

**Max-Planck-Institut für Molekulare Pflanzenphysiologie
Arbeitsgruppe Usadel**

Expression-based Reverse Engineering of Plant Transcriptional Networks

Dissertation

**zur Erlangung des akademischen Grades
"doctor rerum naturalium"
(Dr. rer. nat.)
in der Wissenschaftsdisziplin "Molekulare Pflanzenphysiologie"**

**eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät
der Universität Potsdam**

von

Federico Manuel Giorgi

Potsdam, den 28.07.2011

This work is licensed under a Creative Commons License:
Attribution - Noncommercial - No Derivative Works 3.0 Unported
To view a copy of this license visit
<http://creativecommons.org/licenses/by-nc-nd/3.0/>

Published online at the
Institutional Repository of the University of Potsdam:
URL <http://opus.kobv.de/ubp/volltexte/2011/5676/>
URN <urn:nbn:de:kobv:517-opus-56760>
<http://nbn-resolving.de/urn:nbn:de:kobv:517-opus-56760>

A big computer, a complex algorithm and a long time does not equal science

Robert Gentleman, SSC 2003, Halifax (June 2003)

Abstract

Regulation of gene transcription plays a major role in mediating cellular responses and physiological behavior in all known organisms. The finding that similar genes are often regulated in a similar manner (co-regulated or "co-expressed") has directed several "guilt-by-association" approaches in order to reverse-engineer the cellular transcriptional networks using gene expression data as a compass. This kind of studies has been considerably assisted in the recent years by the development of high-throughput transcript measurement platforms, specifically gene microarrays and next-generation sequencing.

In this thesis, I describe several approaches for improving the extraction and interpretation of the information contained in microarray based gene expression data, through four steps: (1) microarray platform design, (2) microarray data normalization, (3) gene network reverse engineering based on expression data and (4) experimental validation of expression-based guilt-by-association inferences. In the first part test case is shown aimed at the generation of a microarray for *Thellungiella salsuginea*, a salt and drought resistant close relative to the model plant *Arabidopsis thaliana*; the transcripts of this organism are generated on the combination of publicly available ESTs and newly generated ad-hoc next-generation sequencing data. Since the design of a microarray platform requires the availability of highly reliable and non-redundant transcript models, these issues are addressed consecutively, proposing several different technical solutions. In the second part I describe how inter-array correlation artifacts are generated by the common microarray normalization methods RMA and GCRMA, together with the technical and mathematical characteristics underlying the problem. A solution is proposed in the form of a novel normalization method, called tRMA. The third part of the thesis deals with the field of expression-based gene network reverse engineering. It is shown how different centrality measures in reverse engineered gene networks can be used to distinguish specific classes of genes, in particular essential genes in *Arabidopsis thaliana*, and how the use of conditional correlation can add a layer of understanding over the information flow processes underlying transcript regulation. Furthermore, several network reverse engineering approaches are compared, with a particular focus on the LASSO, a linear regression derivative rarely applied before in global gene network reconstruction, despite its theoretical advantages in robustness and interpretability over more standard methods. The performance of LASSO is assessed through several *in silico* analyses dealing with the reliability of the inferred gene networks. In the final part, LASSO and other reverse engineering methods are used to experimentally identify novel genes involved in two independent scenarios: the seed coat mucilage pathway in *Arabidopsis thaliana* and the hypoxic tuber development in *Solanum tuberosum*. In both cases an interesting method complementarity is shown, which strongly suggests a general use of hybrid approaches for transcript expression-based inferences.

In conclusion, this work has helped to improve our understanding of gene transcription regulation through a better interpretation of high-throughput expression data. Part of the network reverse engineering methods described in this thesis have been included in a tool (CorTo) for gene network reverse engineering and annotated visualization from custom transcription datasets.

Contents

| | |
|---|-----------|
| 1. Introduction..... | 1 |
| 1.1 Regulation of Transcription and Systems Biology | 1 |
| 1.2 Transcriptomics | 2 |
| 1.2.1 Transcriptomics from Northern blot to microarrays | 2 |
| 1.2.2 Microarray data preprocessing | 5 |
| 1.2.3 The future of Transcriptomics | 6 |
| 1.3 Gene network reverse engineering | 7 |
| 1.4 Biological scenarios of gene network reverse engineering..... | 12 |
| 1.4.1 The seed coat mucilage pathway | 12 |
| 1.4.2 Hypoxic tuber development in <i>Solanum tuberosum</i> | 14 |
| 1.4.3 Essential genes | 16 |
| 1.5 Summary of the aims of this thesis | 17 |
| 2. Results | 18 |
| 2.1 Generation of a custom microarray platform from next-generation mRNA sequencing data: the <i>Thellungiella salsuginea</i> transcriptome | 18 |
| 2.1.1 Collecting <i>Thellungiella</i> sequences | 18 |
| 2.1.2 Transcript assembly..... | 19 |
| 2.1.3 Transcriptome completion | 23 |
| 2.1.4 Comparative transcriptome considerations between <i>Thellungiella</i> and <i>Arabidopsis</i> | 23 |
| 2.2 Algorithm-driven Artifacts in median polish summarization of microarray data: tRMA..... | 26 |
| 2.2.1 Multi-array preprocessing effects | 26 |
| 2.2.2 Causes of RMA and GCRMA artifact generation | 28 |
| 2.2.3 Median polish inconsistency | 30 |
| 2.2.4 Comparison between RMA and tRMA in biological contexts | 34 |
| 2.2.5 Conclusions on median polish based microarrays normalization methods..... | 36 |
| 2.3 Combining network centrality analysis and conditional correlation: application to essential gene prediction | 37 |
| 2.3.1 Definition of Breaking Potential..... | 37 |

| | |
|--|-----------|
| 2.3.2 Comparison between Breaking Potential and other centralities in <i>Arabidopsis thaliana</i> coexpression networks | 38 |
| 2.3.3 Breaking Potential is a positive predictor for gene essentiality in <i>Arabidopsis thaliana</i> | 38 |
| 2.3.4 Conclusions on Breaking Potential as an essential gene predictor and future perspectives | 41 |
| 2.4 Expression-based gene network reverse engineering | 42 |
| 2.4.1 Custom network reverse engineering and method comparison: the CorTo tool | 42 |
| 2.4.2 Application of the LASSO to gene expression-based modeling | 44 |
| 2.4.3 Comparative analysis of expression-based methods for gene network reverse engineering | 45 |
| 2.5 LASSO and correlation for reverse engineering the seed coat mucilage pathway in <i>Arabidopsis thaliana</i> | 58 |
| 2.5.1 RHM2 expression network analysis | 58 |
| 2.5.2 Network reconstruction based on several mucilage genes | 60 |
| 2.6 LASSO and correlation for reverse engineering the hypoxia-regulated tuber development pathway in <i>Solanum tuberosum</i> | 69 |
| 2.6.1 Identification of hypoxia responsive ERFs in <i>Solanum tuberosum</i> | 69 |
| 2.6.2 StHREs expression during tuber development | 72 |
| 2.6.3 Characterization of StHREs co-regulators in tuber development by Spearman Correlation and the LASSO | 72 |
| 2.6.4 Conclusions on StHRE characterization and co-regulation analysis | 81 |
| 3. Discussion | 82 |
| 3.1 Transcript model characterization for microarray generation | 82 |
| 3.2 Caveats in microarray data normalization | 83 |
| 3.3 Conditional correlation techniques in central gene prediction | 84 |
| 3.4 Gene network reverse engineering | 86 |
| 3.5 Conclusions and future perspectives | 89 |
| 4. Materials and Methods | 91 |
| 4.1 Transcriptome assembly | 91 |
| 4.2 Comparison of Microarray preprocessing methods | 92 |
| 4.2.1 Microarray preprocessing methods | 92 |
| 4.2.2 Microarray datasets | 92 |
| 4.2.3 Permutation of microarrays | 93 |

| | |
|---|------------|
| 4.2.4 Inter-array correlation analysis | 93 |
| 4.2.5 Noise robustness analysis | 93 |
| 4.2.6 Linear model for measuring internal probeset consistency | 93 |
| 4.2.7 Transposed RMA (tRMA) | 94 |
| 4.2.8 AffyComp benchmark | 94 |
| 4.2.9 Sample classification performance | 94 |
| 4.3 Network Centrality and Breaking Potential calculations | 95 |
| 4.4 CorTo tool development..... | 95 |
| 4.5 Gene Network Reconstruction and Comparison | 95 |
| 4.6 Candidate selection through LASSO and Correlation analysis..... | 97 |
| 4.6.1 RHM2 - Selection of candidate genes | 97 |
| 4.6.2 Multi-gene mucilage networks | 97 |
| 4.6.3 StHRE1 and StHRE2a/b networks | 98 |
| 4.7 Sugar screening in <i>Arabidopsis thaliana</i> seed coat mucilage..... | 98 |
| 4.7.1 Seed Staining and Microscopy | 99 |
| 4.8 Gene expression in <i>Solanum tuberosum</i> tubers | 99 |
| 4.8.1 Phylogenetic analysis of StHRE genes | 99 |
| 4.8.2 Tuber growing conditions..... | 99 |
| 4.8.3 Sequencing of StHRE mRNAs | 99 |
| 4.8.4 mRNA extraction..... | 100 |
| 4.8.5 Expression measurement through Realtime RT-qPCR..... | 100 |
| 5. Appendix | 101 |
| 5.1 Sequence length distribution for the <i>Arabidopsis thaliana</i> Transcriptome | 101 |
| 5.2 Average Affymetrix inter-array correlation coefficients at different sample sizes, using three different microarray normalization procedures..... | 102 |
| 5.3 Example of Breaking Potential calculation | 103 |
| 5.4 Breaking Potential and other Centrality measures assessed for essential gene prediction power | 106 |
| 5.5 Full coding sequences of <i>StHRE1</i> , <i>StHRE2a</i> and <i>StHRE2b</i> | 107 |

| | |
|---|------------|
| 5.6 MapMan ontology bin enrichment analysis for top correlators of <i>StHRE1</i> and <i>StHRE2a/b</i> | 108 |
| 5.7 Expression Intensity and Variance for Transcriptional genes and Essential genes in <i>Arabidopsis thaliana</i> Affymetrix microarrays | 110 |
| 5.8 Comparative Network Reconstruction Methods Analysis - Additional Network Quality Assessments | 111 |
| 5.8.1 Example of a Network Degree distribution..... | 111 |
| 5.8.2 Fit to a power law of the Network Degree - adjusted R2 | 112 |
| 5.8.3 Fit to a power law of the Network Degree - Pearson Correlation coefficient of the Network Degree distribution | 114 |
| 5.8.4 Overlap to Protein-Protein interaction networks - Accuracy | 116 |
| 5.8.5 Overlap to Protein-Protein interaction networks - Matthew's coefficient..... | 118 |
| 5.9 Coefficient Distributions for Pearson Correlation and Mutual Information | 120 |
| 5.10 LASSO model for the <i>Arabidopsis</i> gene RHM2..... | 121 |
| 5.11 Mucilage release upon mechanical stress in a <i>Myb5</i> knockout line ... | 122 |
| 5.12 <i>Thellungiella salsuginea</i> seeds upon hydration..... | 123 |
| Publications | 124 |
| Posters | 124 |
| Curriculum vitae | 125 |
| Selbständigkeitserklärung | 126 |
| Acknowledgments | 127 |
| Bibliography | 128 |

