotenoid metabolic process	9.72E-24	7.07 (11998,27,1697,27)
GO:0016117	carotenoid biosynthetic process	9.72E-24	7.07 (11998,27,1697,27)
GO:0006732	coenzyme metabolic process	3.23E-23	3.78 (11998,114,1697,61)
GO:0050793	regulation of developmental process	3.69E-23	3.05 (11998,195,1697,84)
GO:0042440	pigment metabolic process	1.02E-22	3.96 (11998,100,1697,56)
GO:0006006	glucose metabolic process	2.37E-22	3.41 (11998,141,1697,68)
GO:0019752	carboxylic acid metabolic process	2.38E-22	2.07 (11998,585,1697,171)
GO:0043436	oxoacid metabolic process	3.56E-22	2.06 (11998,587,1697,171)
GO:0006082	organic acid metabolic process	3.56E-22	2.06 (11998,587,1697,171)
GO:0019318	hexose metabolic process	7.67E-22	3.26 (11998,154,1697,71)
GO:0016071	mRNA metabolic process	8.11E-22	3.17 (11998,165,1697,74)
GO:0006606	protein import into nucleus	1.45E-21	4.48 (11998,71,1697,45)
GO:1902593	single-organism nuclear import	1.45E-21	4.48 (11998,71,1697,45)

GO:0051170	nuclear import	1.45E-21	4.48 (11998,71,1697,45)
GO:0019684	photosynthesis, light reaction	1.63E-21	6.81 (11998,27,1697,26)
GO:0006913	nucleocytoplasmic transport	2.03E-21	4.05 (11998,89,1697,51)
GO:0051169	nuclear transport	2.03E-21	4.05 (11998,89,1697,51)
GO:0030154	cell differentiation	2.29E-21	3.18 (11998,160,1697,72)
GO:0006886	intracellular protein transport	2.42E-21	2.35 (11998,370,1697,123)
GO:0015031	protein transport	3.54E-21	2.31 (11998,385,1697,126)
GO:0045184	establishment of protein localization	3.54E-21	2.31 (11998,385,1697,126)
GO:0043085	positive regulation of catalytic activity	3.72E-21	5.38 (11998,46,1697,35)
GO:0044093	positive regulation of molecular function	1.26E-20	5.26 (11998,47,1697,35)
GO:0006464	cellular protein modification process	1.91E-20	2.04 (11998,558,1697,161)
GO:0036211	protein modification process	1.91E-20	2.04 (11998,558,1697,161)
GO:0005996	monosaccharide metabolic process	2.28E-20	3.05 (11998,169,1697,73)
GO:0009965	leaf morphogenesis	2.55E-20	6.03 (11998,34,1697,29)
GO:0006007	glucose catabolic process	4.96E-20	3.79 (11998,97,1697,52)
GO:0051567	histone H3-K9 methylation	6.02E-20	4.43 (11998,67,1697,42)
GO:0008283	cell proliferation	7.09E-20	5.62 (11998,39,1697,31)
GO:0009902	chloroplast relocation	1.31E-19	6.55 (11998,27,1697,25)
GO:0051656	establishment of organelle localization	1.31E-19	6.55 (11998,27,1697,25)
GO:0051667	establishment of plastid localization	1.31E-19	6.55 (11998,27,1697,25)
GO:0019320	hexose catabolic process	1.67E-19	3.71 (11998,99,1697,52)
GO:0046365	monosaccharide catabolic process	1.67E-19	3.71 (11998,99,1697,52)
GO:0032506	cytokinetic process	2.36E-19	4.49 (11998,63,1697,40)
GO:0000911	cytokinesis by cell plate formation	2.36E-19	4.49 (11998,63,1697,40)
GO:1902410	mitotic cytokinetic process	2.36E-19	4.49 (11998,63,1697,40)
GO:0065003	macromolecular complex assembly	1.33E-18	3.22 (11998,134,1697,61)
GO:0006626	protein targeting to mitochondrion	2.14E-18	4.48 (11998,60,1697,38)