List of Abbreviations

| | |
|-------|---|
| AIC | Akaike's Information Criterion (coefficient) |
| ERF | Ethylene Response Factor |
| EST | Expressed Sequence Tag |
| HRE | Hypoxia Response Element |
| LASSO | Least Absolute Selection and Shrinkage Operator |
| logFC | logarithmic Fold Change |
| MAS5 | MicroArray Suite 5 |
| MM | Mismatch probe |
| NGS | Next Generation Sequencing |
| ORF | Open Reading Frame |
| PCC | Pearson's Correlation Coefficient |
| PM | Perfect Match probe |
| PPI | Protein-Protein Interaction |
| RMA | Robust Multiarray Algorithm |
| SCC | Spearman's Correlation Coefficient |
| TF | Transcription factor |
| TFA | Trifluoroacetic acid |
| tRMA | transposed Robust Multiarray Algorithm |

1. Introduction

1.1 Regulation of Transcription and Systems Biology

The survival of all living organisms depends on their rapid adaptation in the ever-changing environmental conditions. One of these mechanisms of adaptation resides in the dynamic and selective activation of the genetic information contained in the chromosomes (Lodish et al., 2003). This process, which transcribes genes into transcripts, and transcripts into active proteins, has been described as the "central dogma of molecular biology" (Crick, 1970). In particular, the regulation of the first step of the central dogma, gene transcription, determines how many messenger RNA (mRNA) transcript copies are produced from a gene, and therefore it controls how many active proteins will be synthesized from a particular gene. This control, mainly investigated for the transcription initiation mechanism steps, but applied also over transcript maintenance and turnover, is precisely regulated in all organisms (Figure 1) but particularly important in plants (Taiz and Zeiger, 2006). In fact, these sessile organisms cannot depend on rapid muscular movement to overcome a wide range of potentially harmful events like flooding (Jackson and Colmer, 2005), drought (Zhu, 2002) or nutrient limitations (Lee et al., 2007) and must therefore rely on other mechanisms in order to survive, among which a rapid and dynamic capability to vary transcript expression. Transcriptional control is important not only for dealing with external events (Wilke et al., 1994), but also to carry out physiological processes like growth (Nasmyth and Shore, 1987), cell differentiation (Fitzsimmons and Hagman, 1996), homeostasis (Hastings et al., 2008) and the cell cycle (Mudryj et al., 1991).

No plant species has been studied, so far, in such a detail to be able to understand all transcriptional regulation mechanisms. However, for some of the simplest organisms, e.g. *Escherichia coli*, a nearly complete map of transcriptional regulation has been characterized through different single gene approaches (e.g. via gene knockouts and promoter inductional screenings), indicating which gene controls the transcription of which gene in specific conditions (Gama-Castro et al., 2011). These collections of details of transcriptional regulation phenomena have been gathered over more than 50 years by thousands of biologists, who tried to understand the function and interaction of single genes and cellular components (Madan Babu and Teichmann, 2003). This reductionist approach has recently been coupled with a more holistic approach, with the aim of explaining and predicting the behavior of the entire transcript population as a whole cellular subsystem (Kitano, 2002). A simple look at the *E.coli* transcription interaction network (Gama-Castro et al., 2008) (Figure 2) shows us that only a few regulation units act separated from the rest of the bacterial genes. In fact, it can be observed not only that many different transcriptional control mechanisms are interlaced, but also that transcript and protein levels are influenced by and influence other cellular systems, such post-translational modifications, metabolite levels and compartmentalization, membrane fluidity, cytoskeleton structure, osmolarity, etc. Everything considered, it is possible to achieve a full comprehension of any step of biological regulation, including transcription, only by considering it from a broader perspective, as a cog in a system of intertwined systems. This approach has been called "Systems biology" (Kitano, 2002) and has recently been

massively assisted by technology leaps towards the measurability of entire populations of molecular species, the so-called "omics" disciplines.

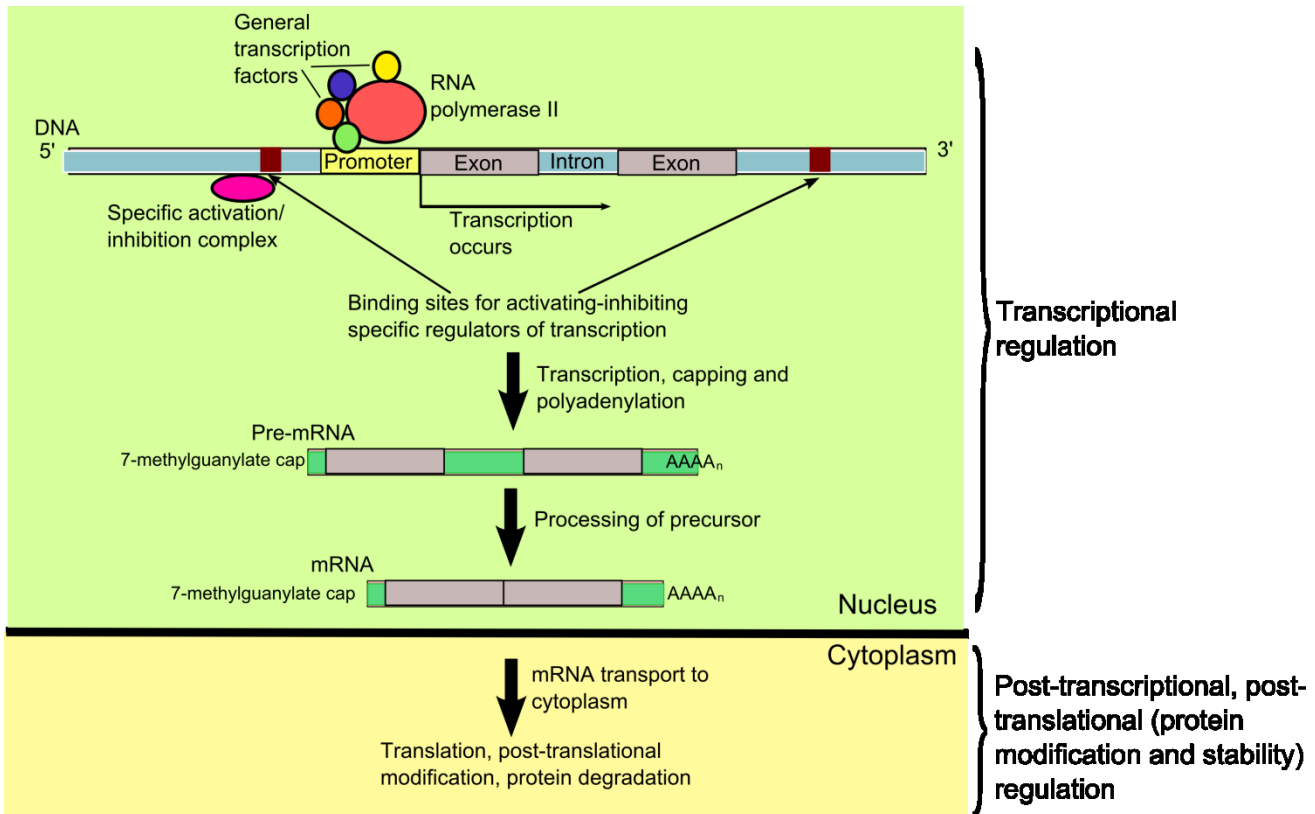


Figure 1 - Gene regulation in eukaryotes, with focus on the transcriptional control. This diagram is, in fact, a simplification, since several further tuning mechanisms for gene regulation are omitted (e.g. microRNA (Hobert, 2008) and differential codon usage (Gouy and Gautier, 1982)). Readapted from (Taiz and Zeiger, 2006)

1.2 Transcriptomics

1.2.1 Transcriptomics from Northern blot to microarrays

Amongst the -omics techniques, the first one that saw widespread use has certainly been Genomics (Cole and Saint Girons, 1994), i.e. the collection of sequence information contained in the inheritable genomic DNA. Currently, almost 2000 genomes from bacterial and eukaryotic organisms have been fully sequenced, and over 5000 sequencing projects are in progress (Lioliou et al., 2009).

However, Genomics is collecting static information, as the genetic information *per se* is usually not affected not affected by cellular events (with a few peculiar exceptions in the area of Epigenetics (Wolffe and Matzke, 1999)). The transcriptome however, i.e. the entire collection of transcripts in a species, is the key link between information encoded in the DNA and observable phenotypes. Particular genes are dynamically activated or repressed in response to an external stimulus, or to a physiological event. The study of the behavior of the mRNA population in response to these perturbations is a methodology of Systems biology, called

Transcriptomics, and this discipline is one of the main instruments today to model and understand the mechanisms of transcriptional regulation as a whole (Kirschner, 2005).

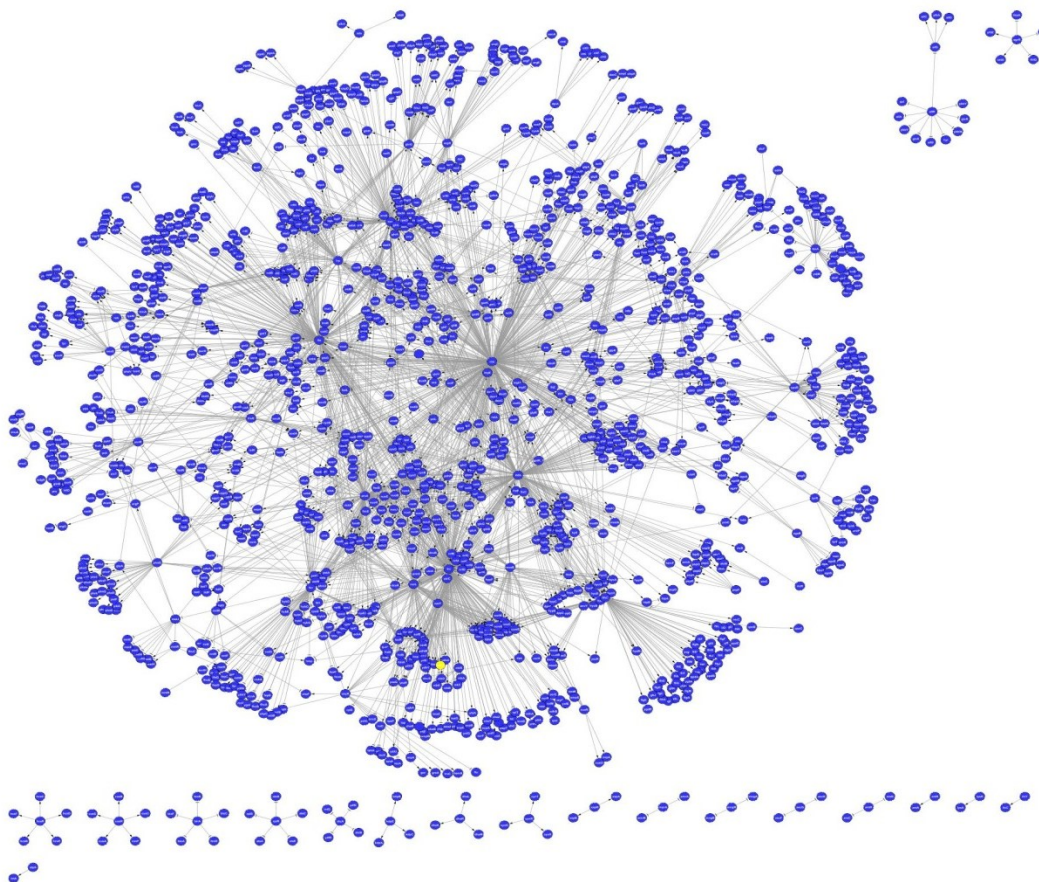


Figure 2 - Network representation of the known *Escherichia coli* regulation of transcription mechanisms. Every blue circle is a gene, every grey connection is a regulative relationship. The network is generated from the collection of experimentally validated genetic interactions collected in (Gama-Castro et al., 2008)

One of the first and most popular techniques to semi-quantitatively measure transcript abundances has certainly been Northern blot (Alwine et al., 1977), a direct evolution of the Southern blot method used for DNA (Southern, 1975). In Northern blot the total RNA population is extracted from a particular tissue or cell sample, then the RNAs are separated via electrophoresis on a gel and transferred to a nylon membrane (hence the term "blotting").

Then, following the Watson and Crick rules of double-helix nucleic acid complementarity (Watson and Crick, 1953), single-stranded labeled DNA probes are hybridized on this membrane (Thomas, 1983) (Kevil et al., 1997). This technique provides a semi-quantitative assessment of the abundance of specific mRNAs in the sample, together with the information about the approximate length of the mRNA annealed by the probe. Northern blot has been extensively used over the years for differential expression analyses and for comparative transcript abundance studies (Durand and Zukin, 1993) and is still used as a benchmark

procedure in molecular biology (Taniguchi et al., 2001) for investigating a limited amount of transcript types (Bor, 2006).

A principle similar to the specific-hybridization mechanism used in the Northern blot has been applied in the subsequent, higher throughput RNA measuring method, namely Real Time Polymerase Chain Reaction (RT-PCR). In RT-PCR, the mRNA population is reverse-transcribed to the more stable cDNA, and the PCR reaction is monitored after every PCR or thermal cycle to assess the increase of amplicons using specific primers and a colorimetric (Nolan et al., 2006) or a fluorescence (Morrison et al., 1998) assay. Nowadays a common run of RT-PCR (having plates with 96 separated reactions) is able to measure around ten times more transcripts at less than half the experimental time required by than Northern blot.

Despite the clear advancements obtained by RT-PCR, only with microarrays it has become possible to approach nearly full coverage of the transcriptome for transcript quantification (Ramsay, 1998). By automating the spotting on a chip of probes for thousands or tens of thousands of genes, high density oligonucleotide microarrays have become available. In microarrays, labeled samples are hybridized on the chip itself (Figure 3).

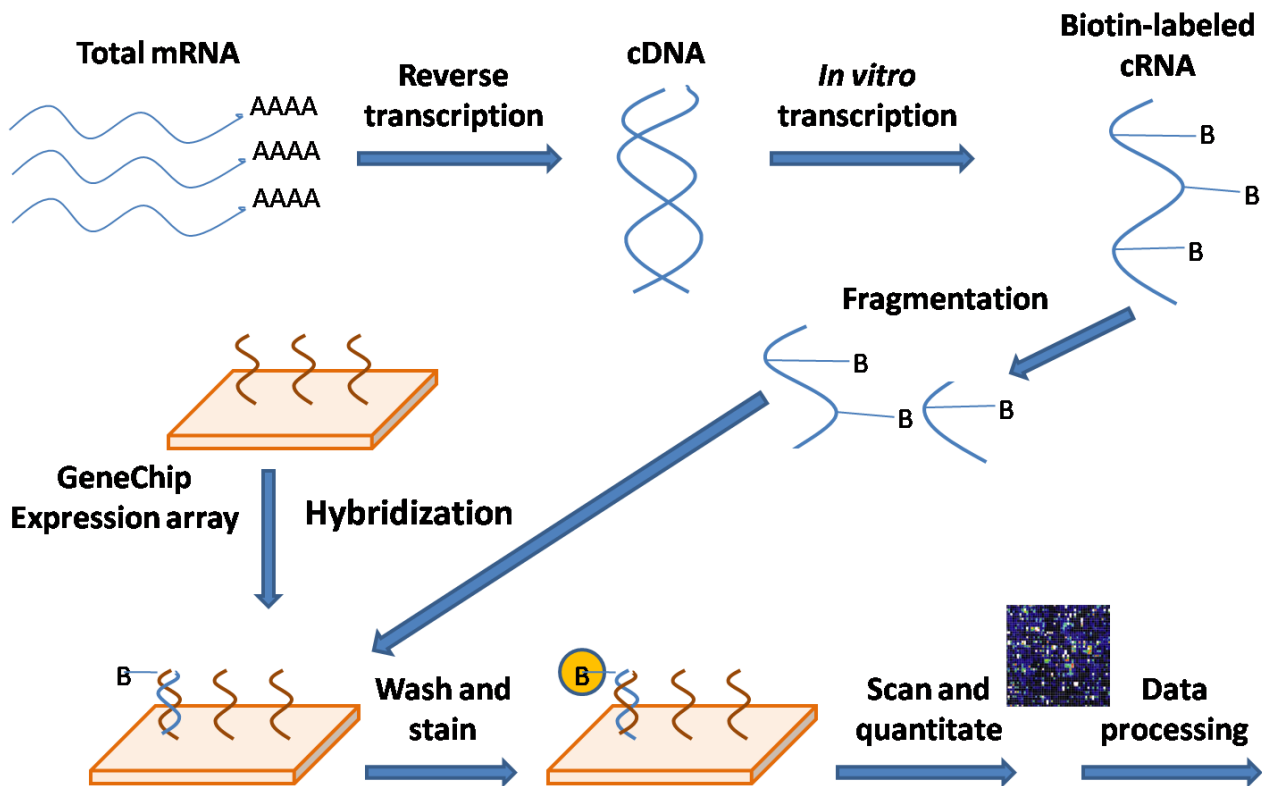


Figure 3 - Diagram of Affymetrix GeneChip type microarrays for mRNA quantification

The ready availability of microarrays has made them a widely used tool in many areas of biological research for quantitative, high-throughput measurements of gene expression. Publicly available databases alone store

a huge (and growing) quantity of microarray experiments (e.g. 338947 samples in Gene Expression Omnibus (Edgar et al., 2002) and 251711 in ArrayExpress (Parkinson et al., 2007)), comprising hundreds of different species.

The first type of microarray to be introduced were the so-called "two color" microarrays (Churchill, 2002), where two differently labeled samples are compared on the same chip and measured via two signal intensity (Red and Green) channels. However, the most popular platform today is arguably the single-channel Affymetrix GeneChip (Figure 3) (Affymetrix). For instance, in Gene Expression Omnibus they represent 97.9% of all arrays available for *Arabidopsis thaliana*, and 99.0% for *Homo sapiens*. In this technology each transcript is typically measured by a set of 11–20 pairs of 25 bases-long probes, collectively referred to as "probeset".

For every "perfect match" probe (PM), the Affymetrix chips contain a "mismatch" counterpart (MM), with a single nucleotide change in the middle of the PM probe sequence. The role of MM probes, located adjacent to the respective PM, is to measure probe-specific background signal associated to any perfect-match signal intensity.

1.2.2 Microarray data preprocessing

In general, the process of obtaining a single gene expression value out of raw probe intensity measurements is called "microarray preprocessing". Three steps are usually required for Affymetrix type arrays: background correction, normalization and summarization. Many different methods or combinations of methods were proposed over the years (Irizarry et al., 2006; Millenaar et al., 2006). The most popular manufacturer-provided method, MAS5 (Hubbell et al., 2002), uses a scale normalization approach, and then corrects the background by subtracting the mean intensity of the lowest 2% spots in every microarray region, and then MM intensities from the respective PM ones. Whenever the MM intensity is higher than the corresponding PM one, in order to avoid negative signal intensities, MAS5 replaces the MM signal with an "idealized mismatch" value (IM) derived from other values in the same probeset. This was a significant improvement over the MAS4 normalization which could result in negative signal intensities (Affymetrix; Zhou and Abagyan, 2003) To extract final probeset intensities, MAS5 calculates a robust average (Tukey's biweight) of all the probes contained in a probeset.