GO:0072655	establishment of protein localization to mitochondrion	2.14E-18	4.48 (11998,60,1697,38)
GO:0032774	RNA biosynthetic process	5.89E-18	4.14 (11998,70,1697,41)
GO:0016052	carbohydrate catabolic process	1.19E-17	2.98 (11998,154,1697,65)
GO:1902582	single-organism intracellular transport	1.44E-17	2.04 (11998,479,1697,138)
GO:0006461	protein complex assembly	1.44E-17	3.16 (11998,132,1697,59)
GO:0006399	tRNA metabolic process	2.34E-17	4.76 (11998,49,1697,33)
GO:0015995	chlorophyll biosynthetic process	3.78E-17	6.48 (11998,24,1697,22)
GO:0046907	intracellular transport	4.77E-17	1.99 (11998,505,1697,142)
GO:0007010	cytoskeleton organization	9.56E-17	3.24 (11998,118,1697,54)
GO:0000226	microtubule cytoskeleton organization	1.38E-16	3.75 (11998,81,1697,43)
GO:0006779	porphyrin-containing compound biosynthetic process	2.50E-16	5.30 (11998,36,1697,27)
GO:0033014	tetrapyrrole biosynthetic process	2.50E-16	5.30 (11998,36,1697,27)
GO:0042793	transcription from plastid promoter	4.75E-16	7.07 (11998,18,1697,18)
GO:0008610	lipid biosynthetic process	4.75E-16	2.14 (11998,366,1697,111)
GO:0034622	cellular macromolecular complex assembly	6.54E-16	3.09 (11998,126,1697,55)
GO:0007017	microtubule-based process	9.26E-16	3.50 (11998,91,1697,45)
GO:0044724	single-organism carbohydrate catabolic process	9.69E-16	2.91 (11998,146,1697,60)
GO:0009073	aromatic amino acid family biosynthetic process	1.22E-15	5.81 (11998,28,1697,23)
GO:0043623	cellular protein complex assembly	1.85E-15	3.10 (11998,121,1697,53)
GO:0065009	regulation of molecular function	2.42E-15	3.54 (11998,86,1697,43)
GO:0008652	cellular amino acid biosynthetic process	2.98E-15	2.85 (11998,149,1697,60)
GO:0017038	protein import	6.12E-15	3.18 (11998,109,1697,49)
GO:0044723	single-organism carbohydrate metabolic process	8.31E-15	1.90 (11998,514,1697,138)
GO:0006275	regulation of DNA replication	1.19E-14	5.30 (11998,32,1697,24)

GO:0050790	regulation of catalytic activity	1.21E-14	3.79 (11998,69,1697,37)
GO:0048645	organ formation	1.36E-14	3.86 (11998,66,1697,36)
GO:0044264	cellular polysaccharide metabolic process	1.49E-14	2.42 (11998,225,1697,77)
GO:0010103	stomatal complex morphogenesis	1.60E-14	4.01 (11998,60,1697,34)
GO:0043933	macromolecular complex subunit organization	1.67E-14	2.68 (11998,166,1697,63)
GO:0006839	mitochondrial transport	1.85E-14	3.49 (11998,83,1697,41)
GO:0008150	biological_process	1.87E-14	1.08 (11998,9556,1697,1464)
GO:0016051	carbohydrate biosynthetic process	2.79E-14	2.22 (11998,286,1697,90)
GO:0048449	floral organ formation	2.85E-14	3.87 (11998,64,1697,35)
GO:0015979	photosynthesis	4.19E-14	5.71 (11998,26,1697,21)
GO:0006520	cellular amino acid metabolic process	4.50E-14	2.35 (11998,238,1697,79)
GO:0032787	monocarboxylic acid metabolic process	1.11E-13	1.98 (11998,403,1697,113)
GO:0006354	DNA-templated transcription, elongation	1.11E-13	4.99 (11998,34,1697,24)
GO:0071822	protein complex subunit organization	1.26E-13	2.63 (11998,164,1697,61)
GO:0040029	regulation of gene expression, epigenetic	1.27E-13	2.95 (11998,120,1697,50)
GO:0005976	polysaccharide metabolic process	1.59E-13	2.30 (11998,243,1697,79)
GO:0033692	cellular polysaccharide biosynthetic process	1.72E-13	2.45 (11998,199,1697,69)
GO:0045036	protein targeting to chloroplast	3.39E-13	5.18 (11998,30,1697,22)
GO:0072596	establishment of protein localization to chloroplast	3.39E-13	5.18 (11998,30,1697,22)
GO:0044255	cellular lipid metabolic process	3.91E-13	1.94 (11998,420,1697,115)
GO:0000271	polysaccharide biosynthetic process	4.15E-13	2.38 (11998,211,1697,71)
GO:0055114	oxidation-reduction process	5.94E-13	2.55 (11998,169,1697,61)
GO:0051726	regulation of cell cycle	9.20E-13	3.77 (11998,60,1697,32)
GO:0010557	positive regulation of macromolecule biosynthetic process	1.63E-12	3.09 (11998,96,1697,42)
GO:0009560	embryo sac egg cell differentiation	1.65E-12	3.49 (11998,71,1697,35)
GO:0051188	cofactor biosynthetic process	1.67E-12	3.71 (11998,61,1697,32)

GO:0016458	gene silencing	2.13E-12	2.95 (11998,108,1697,45)
GO:0005975	carbohydrate metabolic process	2.38E-12	1.73 (11998,609,1697,149)
GO:0006270	DNA replication initiation	2.78E-12	6.28 (11998,18,1697,16)
GO:0034637	cellular carbohydrate biosynthetic process	3.19E-12	2.32 (11998,210,1697,69)
GO:0044262	cellular carbohydrate metabolic process	3.77E-12	2.07 (11998,304,1697,89)
GO:0031328	positive regulation of cellular biosynthetic process	3.77E-12	3.03 (11998,98,1697,42)
GO:0006629	lipid metabolic process	4.73E-12	1.81 (11998,497,1697,127)
GO:0000338	protein deneddylation	5.14E-12	4.31 (11998,41,1697,25)
GO:0010388	cullin deneddylation	5.14E-12	4.31 (11998,41,1697,25)
GO:0065007	biological regulation	7.14E-12	1.40 (11998,1612,1697,320)
GO:0050789	regulation of biological process	9.85E-12	1.42 (11998,1470,1697,296)
GO:0010604	positive regulation of macromolecule metabolic process	1.25E-11	2.94 (11998,101,1697,42)
GO:0031325	positive regulation of cellular metabolic process	1.25E-11	2.94 (11998,101,1697,42)
GO:0010628	positive regulation of gene expression	1.46E-11	3.06 (11998,90,1697,39)
GO:1902680	positive regulation of RNA biosynthetic process	1.46E-11	3.06 (11998,90,1697,39)
GO:0045893	positive regulation of transcription, DNA-templated	1.46E-11	3.06 (11998,90,1697,39)
GO:0051254	positive regulation of RNA metabolic process	1.46E-11	3.06 (11998,90,1697,39)
GO:0006790	sulfur compound metabolic process	1.92E-11	2.53 (11998,151,1697,54)
GO:0045935	positive regulation of nucleobase-containing compound metabolic process	2.22E-11	3.03 (11998,91,1697,39)
GO:0051173	positive regulation of nitrogen compound metabolic process	2.22E-11	3.03 (11998,91,1697,39)
GO:0006260	DNA replication	2.27E-11	4.11 (11998,43,1697,25)
GO:0009887	organ morphogenesis	2.57E-11	2.93 (11998,99,1697,41)
GO:0006091	generation of precursor metabolites and energy	2.77E-11	2.84 (11998,107,1697,43)

GO:0044272	sulfur compound biosynthetic process	4.31E-11	2.56 (11998,141,1697,51)
GO:0003006	developmental process involved in reproduction	5.26E-11	1.74 (11998,519,1697,128)
GO:0022402	cell cycle process	7.32E-11	2.43 (11998,160,1697,55)
GO:0006306	DNA methylation	8.01E-11	3.83 (11998,48,1697,26)
GO:0006305	DNA alkylation	8.01E-11	3.83 (11998,48,1697,26)
GO:0044728	DNA methylation or demethylation	8.01E-11	3.83 (11998,48,1697,26)
GO:0035304	regulation of protein dephosphorylation	8.43E-11	5.09 (11998,25,1697,18)
GO:0022414	reproductive process	8.95E-11	1.72 (11998,539,1697,131)
GO:0006778	porphyrin-containing compound metabolic process	9.12E-11	3.37 (11998,65,1697,31)
GO:0033013	tetrapyrrole metabolic process	9.12E-11	3.37 (11998,65,1697,31)
GO:0006605	protein targeting	1.23E-10	2.02 (11998,283,1697,81)
GO:0006304	DNA modification	1.48E-10	3.75 (11998,49,1697,26)
GO:0009886	post-embryonic morphogenesis	1.56E-10	2.68 (11998,116,1697,44)
GO:0022412	cellular process involved in reproduction in multicellular organism	1.79E-10	3.05 (11998,81,1697,35)
GO:0009653	anatomical structure morphogenesis	2.28E-10	1.96 (11998,311,1697,86)
GO:0035303	regulation of dephosphorylation	2.38E-10	4.89 (11998,26,1697,18)
GO:0051234	establishment of localization	3.05E-10	1.45 (11998,1168,1697,239)
GO:0051052	regulation of DNA metabolic process	4.06E-10	3.36 (11998,61,1697,29)
GO:0044702	single organism reproductive process	5.70E-10	1.77 (11998,435,1697,109)
GO:0006342	chromatin silencing	6.60E-10	2.80 (11998,96,1697,38)
GO:0045814	negative regulation of gene expression, epigenetic	6.60E-10	2.80 (11998,96,1697,38)
GO:0019220	regulation of phosphate metabolic process	8.34E-10	4.28 (11998,33,1697,20)
GO:0051174	regulation of phosphorus metabolic process	8.34E-10	4.28 (11998,33,1697,20)
GO:0009639	response to red or far red light	1.25E-09	3.17 (11998,67,1697,30)
GO:0048646	anatomical structure formation involved in	1.33E-09	2.74 (11998,98,1697,38)