Many alternative techniques have challenged MAS5 supremacy for preprocessing. Being a single-array technique, MAS5 doesn't model probes' behavior across different samples, and therefore suffers from high variance and is theoretically less robust than algorithms taking multiple arrays into account (Irizarry et al., 2003) (Wu and Irizarry, 2004). On the other hand, MAS5 normalization doesn't depend on the nature of the samples analyzed, and therefore will yield identical results for a given microarray in any dataset considered. Two of the most popular multi-array normalization techniques are RMA (Irizarry et al., 2003) and GCRMA (Wu and Irizarry, 2005). RMA doesn't use any information contained in MM probes, and calculates background signal by performing a modeled global correction of all PM intensities. Then it applies a quantile normalization step and a median polish summarization, which accounts for probe intensities over multiple arrays. GCRMA applies the same normalization and summarization steps as RMA, but it differs in the background correction

method, which is based on the probe sequence. Other multi-array methods which don't discard MM intensities exist, one of them being dChip (Li and Wong, 2001). However, I will focus here on the likely most popular microarray normalization methods, specifically RMA, GCRMA and MAS5 (Gentleman et al., 2005; Bolstad, 2008). Their popularity is illustrated by the fact that they are the most applied normalization techniques in online databases (Usadel et al., 2009). To assess the properties of these different preprocessing techniques, most benchmarks were specific for differential gene expression scenarios, the original purpose for which microarrays were developed (Schena et al., 1995). To do so, golden set spike-in samples were used, with known concentrations of transcripts (Cope et al., 2004) (Irizarry et al., 2006), or Real Time PCR measurements were performed for a comparison (Gyorffy et al., 2009). The outcome of these benchmarks has not identified any technique as the top performer, although single-array techniques such as MAS5 have been outperformed by multi-array ones such as RMA (Irizarry et al., 2003; Therneau and Ballman, 2008; Gyorffy et al., 2009).

However, biological investigation was not limited to the analysis of differentially expressed genes. Indeed, many different approaches to biological investigation have relied on microarrays, ranging from gene and sample clustering (Golub et al., 1999) to gene-gene network reverse-engineering (Basso et al., 2005), from sample classification (Nielsen et al., 2007) to global transcript models (Usadel et al., 2008). The field of microarray data correlation and clustering based on the principle of coexpression has developed at a quite considerable pace (Boutros and Okey, 2005); despite this, the effects of preprocessing on coexpression analyses have been generally overlooked, with a few exceptions. (Harr and Schlotterer, 2006) used bacterial operons to validate the different normalization techniques for correlation analysis and concluded that a combination of different methods works best. On the other hand, (Lim et al., 2007) have pointed out how the use of the multi-array techniques RMA and GCRMA can yield inter-array correlation artifacts and generally lower quality networks than the older MAS5. In particular, a specific step in GCRMA background correction (the gene-specific binding correction, or GSB) has been identified as partially responsible for the spurious correlations generated by GCRMA. Notably however, the correction of this step is not sufficient to remove all artifact effects, and no explanation was provided for artifacts produced by RMA.

1.2.3 The future of Transcriptomics

Although vast, the number of transcripts measurable by any microarray platform is limited to the amount of probes present on the chip itself, and therefore relies on the prior knowledge about genes and their sequences. This has been partially overcome with the introduction of "tiling" arrays, which cover almost the entire genome of an organism with specific probes (Mockler and Ecker, 2005), allowing to measure the transcription of intergenic regions (Kapranov et al., 2007) and assess differences between splice variants (Wang et al., 2003). However, it remains challenging for microarrays to measure transcripts in organisms whose genomic sequence is not known; in these cases, only cross-hybridization on a microarray designed for a different species is possible, with obvious issues given by sequence evolutionary divergence (Lu et al., 2009).

Recently, the development of high-throughput next generation sequencing (NGS) technologies (Sultan et al., 2008) has started a revolution in the field of Transcriptomics (Wang et al., 2009). The application of NGS to the field of transcript quantification takes the collective name of "RNA-Seq" and is based on three major techniques (Mardis, 2008; Wang et al., 2009), specifically 454 "pyrosequencing" (www.454.com) and the similar Illumina (www.illumina.com) and SOLiD (www.appliedbiosystems.com) "sequencing by synthesis" methods. These three techniques, despite their chemical differences, share the capability to obtain vast amounts of sequences, or reads, in short time, for example allowing the generation of 100 millions of nucleotide bases in roughly 7 hours (454 FLX, (Mardis, 2008)). Providing appropriate gene models (which can also be generated via NGS (Bai et al., 2011)), the reads generated by these techniques can be aligned and the transcript abundance can therefore be estimated for each gene in a discrete way. In plant science, the capability of RNA-Seq to be independent from pre-existing genomic knowledge about a particular organism has opened the possibility to assess global transcript variation in nonmodel species, e.g. the orchid *Phalaenopsis* (Hsiao et al., 2011) or the insect-eating plant *Sarracenia* (Srivastava et al., 2011). Another advantage of RNA-Seq compared to microarrays is its capability to obtain sequences without pre-designing a matching probe, and therefore allowing the detection of splicing variants, point mutations and microRNAs.

All together, it is logical to expect that RNA-Seq is going to replace microarrays in the near future (Sultan et al., 2008). However, microarrays are not only still cheaper, but are still highly competitive with RNA-Seq in several scenarios (Agarwal et al., 2010), such as the characterization of differential gene expression between male and female *Drosophila pseudoobscura* fruit flies (Malone and Oliver, 2011). Furthermore, the incredibly high amount of information collected over more than 15 years with microarrays in countless different species and conditions cannot be ignored (Edgar et al., 2002). In this intermediate period, next generation sequencing and microarrays could very well be complementary techniques, rather than be considered as competitors. Since all types of microarray platforms require a specific hybridization to occur between the sample mRNAs/cDNAs and the probes on the chip, it is mandatory to select specific and reliable probe sequences: this design is usually performed based on the current transcript information available for the selected organism (Gasieniec et al., 2006), but could be massively sped up by a NGS-based transcript population definition. This NGS-microarray combined principle has been tried before, for example 454 pyrosequencing has been used to qualitatively draft a transcriptome with the purpose of designing novel microarrays in the poorly characterized butterfly *Melitaea cinxia* (Vera et al., 2008). The task is however paved with technical issues, such as the necessity to assemble smaller sequences into larger representative transcript models, and conceptual issues, like the problem of normalizing the different transcript abundances in order to be able to detect rare RNAs.

1.3 Gene network reverse engineering

Transcriptional coordination, also called co-expression or co-regulation, has been observed in several biological contexts between functionally related genes (Stuart et al., 2003; Yu et al., 2003). Thus, co-expression has been successfully exploited in a range of model organisms, including yeast (Yu et al., 2003), human (Lee et al., 2004) and other mammals (Wolfe et al., 2005). Consequently, using this "guilt by association" approach, transcriptome-wide gene function inference and biological pathway discovery has been

possible (Wei et al., 2006; Yonekura-Sakakibara et al., 2008; Usadel et al., 2009). For example, cellulose synthase genes (CESAs) have been showed to be co-expressed in *Arabidopsis thaliana* (Figure 4) and to be interacting in the same cellulose biosynthetic pathway; following this principle, two further genes were found to be coexpressed with the CESAs and characterized as displaying cellulose synthesis deficiencies (Brown et al., 2005; Persson et al., 2005).

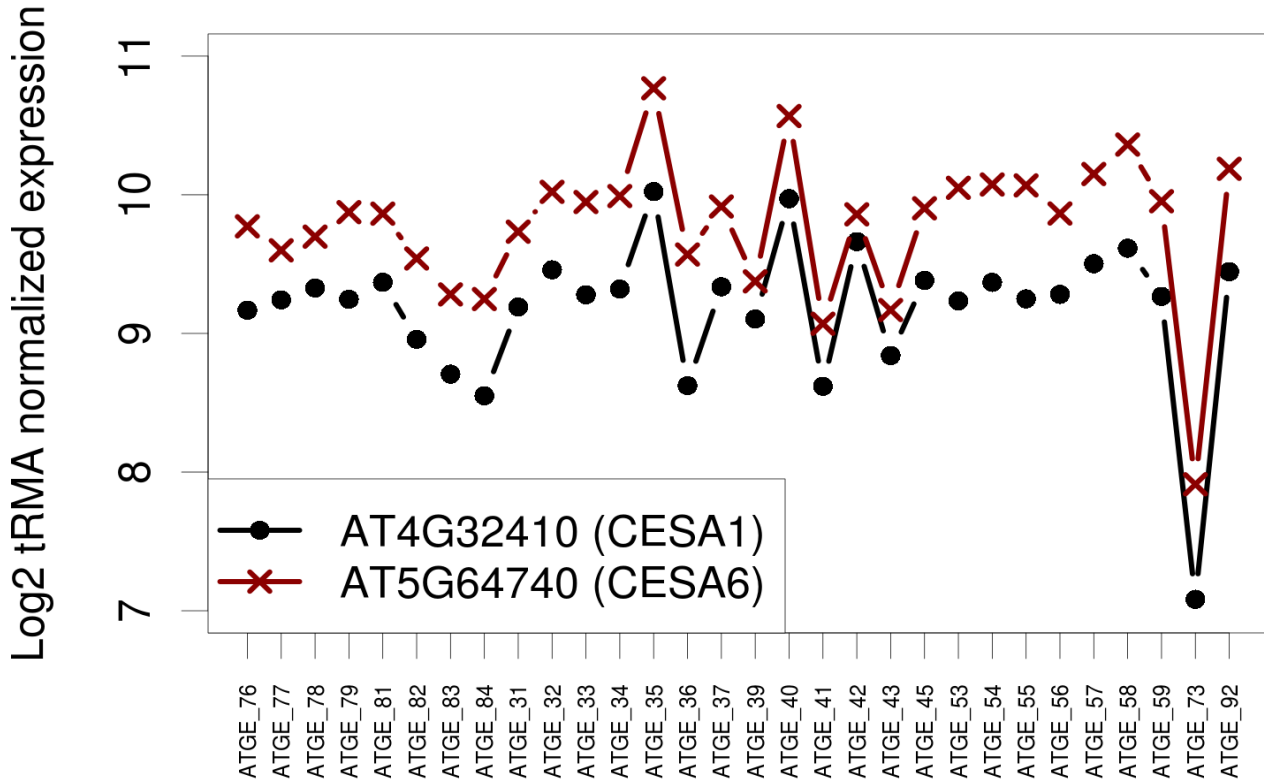


Figure 4 - Co-regulation of two cellulose synthase genes in *Arabidopsis thaliana* across several seeds and siliques tRMA-normalized (Paragraph 4.2.7) microarray samples from (Schmid et al., 2005)

Ideally, the large amount of transcript data publicly available (Parkinson et al., 2007; Leinonen et al., 2011) would be a gold mine for transcriptome-wide co-expression screenings, also considering that Metabolomics and Proteomics datasets are currently approaching near full genome scale (Weckwerth, 2010; Dunn, 2011). Unfortunately for these "systems" scopes, most experiments found in the literature have a single and precise biological question, e.g. which genes are up-regulated upon challenge with a specific abiotic stress. Such questions can be tackled via differential expression analysis, and solved due to recent advances in statistical analysis of microarray data. However, it is not immediately clear how different experimental series can be combined to reveal novel unaccounted information, not necessarily directly pertaining to the experimental question at hand. Thankfully, some successes have come from early work done using simple clustering (Eisen et al., 1998) or correlation approaches to infer biological themes or -more recently- to apply machine learning techniques to different experiments in order to infer the biological function of candidate genes (Brown et al.,

2000), the requirement for certain genes for a viable organism (Mutwil et al., 2010) or to predict the subcellular localization of genes (Ryngajllo et al., unpublished).

In the recent biology history, this has been complemented by network representations which have been successfully employed to capture various cellular relationships, ranging from protein-protein interactions (Breitkreutz et al., 2007) to gene regulations (Gama-Castro et al., 2008) and metabolic conversions (Yamada and Bork, 2009). In these networks, biological entities (e.g. genes, proteins, and metabolites) are represented as vertices, and their interactions are represented as edges. Biological networks can be assembled based principally on two different methods: (1) experimental evidence (i.e., existing knowledge) on the relationships between the considered entities, usually stored in a database form (e.g. (Wingender et al., 2000; Peri et al., 2003; Breitkreutz et al., 2007; Caspi et al., 2008)) and (2) network reconstruction from data profiles (Hartemink, 2005). Since direct evidence on protein-protein interaction and direct transcription control is experimentally time- and cost- consuming to obtain, many studies have focused on inferring large-scale biological relationships from expression data, in an approach called network reverse engineering (He et al., 2009). This approach aims at revealing the complete structure of relationships between molecular species within a biological system by applying suitable similarity measures (e.g., correlation, Euclidean distance, Mutual information (D'haeseleer et al., 2000)) or by using more sophisticated algorithms, e.g. probabilistic graphical models (Friedman, 2004).

In this framework, microarray datasets and transcriptional measurements obtained from comparable platforms (Schmittgen et al., 2008; Wang et al., 2009), have led to the prolific application of reverse engineering in the context of gene expression (He et al., 2009). These approaches have proven useful in both small-scale scenarios, e.g. for determining novel drug targets in the human B-cell leukemia gene network (Basso et al., 2005) and large-scale studies, e.g. for validating the transcription network of *Escherichia coli* (Faith et al., 2007) or *Arabidopsis thaliana* (Mutwil et al., 2010).

When similarity measures are applied in reconstruction of gene regulatory networks, two possibilities can be considered. The first consists in assessing the co-regulation of two genes via *direct* methods, that quantify the relationship without considerations over the rest of the gene population. The second implies assessing a relationship via *conditional* methods, which try to filter out indirect effects from each gene-gene pair by removing the effect of the other genes measured, process that is called *conditioning* (D'haeseleer et al., 2000; Zampieri et al., 2008).

The most known direct methods are Pearson correlation and Spearman correlation, which applies Pearson correlation after transforming the values of the variables to be correlated into ranks. Pearson correlation is able to assess direct, linear relationships (Butte and Kohane, 1999), while the rank-transformation of Spearman correlation makes it able to detect also non-linear (but monotonic) relationships and arguably more robust to outliers (Usadel et al., 2009). Another direct method, Mutual Information, has also seen a broad application in gene network reconstruction (Butte and Kohane, 2000; Daub et al., 2004; Margolin et al., 2006); Mutual Information tries to predict the behavior of one gene via the expression of another one, based on the

interpretation of the informational entropies of the two expression patterns: this is achieved by discrete binning of the distributions and is able to assess also non-linear and complex interactions. A fourth widely applied direct method is Euclidean Distance (Wen et al., 1998), which simply tries to calculate the relative distances of genes considered as points in multi-dimensional spaces (where every dimension is a measurement).

However powerful, direct methods lack the capability to grasp one layer of understanding of co-regulation networks. As depicted in Figure 5, if one (or more) intermediate genes (gene Z) exist between two genes for which we want to assess co-regulation (gene X and gene Y), an indirect correlation is observable. Conditional methods deal with this phenomenon, taking any X-Y relationship and conditioning it to one or several other genes in the same dataset (de la Fuente et al., 2004; Frenzel and Pompe, 2007). Simply put, they specify the degree of relationship between two genes X and Y when the effect of a third variable Z (or several other variables) is removed.

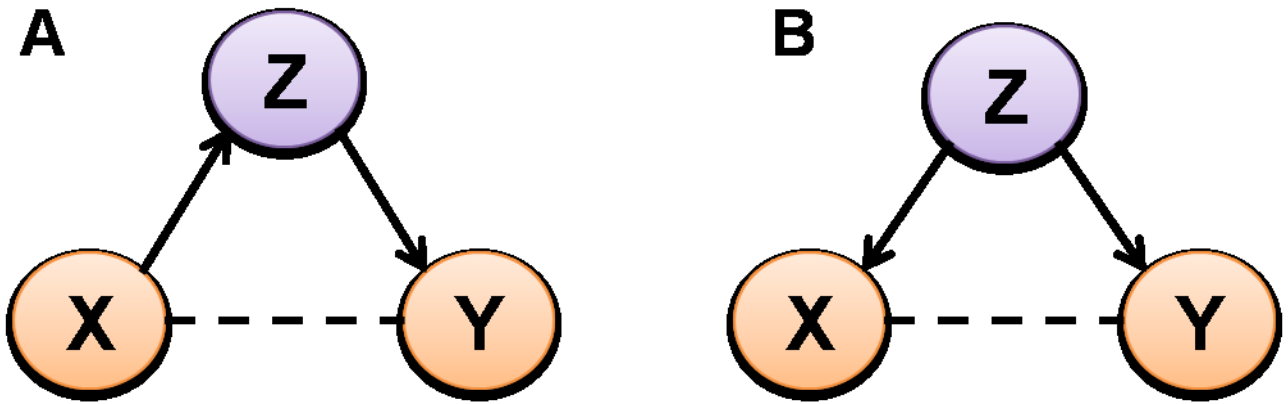


Figure 5 - Examples of indirect gene relationships yielding correlation. (A) gene X activates gene Z, which in turn activates gene Y. Correlation will be observable between gene X and Y (dashed line). (B) Gene Z is the common activator of gene X and gene Y, which therefore will appear as (indirectly) co-regulated.

Both correlation-based (de la Fuente et al., 2004) and Mutual Information-based (Frenzel and Pompe, 2007) direct methods have conditional (or partial) counterparts. In the simplest case, Conditional Pearson correlation (also called Partial Pearson correlation or higher order Pearson correlation), can be determined for two genes X and Y based on the standard direct Pearson correlation coefficient (or zeroth order Pearson correlation coefficient, Equation 1). Conditional correlation coefficients can then be derived directly from standard correlation coefficients, as in Equation (2) (de la Fuente et al., 2004).

$$\text{zeroth order correlation: } r_{xy} = \frac{\text{cov}(xy)}{\sqrt{\text{var}(x)\text{var}(y)}}, \quad (1)$$

$$\text{first order correlation: } r_{xy.z} = \frac{r_{xy} - r_{xz} r_{yz}}{\sqrt{(1-r_{xz}^2)(1-r_{yz}^2)}}, \quad (2)$$

$$\text{second order correlation: } r_{xy.zq} = \frac{r_{xy.z} - r_{xq.z} r_{yq.z}}{\sqrt{(1-r_{xq.z}^2)(1-r_{yq.z}^2)}} \quad (3)$$

If the Partial correlation coefficient calculated decreases significantly when compared to the original zeroth order correlation coefficient, then gene Z can be hypothesized as one of the common causes or as an intermediate variable in sequential pathways (as in Figure 5). The conditional coefficients for a dataset of n genes can be calculated up to the order of $n-2$, removing the effect of additional genes from a particular correlation. This can be done simply by expanding the following Equation 2 to an higher order correlation (Equation 3 for second order correlation) and continue incrementally.

Conditional methods have been used in genome-wide gene network reconstruction (Schäfer and Strimmer, 2005; Veiga et al., 2007), and combined with information theory approaches (Reverter and Chan, 2008). The conditional approach is particularly fit in reverse engineering pathways of genetic regulation, such as signal transduction cascades, where they manage to identify indirect correlation effects (Zampieri et al., 2008).

However powerful, the applicability higher order of Conditional/Partial methods, based on sequentially growing formulas as in Equation 2 and 3, is limited due to underlying algorithmic complexity leading to very long runtimes when increasing the conditioning order above one (de la Fuente et al., 2004). This is a major drawback, since a full conditioning against all other elements would be desirable in order to obtain truly direct connections between genes. Conditioning approaches based on the classical framework discussed above needed more data points (*i.e.* samples) than variables (*i.e.* genes), a scenario rarely found for microarray datasets. However, it was proposed to extract the fully conditioned data using several approximate approaches to obtain numerically stable solutions, for example by transforming the correlation matrix to an equivalent shrunken version (Opgen-Rhein and Strimmer, 2007). This solution, while approximate in its nature, allowed the total reconstruction of gene-gene networks (sometimes referred to as graphical Gaussian model) on the full genome level, mathematically removing the effect of all other measured genes (Ma et al., 2007). An analogous approach was taken by (Friedman et al., 2000) that tried to rebuild networks using probabilistic (Bayesian) approaches. While these approaches initially limited the measurements to a few discrete values, recent developments have allowed more flexibility (Werhli and Husmeier, 2007). Despite these recent developments, the performance of these reconstruction approaches in unveiling the transcriptional control mechanisms is still poorly understood. Moreover, an intrinsic flaw of probabilistic networks is that they necessarily have to rely on *a priori* assumptions, assuming for example the joint distribution of the variables in the dataset. Despite these shortcomings, (Werhli et al., 2006) analyzed the performance of several probabilistic and correlation-based approaches. They came to the conclusion that, in both simulated networks and well studied real pathways, probabilistic and partial correlation methods performed better than simple correlation, but only by a small margin. In fact, simple correlation networks have been very successful in the field of functional gene annotation (Mutwil et al., 2011) or biomarker prediction (*e.g.* (Qiu et al., 2007)), since in these cases the question of true direct connections is of lesser interest. However, in some circumstances it may be of central importance to detect the real connections, for example a common transcription factor activating a cluster of similar co-regulated isozymes, rather than connecting everything activated by the same transcriptional mechanism. Recently, global network properties, like the

number of connections for every gene (network degree) have been associated to biological properties of the genes themselves. For example, it has been shown that products of essential genes tend to interact with more partners than non-lethal ones in yeast (see also Paragraph 1.4.3) (Yu et al., 2004) and that disease genes tend to be found in particularly dense and interconnected areas of the human protein-protein interaction network (Rambaldi et al., 2008). However, no connection has been made so far between network degree and gene properties after removal of indirect connections via conditional approaches, giving the initial hint to the analysis described in Paragraph 2.3. It can also be expected that the successful field of network pattern evaluation (Milo et al., 2002) could benefit from the capability to detect indirect correlations, since the effect of these spurious connections in biological coexpression networks has never been studied in relation to topology or centrality. Another way to improve the reliability of an edge (*i.e.* a connection between two genes) are cross-species approaches profiting from conserved edges and network architectures between core biological functions (Stuart et al., 2003; Tsaparas et al., 2006). However, the problem of improving network reconstruction and biological validation still remains unsolved to this day. Another technique still not broadly applied for gene network reverse engineering (although some exploratory studies exist, such as (Gustafsson et al., 2009)) is notably the Least Absolute Shrinkage and Selection Operator, or the LASSO (Tibshirani, 1996). The LASSO is a linear regression-based approach, which can work in scenarios with more variables than samples (like e.g. microarrays), providing a robust set of interactions and the capability of removing indirect connections, like conditional correlation (Hastie et al., 2001). More details on the LASSO and our implementation of it into gene network analysis will be provided in Paragraph 2.4.2).