	morphogenesis		
GO:0015994	chlorophyll metabolic process	1.39E-09	3.47 (11998,53,1697,26)
GO:0048453	sepal formation	1.39E-09	3.78 (11998,43,1697,23)
GO:0048451	petal formation	1.39E-09	3.78 (11998,43,1697,23)
GO:0048522	positive regulation of cellular process	1.41E-09	2.33 (11998,158,1697,52)
GO:0031399	regulation of protein modification process	1.76E-09	4.16 (11998,34,1697,20)
GO:0006655	phosphatidylglycerol biosynthetic process	2.14E-09	5.50 (11998,18,1697,14)
GO:0046471	phosphatidylglycerol metabolic process	2.14E-09	5.50 (11998,18,1697,14)
GO:0006259	DNA metabolic process	2.41E-09	2.15 (11998,197,1697,60)
GO:0006346	methylation-dependent chromatin silencing	2.94E-09	3.47 (11998,51,1697,25)
GO:0009640	photomorphogenesis	2.94E-09	3.47 (11998,51,1697,25)
GO:0006414	translational elongation	2.95E-09	4.62 (11998,26,1697,17)
GO:0019222	regulation of metabolic process	3.19E-09	1.51 (11998,823,1697,176)
GO:0010629	negative regulation of gene expression	4.70E-09	2.39 (11998,136,1697,46)
GO:0010605	negative regulation of macromolecule metabolic process	4.87E-09	2.34 (11998,145,1697,48)
GO:0072521	purine-containing compound metabolic process	6.92E-09	4.45 (11998,27,1697,17)
GO:0006164	purine nucleotide biosynthetic process	7.06E-09	5.21 (11998,19,1697,14)
GO:0072522	purine-containing compound biosynthetic process	7.06E-09	5.21 (11998,19,1697,14)
GO:0032268	regulation of cellular protein metabolic process	8.04E-09	3.33 (11998,53,1697,25)
GO:0009892	negative regulation of metabolic process	1.03E-08	2.29 (11998,148,1697,48)
GO:0048869	cellular developmental process	1.11E-08	1.85 (11998,314,1697,82)
GO:0010389	regulation of G2/M transition of mitotic cell cycle	1.19E-08	5.41 (11998,17,1697,13)
GO:1902749	regulation of cell cycle G2/M phase transition	1.19E-08	5.41 (11998,17,1697,13)
GO:0010564	regulation of cell cycle process	1.38E-08	4.52 (11998,25,1697,16)

GO:0016114	terpenoid biosynthetic process	1.39E-08	2.97 (11998,69,1697,29)
GO:0016572	histone phosphorylation	1.86E-08	5.66 (11998,15,1697,12)
GO:0031047	gene silencing by RNA	2.17E-08	2.71 (11998,86,1697,33)
GO:0009791	post-embryonic development	2.28E-08	2.51 (11998,107,1697,38)
GO:1901659	glycosyl compound biosynthetic process	2.79E-08	3.46 (11998,45,1697,22)
GO:0006721	terpenoid metabolic process	3.02E-08	2.89 (11998,71,1697,29)
GO:1901990	regulation of mitotic cell cycle phase transition	3.72E-08	5.11 (11998,18,1697,13)
GO:1901987	regulation of cell cycle phase transition	3.72E-08	5.11 (11998,18,1697,13)
GO:0006351	transcription, DNA-templated	4.27E-08	3.63 (11998,39,1697,20)
GO:0044283	small molecule biosynthetic process	4.52E-08	1.67 (11998,441,1697,104)
GO:0044707	single-multicellular organism process	4.88E-08	1.78 (11998,329,1697,83)
GO:0006163	purine nucleotide metabolic process	5.34E-08	4.71 (11998,21,1697,14)
GO:0009072	aromatic amino acid family metabolic process	6.29E-08	2.81 (11998,73,1697,29)
GO:0016143	S-glycoside metabolic process	7.56E-08	3.31 (11998,47,1697,22)
GO:0019760	glucosinolate metabolic process	7.56E-08	3.31 (11998,47,1697,22)
GO:0019757	glycosinolate metabolic process	7.56E-08	3.31 (11998,47,1697,22)
GO:0016144	S-glycoside biosynthetic process	1.25E-07	3.45 (11998,41,1697,20)
GO:0019758	glycosinolate biosynthetic process	1.25E-07	3.45 (11998,41,1697,20)
GO:0019761	glucosinolate biosynthetic process	1.25E-07	3.45 (11998,41,1697,20)
GO:1901657	glycosyl compound metabolic process	1.26E-07	2.85 (11998,67,1697,27)
GO:0032501	multicellular organismal process	1.44E-07	1.74 (11998,342,1697,84)
GO:0000469	cleavage involved in rRNA processing	1.58E-07	7.07 (11998,8,1697,8)
GO:0090501	RNA phosphodiester bond hydrolysis	1.58E-07	7.07 (11998,8,1697,8)
GO:0009767	photosynthetic electron transport chain	1.58E-07	7.07 (11998,8,1697,8)
GO:0000741	karyogamy	1.92E-07	4.99 (11998,17,1697,12)
GO:0048284	organelle fusion	1.92E-07	4.99 (11998,17,1697,12)
GO:1901575	organic substance catabolic process	2.40E-07	1.56 (11998,545,1697,120)