1.4 Biological scenarios of gene network reverse engineering

As previously stated, network approaches have been extensively applied in several biological studies following the "guilt-by-association" approach, *i.e.* using a gene, known to be involved in a specific pathway, function or biological process, as a "bait" to infer more counterparts in the same context (Wolfe et al., 2005; Aoki et al., 2007; Peng and Weselake, 2011). Using coexpression and network approaches many new genes have been characterized (e.g. in the cellulose synthase pathway (Persson et al., 2005; Mutwil et al., 2010), during starch biosynthesis (Kossmann et al., 1991), etc.) and even entire families of genes have been associated to a putative function (e.g., the cytochrome P450 superfamily (Ehlting et al., 2006)). I will now focus on two still not fully understood plant biological pathways, which have constituted my test cases for evaluating novel gene network reverse engineering approaches in the course of this dissertation. Finally, I will talk about essential genes (*i.e.* genes whose deletion determines the organism's death) and their properties in biological and expression-based networks.

1.4.1 The seed coat mucilage pathway

Higher plants' seed coat is composed of specialized tissues that provide protection to the embryo and assist in seed germination and dispersal. In some plant species, including *Arabidopsis thaliana*, the epidermal cells of the seed coat host a considerable amount of mucilage, containing large quantities of relatively unbranched pectin (Macquet et al., 2007). When dry *Arabidopsis* seeds are placed in an aqueous environment, the mucilage is released (extruded) and completely envelops the seed (Macquet et al., 2007). This pectinaceous

system has been suggested to be important for seed hydration and germination, attachment to soil components and for preventing gas exchanges (Caesele et al., 1981; Boesewinkel and Bouman, 1995), although seeds deprived of mucilage are still viable. Seed coat mucilage is produced and released by a specific family of cells, characterized by a typical volcano-like shape (Western, 2006). The seed coat secretory cells undergo a consistent differentiation process during seed development, which is accompanied by different stages of cell wall and pectin production. Mucilage is generally considered a superb model in which to study pectin biosynthesis during development, because, unlike mature tissues, it can be easily extracted without waiting for plants to be fully grown and without killing the seeds, since *Arabidopsis* plants can live even without this pectinaceous matrix (Western, 2006).

Mutations in a number of genes have been associated to altered mucilage production and/or release in the *Arabidopsis* seed coat. These include several transcription factors and development regulators, such as *AP2*, *TTG1*, *GL2*, *TT2*, *TT8*, *EGL3*, *MYB5*, *MYB61*. Furthermore, through screening of mucilage-defective mutants, five "**MU**cilage-**M**odified" (MUM) loci have been identified, which seem to act specifically in certain steps of mucilage production and release (Western et al., 2001). *Mum3* and *mum5* mutants show mucilage of altered composition, while *MUM1* (also known as *LUH* or Leunig Homolog (Huang et al., 2011)) and *MUM2* (also known as *BGAL6* or Beta-Galactosidase 6) have a key role in mucilage release. The sub-pathway controlled by *LUH*, regulating the expression of the genes *BGAL6*, *BXL1* and *SBT1.7*, is required for modifying the branching structure of pectins (Huang et al., 2011). *MUM4* is clearly acting in a biosynthetic step of mucilage production as cloning of the underlying gene revealed this to be *RHM2* (Rhamnose Biosynthesis 2) which is coding for a UDP-L-rhamnose synthase (Usadel et al., 2004). In addition to these genes, eight enhancer loci (called MEN: Mum-ENhancers) have been identified in the context of an already present *RHM2* inactivation, showing reduced mucilage production and release (Arsovski et al., 2009). It is noteworthy to observe that, among this collection of cloned loci, only a few have been associated to the biosynthesis of seed coat mucilage, and concomitantly much more has been discovered on the upstream signaling cascades (Figure 6, inferred from (Arsovski et al., 2009) and (Huang et al., 2011)). In the general picture, half of the characterized genes seem to have an exclusive transcriptional regulation function, while the other half possess enzymatic capabilities (see Table 8, page 61, for a summary). Also, the nature of the mucilage deficiency caused by knocking out these genes is not unique: for some mutants (e.g. *mum4*) the mucilage release can be triggered by addition of EDTA, which acts possibly by removing Ca^{2+} ions from the cell wall, while the inactivation of other genes (e.g. the upstream regulator *AP2*) impairs the very synthesis of mucilage (Arsovski et al., 2009).

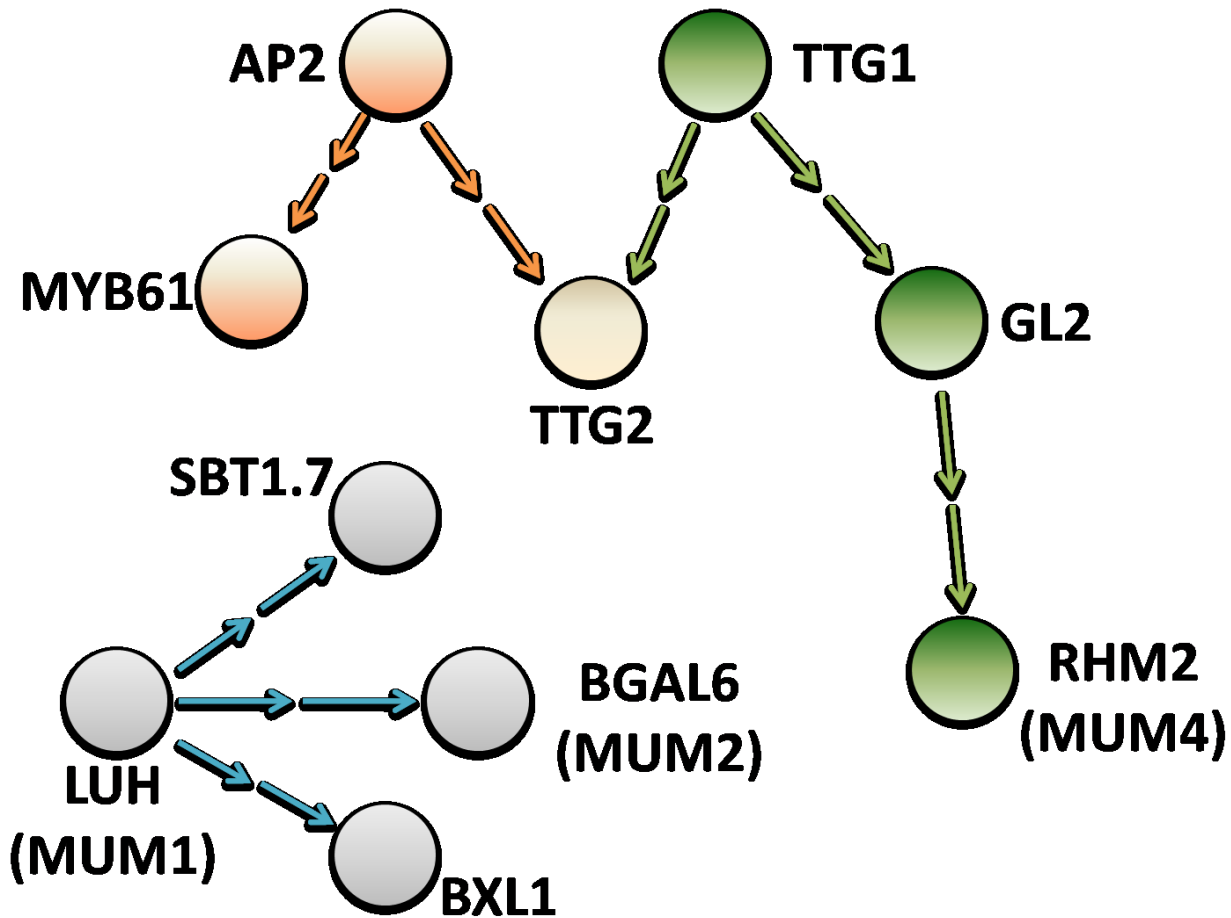


Figure 6 - Gene regulatory network for the seed coat mucilage pathway in *Arabidopsis thaliana* (inferred from (Arsovski et al., 2009) and (Huang et al., 2011))

Although several genes are constantly found to be involved in the mucilage synthesis and/or modification (Arsovski et al., 2010), several players are still missing in the network summarized above (Figure 6). The existence of genes known to be involved in this pathway and transcriptionally measurable via the *Arabidopsis thaliana* Affymetrix microarray (Affymetrix), together with the possibility to experimentally confirm alterations in the seed coat mucilage sugar composition via chromatography (Ip et al., 1992), make this scenario an ideal candidate for expression-based gene network reverse engineering (results of this approach in Paragraph 2.5).

1.4.2 Hypoxic tuber development in *Solanum tuberosum*

Seeds are also among those plant tissues characterized by physiological hypoxia, i.e. oxygen shortage (Borisjuk and Rolletschek, 2009), together with other bulky or high metabolism tissues such as fruits (Banks, 1983), seedlings (Van Dongen et al., 2003) and tubers (Geigenberger et al., 2000).