GO:0009637	response to blue light	3.47E-07	5.18 (11998,15,1697,11)
GO:0051246	regulation of protein metabolic process	3.61E-07	2.85 (11998,62,1697,25)
GO:0044765	single-organism transport	4.47E-07	1.38 (11998,1037,1697,202)
GO:0000096	sulfur amino acid metabolic process	4.78E-07	2.36 (11998,105,1697,35)
GO:0009790	embryo development	5.07E-07	2.69 (11998,71,1697,27)
GO:0009793	embryo development ending in seed dormancy	5.09E-07	2.88 (11998,59,1697,24)
GO:0006333	chromatin assembly or disassembly	5.92E-07	4.12 (11998,24,1697,14)
GO:0010558	negative regulation of macromolecule biosynthetic process	5.97E-07	2.22 (11998,124,1697,39)
GO:1902679	negative regulation of RNA biosynthetic process	5.97E-07	2.22 (11998,124,1697,39)
GO:0045892	negative regulation of transcription, DNA-templated	5.97E-07	2.22 (11998,124,1697,39)
GO:2000113	negative regulation of cellular macromolecule biosynthetic process	5.97E-07	2.22 (11998,124,1697,39)
GO:0051253	negative regulation of RNA metabolic process	5.97E-07	2.22 (11998,124,1697,39)
GO:0045934	negative regulation of nucleobase-containing compound metabolic process	7.51E-07	2.21 (11998,125,1697,39)
GO:0044712	single-organism catabolic process	8.18E-07	1.61 (11998,427,1697,97)
GO:1901607	alpha-amino acid biosynthetic process	9.93E-07	2.24 (11998,117,1697,37)
GO:0031324	negative regulation of cellular metabolic process	1.08E-06	2.13 (11998,136,1697,41)
GO:0000478	endonucleolytic cleavage involved in rRNA processing	1.12E-06	7.07 (11998,7,1697,7)
GO:0090502	RNA phosphodiester bond hydrolysis, endonucleolytic	1.12E-06	7.07 (11998,7,1697,7)
GO:0009773	photosynthetic electron transport in photosystem I	1.12E-06	7.07 (11998,7,1697,7)
GO:0006405	RNA export from nucleus	1.17E-06	3.96 (11998,25,1697,14)
GO:0050658	RNA transport	1.17E-06	3.96 (11998,25,1697,14)
GO:0050657	nucleic acid transport	1.17E-06	3.96 (11998,25,1697,14)
GO:0051236	establishment of RNA localization	1.17E-06	3.96 (11998,25,1697,14)

GO:0051172	negative regulation of nitrogen compound metabolic process	1.18E-06	2.17 (11998,127,1697,39)
GO:0009890	negative regulation of biosynthetic process	1.18E-06	2.17 (11998,127,1697,39)
GO:0031327	negative regulation of cellular biosynthetic process	1.18E-06	2.17 (11998,127,1697,39)
GO:0009891	positive regulation of biosynthetic process	1.20E-06	2.08 (11998,146,1697,43)
GO:0031048	chromatin silencing by small RNA	1.49E-06	2.74 (11998,62,1697,24)
GO:0060255	regulation of macromolecule metabolic process	1.72E-06	1.45 (11998,682,1697,140)
GO:0006810	transport	1.75E-06	1.34 (11998,1143,1697,216)
GO:0070646	protein modification by small protein removal	1.79E-06	2.55 (11998,75,1697,27)
GO:0051049	regulation of transport	1.98E-06	3.34 (11998,36,1697,17)
GO:0008299	isoprenoid biosynthetic process	2.01E-06	2.44 (11998,84,1697,29)
GO:0071555	cell wall organization	2.10E-06	2.06 (11998,144,1697,42)
GO:0000097	sulfur amino acid biosynthetic process	2.15E-06	2.29 (11998,102,1697,33)
GO:0016053	organic acid biosynthetic process	2.20E-06	1.69 (11998,309,1697,74)
GO:0046394	carboxylic acid biosynthetic process	2.20E-06	1.69 (11998,309,1697,74)
GO:0051168	nuclear export	2.20E-06	3.81 (11998,26,1697,14)
GO:0048519	negative regulation of biological process	2.26E-06	1.74 (11998,277,1697,68)
GO:0043269	regulation of ion transport	2.56E-06	3.43 (11998,33,1697,16)
GO:0009893	positive regulation of metabolic process	3.19E-06	2.01 (11998,151,1697,43)
GO:0007346	regulation of mitotic cell cycle	3.24E-06	3.54 (11998,30,1697,15)
GO:0006720	isoprenoid metabolic process	3.46E-06	2.38 (11998,86,1697,29)
GO:0048518	positive regulation of biological process	4.05E-06	1.82 (11998,218,1697,56)
GO:0016310	phosphorylation	5.29E-06	1.79 (11998,225,1697,57)
GO:0010155	regulation of proton transport	5.35E-06	4.32 (11998,18,1697,11)
GO:0009056	catabolic process	6.07E-06	1.44 (11998,638,1697,130)