Hypoxia is in fact not only an external stress condition for aerobic living organisms, but also a physiological event during the development of multicellular organisms (Fukao and Bailey-Serres, 2004). For instance, mammalian embryo development takes place at low oxygen levels *in vivo*, and hypoxic conditions significantly

contribute to its correct progression *in vitro* (Forristal et al.). In plants, it has been known for over 90 years (Magness, 1920) that certain organs experience low oxygen conditions during growth and development (Geigenberger, 2003). Hypoxia, both in stress and developmental conditions, is associated with wide changes of transcript (Lasanthi-Kudahettige et al., 2007) and metabolite abundances (Biais et al., 2009), leading in general to a suppression of metabolic activities as an adaptive response to save ATP and decrease oxygen consumption (Geigenberger et al., 2000).

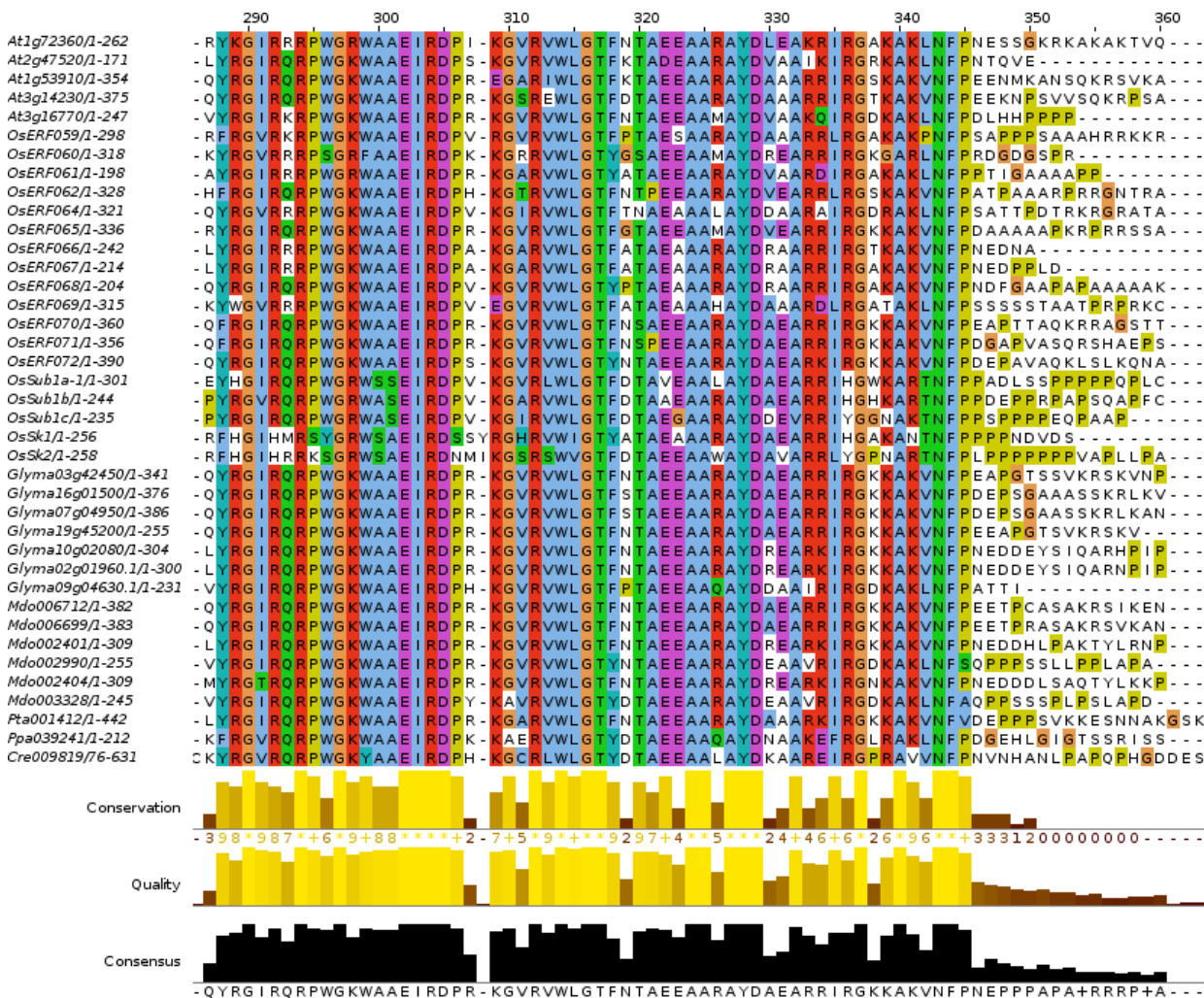


Figure 7 - Multiple sequence alignment of the conserved ERF domain from several hypoxia-related ERF proteins in *Arabidopsis thaliana* (At), *Oryza sativa* (Os), *Glycine max* (Glyma), *Malus domestica* (Md), *Pinus taeda* (Pta), *Physcomitrella patens* (Ppa) and *Chlamydomonas reinhardtii* (Cre). Multiple alignment was performed with MUSCLE (Edgar, 2004).

Transcript profiling of different plant species under low oxygen conditions revealed that several transcription factor (TF) gene families induced by low oxygen are conserved across the plant kingdom (Mustroph et al., 2010). The similar behavior of hypoxia-responsive TFs, together with the high conservation of specific protein domains, suggests a conserved role in low oxygen signaling in higher plants. Moreover, parallel data mining from whole-transcriptome profiling in *Arabidopsis* and rice under low oxygen conditions (Christianson et al.,

2010; Narsai et al., 2010) and QTL analyses led to the identification and characterization of genes putatively involved in the transduction of the hypoxic signal (Xu et al., 2006) (Fukao and Bailey-Serres, 2008). In the animal world, HIF1 is a major regulator of the hypoxic pathway, leading to programmed cell death or apoptosis (Carmeliet et al., 1998). In plants, no HIF1-like gene is present, however a vast cross-species family of hypoxia-related transcription factors called Ethylene Response Factors (ERFs) has been described (Nakano et al., 2006; Licausi et al., 2010). The name derives from early studies showing their responsiveness to ethylene (Ohme-Takagi and Shinshi, 1995), and it refers also to their 50-60 amino acid domain, highly conserved across plants (Sakuma et al., 2002) (Figure 7). Transcriptional regulators belonging to the ERF family play a role in development (Boutillier et al., 2002) and reaction to biotic (Yamamoto et al., 1999) and abiotic stresses (Nakano et al., 2006). A DNA element named the GCC box (AGCCGCC) has been shown to be recognized by most members of the ERF family (Ohme-Takagi and Shinshi, 1995), although a vast number of other motifs have been recently shown to be physically bound by different ERF transcription factors (Sasaki et al., 2007) (Maeo et al., 2009).

While for *Arabidopsis* and rice much is known on transcriptional phenomena related to hypoxia, the characterization of these is still missing in potato (*Solanum tuberosum*), organism which shares the general hypoxic susceptibility of *Arabidopsis* (Table 1). Therefore, we identified in the potato tuber development another ideal scenario for gene network reverse engineering focusing on hypoxia-related expression measurements. Compared to the previous pathway (the seed coat mucilage synthesis and release), in this case no gene bait can be used for guilt-by-association analyses, since no potato ERF has been characterized yet. Therefore, it is necessary to identify the functional orthologs of *Arabidopsis* ERFs prior to the bioinformatical analysis. In particular, two of the *Arabidopsis* ERF genes, *HRE1* and *HRE2*, have been shown to be induced by hypoxia and are required for low oxygen tolerance (Licausi et al., 2010), representing the ideal baits for our pathway reconstruction analysis.

| Species | Tolerance of oxygen deficiency | Seed germination under anoxia |
|---|--------------------------------|-------------------------------|
| Thale cress (<i>Arabidopsis thaliana</i>) | Poor | No |
| Maize (<i>Zea mays</i>) | Poor to intermediate | No |
| Potato (<i>Solanum tuberosum</i>) | Poor | No |
| Rice (<i>Oryza sativa</i>) | Intermediate to strong | Yes |
| Barnyard grass (<i>Echinochloa</i> spp.) | Intermediate to strong | Yes |
| Marsh dock (<i>Rumex palustris</i>) | Strong | No |

Table 1 - Comparative hypoxia tolerance of several plant species. Readapted from (Fukao and Bailey-Serres, 2004)

1.4.3 Essential genes

Analysis on both data-inferred and annotation-based networks has also been used for identifying particular classes of genes whose presence is absolutely necessary for an organism to survive, known as "essential genes" (Jeong et al., 2001; Zotenko et al., 2008). The characterization of these genes will allow to identify the

minimal transcription machinery needed for plants to germinate (Tzafrir et al., 2004; Meinke et al., 2008) and has been described elsewhere as “the most important task of genomics-based target validation” (Chalker and Lunsford, 2002) (Cole, 2002).

However, experimentally screening for lethal gene disruptions is challenging and time consuming, even in model species. Several sequence-based properties of essential genes have been exhaustively investigated: for instance, essential genes tend to evolve more slowly than their non-essential counterparts (Hurst and Smith, 1999), and their sequence tends to be devoid of codons for rare amino-acids, possibly to minimize translational stalling (Seringhaus et al., 2006).

In addition, essential genes have peculiar topological properties in biological networks. In *Saccharomyces cerevisiae*, proteins coded by essential genes tend to have a higher Degree centrality in a protein-protein interaction network (i.e., more distinct interactors) than non-essential ones (Jeong et al., 2001). These genes show a particularly high Degree centrality also in networks extracted from transcript coexpression analysis (Mutwil et al., 2010). The propensity towards essentiality for genes having a high degree centrality also seems to hold for other centrality measures, specifically closeness and betweenness (Hahn and Kern, 2005). Some studies (Carlson et al., 2006; Wunderlich and Mirny, 2006) have combined different sequence-based, expression based and network based properties for efficient essential gene prediction.

1.5 Summary of the aims of this thesis

In the first part of the Results section of this thesis, I will focus on the measurement of transcript levels using microarrays. I will first describe a general framework for transcriptome characterization based on the information available in a joint population of NGS and EST sequences for the salt resistant plant *Thellungiella salsuginea*, in order to provide reliable transcript models for microarray platform design (Paragraph 2.1). Then, I will identify, define and partially overcome sample correlation issues in microarray normalization, and propose a novel normalization approach called transposed RMA or tRMA (Paragraph 2.2).