GO:0070647	protein modification by small protein conjugation or removal	6.35E-06	2.22 (11998,102,1697,32)
GO:0009664	plant-type cell wall organization	7.03E-06	2.08 (11998,126,1697,37)
GO:0032879	regulation of localization	8.36E-06	2.96 (11998,43,1697,18)
GO:0071702	organic substance transport	8.64E-06	1.40 (11998,726,1697,144)
GO:0006406	mRNA export from nucleus	1.00E-05	3.86 (11998,22,1697,12)
GO:0006418	tRNA aminoacylation for protein translation	1.00E-05	3.86 (11998,22,1697,12)
GO:0043038	amino acid activation	1.00E-05	3.86 (11998,22,1697,12)
GO:0043039	tRNA aminoacylation	1.00E-05	3.86 (11998,22,1697,12)
GO:0051028	mRNA transport	1.00E-05	3.86 (11998,22,1697,12)
GO:0010075	regulation of meristem growth	1.06E-05	2.54 (11998,64,1697,23)
GO:0008380	RNA splicing	1.39E-05	2.09 (11998,115,1697,34)
GO:0000413	protein peptidyl-prolyl isomerization	1.75E-05	5.14 (11998,11,1697,8)
GO:0048481	ovule development	2.15E-05	3.89 (11998,20,1697,11)
GO:0018193	peptidyl-amino acid modification	2.49E-05	3.40 (11998,27,1697,13)
GO:0048509	regulation of meristem development	2.51E-05	2.43 (11998,67,1697,23)
GO:0048878	chemical homeostasis	2.54E-05	2.77 (11998,46,1697,18)
GO:0007051	spindle organization	3.10E-05	5.50 (11998,9,1697,7)
GO:0051225	spindle assembly	3.10E-05	5.50 (11998,9,1697,7)
GO:0070925	organelle assembly	3.10E-05	5.50 (11998,9,1697,7)
GO:0006468	protein phosphorylation	3.29E-05	1.81 (11998,180,1697,46)
GO:0045229	external encapsulating structure organization	3.56E-05	1.84 (11998,165,1697,43)
GO:0000377	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	3.57E-05	2.90 (11998,39,1697,16)
GO:0000375	RNA splicing, via transesterification reactions	3.57E-05	2.90 (11998,39,1697,16)
GO:1901605	alpha-amino acid metabolic process	4.16E-05	1.83 (11998,166,1697,43)
GO:0006636	unsaturated fatty acid biosynthetic process	4.89E-05	6.06 (11998,7,1697,6)
GO:0033559	unsaturated fatty acid metabolic process	4.89E-05	6.06 (11998,7,1697,6)

GO:0000398	mRNA splicing, via spliceosome	5.04E-05	2.95 (11998,36,1697,15)
GO:0050801	ion homeostasis	5.14E-05	2.73 (11998,44,1697,17)
GO:0042274	ribosomal small subunit biogenesis	5.63E-05	7.07 (11998,5,1697,5)
GO:0042545	cell wall modification	6.38E-05	1.98 (11998,118,1697,33)
GO:0042991	transcription factor import into nucleus	6.40E-05	3.17 (11998,29,1697,13)
GO:0071669	plant-type cell wall organization or biogenesis	7.29E-05	1.85 (11998,149,1697,39)
GO:0015931	nucleobase-containing compound transport	7.37E-05	2.87 (11998,37,1697,15)
GO:0048638	regulation of developmental growth	7.86E-05	2.12 (11998,90,1697,27)
GO:0006261	DNA-dependent DNA replication	8.49E-05	3.72 (11998,19,1697,10)
GO:0006312	mitotic recombination	8.49E-05	3.72 (11998,19,1697,10)
GO:0048856	anatomical structure development	9.75E-05	1.55 (11998,320,1697,70)
GO:0055080	cation homeostasis	1.06E-04	2.79 (11998,38,1697,15)
GO:0048523	negative regulation of cellular process	1.35E-04	1.71 (11998,190,1697,46)
GO:0006094	gluconeogenesis	1.35E-04	2.56 (11998,47,1697,17)
GO:0019319	hexose biosynthetic process	1.35E-04	2.56 (11998,47,1697,17)
GO:0010073	meristem maintenance	1.48E-04	3.54 (11998,20,1697,10)
GO:0007389	pattern specification process	1.71E-04	2.07 (11998,89,1697,26)
GO:0040008	regulation of growth	1.77E-04	2.03 (11998,94,1697,27)
GO:0043647	inositol phosphate metabolic process	1.82E-04	2.50 (11998,48,1697,17)
GO:0071554	cell wall organization or biogenesis	2.09E-04	1.69 (11998,188,1697,45)
GO:0010264	myo-inositol hexakisphosphate biosynthetic process	2.16E-04	2.75 (11998,36,1697,14)
GO:0032958	inositol phosphate biosynthetic process	2.16E-04	2.75 (11998,36,1697,14)
GO:0033517	myo-inositol hexakisphosphate metabolic process	2.16E-04	2.75 (11998,36,1697,14)
GO:0046173	polyol biosynthetic process	2.16E-04	2.75 (11998,36,1697,14)
GO:0019725	cellular homeostasis	2.16E-04	2.75 (11998,36,1697,14)
GO:0009069	serine family amino acid metabolic process	2.20E-04	2.28 (11998,62,1697,20)

GO:0042592	homeostatic process	2.25E-04	2.14 (11998,76,1697,23)
GO:0018208	peptidyl-proline modification	2.48E-04	3.37 (11998,21,1697,10)
GO:0019344	cysteine biosynthetic process	2.84E-04	2.36 (11998,54,1697,18)
GO:0006873	cellular ion homeostasis	3.12E-04	2.79 (11998,33,1697,13)
GO:0030003	cellular cation homeostasis	3.12E-04	2.79 (11998,33,1697,13)
GO:0055082	cellular chemical homeostasis	3.12E-04	2.79 (11998,33,1697,13)
GO:0006397	mRNA processing	3.54E-04	2.46 (11998,46,1697,16)
GO:0006534	cysteine metabolic process	3.67E-04	2.31 (11998,55,1697,18)
GO:0019751	polyol metabolic process	3.67E-04	2.31 (11998,55,1697,18)
GO:0031323	regulation of cellular metabolic process	3.86E-04	1.32 (11998,690,1697,129)
GO:0006766	vitamin metabolic process	4.01E-04	3.77 (11998,15,1697,8)
GO:0046364	monosaccharide biosynthetic process	4.16E-04	2.36 (11998,51,1697,17)
GO:0009070	serine family amino acid biosynthetic process	4.71E-04	2.27 (11998,56,1697,18)
GO:0010468	regulation of gene expression	5.52E-04	1.33 (11998,634,1697,119)
GO:0009855	determination of bilateral symmetry	6.10E-04	2.36 (11998,48,1697,16)
GO:0009799	specification of symmetry	6.10E-04	2.36 (11998,48,1697,16)
GO:0008033	tRNA processing	8.77E-04	3.81 (11998,13,1697,7)
GO:0010114	response to red light	9.21E-04	5.05 (11998,7,1697,5)
GO:0006400	tRNA modification	9.21E-04	5.05 (11998,7,1697,5)

Reference

1. Aoki, K., Ogata, Y. & Shibata, D. Approaches for extracting practical information from gene co-expression networks in plant biology. *Plant Cell Physiol*. **48**, 381–90 (2007).
2. Arabidopsis, T. & Initiative, G. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**, 796–815 (2000).
3. Blanc, G., Barakat, a, Guyot, R., Cooke, R. & Delseny, M. Extensive duplication and reshuffling in the Arabidopsis genome. *Plant Cell* **12**, 1093–101 (2000).
4. Blanc, G., Hokamp, K. & Wolfe, K. H. A Recent Polyploidy Superimposed on Older Large-Scale Duplications in the Arabidopsis Genome. *Genome Research* **13**, 137–144 (2003).
5. Blanc, G. & Wolfe, K. H. Functional Divergence of Duplicated Genes Formed by Polyploidy during Arabidopsis Evolution. *Plant Cell* **16**, 1679–1691 (2004).
6. Bron, C. & Kerboscht, J. Algorithm 457 Finding All Cliques of an Undirected Graph [H]. *Communications of the* ACM **16**, (1973).
7. Brooke-Powell, E. T., Mandal, T. N., Ajioka, J. W. & Elizabeth, T. Use of Transcriptor Reverse Transcriptase in Microarray Analysis. *Biochemica* **1**, 27–30 (2004).
8. Chiusano, M. L., D'Agostino, N., Traini, A., Licciardello, C., Raimondo, E., Aversano, M., Frusciante, L., Monti, L. ISOL@: an Italian SOLAnaceae genomics resource. *BMC Bioinformatics* **9 Suppl 2**, S7 (2008).
9. Craigon, D. J., James, N., Okyere, J., Higgins, J., Jotham, J., May, S. NASCArrays: a repository for microarray data generated by NASC's transcriptomics service. *Nucleic Acids Res.* **32**, D575–D577 (2004).

10. De Bodt, S., Carvajal, D., Hollunder, J., Van den Cruyce, J., Movahedi, S., Inzé, D. CORNET: a user-friendly tool for data mining and integration. *Plant Physiol.* **152**, 1167–1179 (2010).
11. Eckardt, N. A. Cellulose Synthesis Takes the CesA Train. *Plant Cel.* **15**, 1685–1688 (2003).
12. Eden, E., Navon, R., Steinfeld, I., Lipson, D. & Yakhini, Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* **10**, 48 (2009).
13. Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, a W. From molecular to modular cell biology. *Nature* **402**, C47–52 (1999).
14. Hoheisel, J. D. Microarray technology: beyond transcript profiling and genotype analysis. *Nat. Rev. Genet.* **7**, 200–10 (2006).
15. Ihmels, J., Levy, R. & Barkai, N. Principles of transcriptional control in the metabolic network of Saccharomyces cerevisiae. *Nat. Biotechnol*. **22**, 86–92 (2004).
16. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. **28**, 27–30 (2000).
17. Koonin, E. V. Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet*. **39**, 309–38 (2005).
18. Koornneef, M. & Meinke, D. The development of Arabidopsis as a model plant. *Plant J.* **61**, 909–21 (2010).
19. Manfield, I. W., Jen, C. H., Pinney, J. W., Michalopoulos, I., Bradford, J. R., Gilmartin, P. M., Westhead, D.R. Arabidopsis Co-expression Tool (ACT): web server tools for microarray-based gene expression analysis. *Nucleic Acids Res*. **34**, W504–W509 (2006).
20. Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C. & Morris, Q. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biol.* **9 Suppl 1**, S4 (2008).

21. Mutwil, M., Klie, S., Tohge, T., Giorgi, F. M., Wilkins, O., Campbell, M. M., Fernie, A. R., Usadel, B., Nikoloski, Z., Persson, S. PlaNet: combined sequence and expression comparisons across plant networks derived from seven species. *Plant Cell* **23**, 895–910 (2011).
22. Mutwil, M., Obro, J., Willats, W. G. T. & Persson, S. GeneCAT-novel webtools that combine BLAST and co-expression analyses. *Nucleic Acids Res*. *36*, W320–W326 (2008).
23. Obayashi, T. & Kinoshita, K. Coexpression landscape in ATTED-II: usage of gene list and gene network for various types of pathways. *J. Plant Res*. **123**, 311–9 (2010).
24. Ogata, Y., Suzuki, H., Sakurai, N. & Shibata, D. CoP: a database for characterizing co-expressed gene modules with biological information in plants. *Bioinformatics* **26**, 1267–1268 (2010).
25. Ohno, S. "Evolution by gene duplications." *Book* (1970)
26. Oliver, S. Guilt-by-association goes global. *Nature* **403**, 601–3 (2000).
27. Palla, G., Derényi, I., Farkas, I. & Vicsek, T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **435**, 814–8 (2005).
28. Quackenbush, J. Genomics. Microarrays-guilt by association. *Science* **302**, 240–1 (2003).
29. Redman, J. C., Haas, B. J., Tanimoto, G. & Town, C. D. Development and evaluation of an Arabidopsis whole genome Affymetrix probe array. *Plant J.* **38**, 545–61 (2004).
30. Rhee, S. Y. 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 (2003).

31. Rost, B. Twilight zone of protein sequence alignments. *Protein Eng.* **12**, 85–94 (1999).
32. Sangiovanni, M., Vigilante, A. & Chiusano, M. L. Exploiting a Reference Genome in Terms of Duplications: The Network of Paralogs and Single Copy Genes in Arabidopsis thaliana. *Biology (Basel).* **2**, 1465–1487 (2013).
33. Schmid, M., Davison, T. S., Henz, S. R., Pape, U. J., Demar, M., Vingron, M., Schölkopf, B., Weigel, D., Lohmann, J. U. A gene expression map of Arabidopsis thaliana development. *Nat. Genet.* **37,** 501–6 (2005).
34. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–504 (2003).
35. Slonim, D. K. & Yanai, I. Getting started in gene expression microarray analysis. *PLoS Comput. Biol.* **5**, e1000543 (2009).
36. Srinivasasainagendra, V., Page, G. P., Mehta, T., Coulibaly, I. & Loraine, A. E. CressExpress: a tool for large-scale mining of expression data from Arabidopsis. *Plant Physiol*. **147**, 1004–1016 (2008).
37. Steinhauser, D., Usadel, B., Luedemann, A., Thimm, O. & Kopka, J. CSB.DB: a comprehensive systems-biology database. *Bioinformatics* **20**, 3647–51 (2004).
38. Toufighi, K., Brady, S. M., Austin, R., Ly, E. & Provart, N. J. The Botany Array Resource: e-Northerns, Expression Angling, and promoter analyses. *Plant J.* **43**, 153–163 (2005).
39. Tse-Wen Chang. Binding of cells to matrixes of distinct antibodies coated on solid surface. *J. Immunol. Methods* **65**, 217–223 (1983).

40. Van de Peer, Y., Maere, S. & Meyer, A. The evolutionary significance of ancient genome duplications. *Nat. Rev. Genet.* **10**, 725–32 (2009).
41. Van de Peer, Y., & Meyer, A. "Large-scale gene and ancient genome duplications" Chapter 6 in *The evolution of the genome* (2005)
42. Vandepoele, K., Quimbaya, M., Casneuf, T., De Veylder, L. & Van de Peer, Y. Unraveling transcriptional control in Arabidopsis using cis-regulatory elements and coexpression networks. *Plant Physiol*. **150**, 535–546 (2009).
43. Vision, T. J. The Origins of Genomic Duplications in Arabidopsis. *Science* **290**, 2114–2117 (2000).
44. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 (2010).
45. Zaidi, S. K., Young, D. W., Choi, J. Y., Pratap, J., Javed, A., Montecino, M., Stein, J. L., Lian, J. B., van Wijnen, A. J. Intranuclear trafficking: organization and assembly of regulatory machinery for combinatorial biological control. *J. Biol. Chem.* **279**, 43363–6 (2004).
46. Zimmermann, P., Hirsch-Hoffmann, M., Hennig, L. & Gruissem, W. GENEVESTIGATOR. Arabidopsis microarray database and analysis toolbox. *Plant Physiol*. **136**, 2621–2632 (2004).