The second part of the Results section focuses on expression-based gene network reverse engineering approaches to better understand the transcriptional control phenomena in plants. In Paragraph 2.3, I will attempt to improve the detectability of essential genes based on expression data and a conditional correlation based approach. In Paragraph 2.4, I will describe a comparative *in silico* analysis of several gene network reverse engineering approaches based on a large *Arabidopsis thaliana* microarray dataset, focusing on the set up and application of the linear regression LASSO. In this Paragraph I also show the generation of a comprehensive network reconstruction tool to perform these kinds of analysis.

In the last part, I will transport the network reconstruction methods comparison from a purely theoretical perspective to two incompletely understood *in vivo* scenarios. In Paragraph 2.5 I will describe how LASSO and Correlation can extract complementary sets of genes involved in the *Arabidopsis thaliana* seed coat mucilage pathway, based on known gene baits affecting this process. In Paragraph 2.6 I will test LASSO and Correlation in *Solanum tuberosum* tuber development; here, I will also describe how the gene baits were identified by sequence similarity with *Arabidopsis thaliana* prior to the expression-based analysis.

2. Results

2.1 Generation of a custom microarray platform from next-generation mRNA sequencing data: the *Thellungiella salsuginea* transcriptome

2.1.1 Collecting *Thellungiella* sequences

Thellungiella salsuginea (common name: salt cress) is an extremophile plant, possessing an exceptionally high resistance to salt, drought and low temperatures. Despite these features, *Thellungiella* is morphologically and genetically very close to *Arabidopsis thaliana* (sharing 90-95% nucleotide sequence for housekeeping genes (Inan et al., 2004)). The study of *Thellungiella* in comparison to the broad knowledge available for *Arabidopsis* will possibly shed light on the mechanisms of tolerance to extreme conditions, with great potential applications to crop research. A number of *Thellungiella* expressed sequence tags (ESTs) have been generated by different studies (e.g. (Wang et al., 2004)) using standard Sanger sequencing, but presently, neither the genome nor the transcriptome of this extremophile have been fully examined.

I started by collecting all transcript sequence information publicly available for *Thellungiella salsuginea* on GenBank (Benson et al., 2008), obtaining in a collection of 38,022 sequences. This was merged with 6,529 ESTs produced by (Wang et al., 2004), for a total of 44,551 sequences. Despite this apparently high number of sequences, the internal redundancy plus the lack of a full coverage of experimental conditions make these libraries insufficient to describe the full range of *Thellungiella* transcripts.

To complete the picture, a considerable number of different salt, drought and temperature conditions were tested on a pool of *Thellungiella* plants by our collaborator Dr. Yang Ping Lee (Max Planck Institute for Molecular Plant Physiology, Golm). The polyA tagged (i.e. mRNAs) transcripts extracted from these plants were then pooled and sequenced using 454 technology (Cheung et al., 2006). Two different libraries were produced, an "unnormalized" one, where all mRNAs were pooled and sequenced without altering their relative quantities, and a "normalized" one, where the most abundant transcripts were reduced in number (Soares et al., 1994). The purpose of a normalized library, in a qualitative study like a transcriptome characterization, is to obtain the same amount of information with less sequence reads, and hopefully to detect rare transcripts.

The normalized library contains around 400,000 reads, roughly half the size of the unnormalized library. The normalized library seems to be composed by shorter fragments than the unnormalized one, hinting that a certain fragmentation was introduced during the transcript treatment (Figure 8).

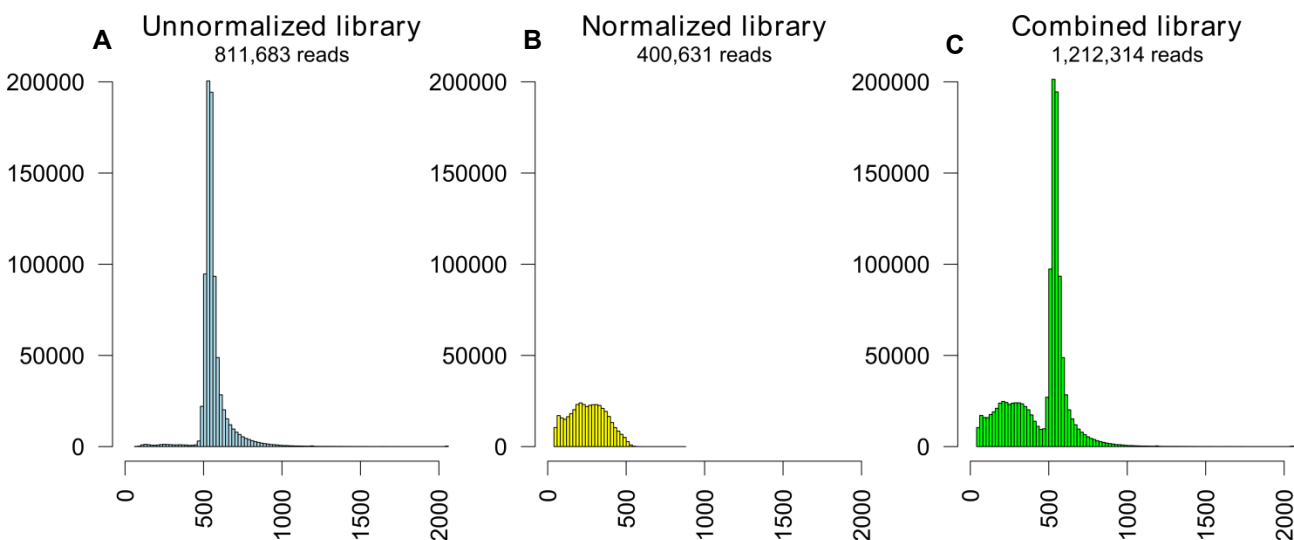


Figure 8 - Read length distribution for the unnormalized library (A), normalized library (B) and a combination of the two (C). On the x axis, the length of the reads is displayed, while on the y axis, the number of reads is indicated. The nucleotide degradation of the normalized library is expected to happen by following the normalization protocol (Soares et al., 1994).

In order to remove potential contaminating reads from the picture, I applied a BLAST-based pipeline highlighting all reads with a high identity first hit outside the *Streptophyta* phylum (see Paragraph 4.1 for details), resulting in both libraries containing only a very low fraction (0.6%) of contaminant reads. Only 10% of these reads belong to "expected" contaminant organisms, such as the plant pathogen *Acyrtosiphon pisum* or *Agrobacterium*, and most of them actually map to a plethora of species from all kingdoms of life. Therefore, since we cannot discard the possibility that these reads are indeed arising from peculiar *Thellungiella* transcripts, that only by chance have a high similarity with non-*Streptophyta* organisms, we included them in the subsequent analysis, and marked as "putative contaminants" the 304 transcript models generated through these reads.

2.1.2 Transcript assembly

However longer than reads obtained by alternative next-gen sequencing methods, 454 reads are generally shorter (<1000 nucleotides) than most of the mRNAs present in the Transcriptome. For example, the average transcript length for *Arabidopsis thaliana* is around 1200 nucleotides (see Appendix, page 101). Therefore, raw reads need to be processed by a transcript *assembly* pipeline, which removes ultralow frequency sequences (most likely sequencing errors) and merges identically overlapping reads into "contigs" (representing transcript models). To do so, I merged the two libraries with the ESTs publicly available for *Thellungiella* and proceeded testing different assembly strategies among those capable to integrate different input data. The first, called MIRA (Chevreux, 2005) implements a modified greedy method (Miller et al., 2010) and is optimized for hybrid input assemblies. The second, included in the CLC Workbench (www.clcbio.com), is a proprietary, commercial assembler optimized for speed. The third, iAssembler (bioinfo.bti.cornell.edu/tool/iAssembler/), is a hybrid method which combines the output of MIRA and CAP3 (Huang and Madan, 1999) pipelines to reduce assembly errors. It must be noted that the commonly accepted favoured method for 454 assembly, the

commercial Roche assembler Newbler (Margulies et al., 2005), was not available for testing. Unfortunately, iAssembler couldn't provide any analyzable output in a reasonable amount of time, and we therefore decided to discard this method.

| Assembler | Assembled reads | Contigs | Largest contig | Average contig length | N50 | Arabidopsis peptides found | Assembly time |
|------------------------------|----------------------|---------|----------------|-----------------------|-----|----------------------------|---------------|
| MIRA 3.1.15 | 1,060,666 (84.4%) | 46,220 | 5,736 | 567.4 | 646 | 23,029 (68.9%) | 26h |
| CLC Workbench 4.0 | 1,079,855 (85.9%) | 39,438 | 2,965 | 545.3 | 625 | 22,133 (66.2%) | 0.5h |
| iAssembler 1.0 | | | | | | | >600h |

Table 2 - Comparison of assembly pipelines on a *Thellungiella salsuginea* collection of 1,212,314 sequences.

MIRA uses slightly less reads than CLC (Table 2), with both methods discarding around 15% of the combined 454/EST data. A manual inspection of the discarded reads shows a high amount of unique and low quality sequences. MIRA assembles the reads in more "contigs", or transcript models (46,220) than CLC (39,438). The MIRA contigs tend to be generally larger (average length for MIRA is 567, for CLC 545) with a lower amount of short fragments, as assessed by the N50 parameter (646 for MIRA, 625 for CLC). The N50 is a widely used statistical instrument for assembly validation and it indicates, in a contig population, the contig size above which 50% of the total sequence nucleotides are contained. In this case, an N50 of 646 for MIRA means that half of the nucleotides assembled are contained in contigs larger than 646, and this indicates that the MIRA transcript models are slightly more "realistic" than CLC ones. About the transcript "completeness" of a contig population, it is in theory impossible to assess it while lacking information on the genome of the studied organism, like in the case of *Thellungiella*. However the highly curated transcriptome of a close relative (i.e. *Arabidopsis thaliana*) is available, and therefore we used this as a test for the degree of completion of the transcript populations obtained through the assembly processes. The coverage of *Arabidopsis thaliana* proteome was performed via BLASTX with a mild E-value threshold (10^{-10}), using TAIR9 (Swarbreck et al., 2008) peptides (33,410). I decided to use peptides since sequence divergence is more strictly controlled than nucleotides (due to the degeneration of the codon code), however similar conclusions can be drawn using *Arabidopsis* cDNA sequences (data not shown). MIRA seems to perform a little better than CLC, providing a set of contigs (i.e. transcript models) which achieves a coverage of around 69% of the *Arabidopsis thaliana* transcriptome, compared to 66% of CLC. All considered, the output of this comparative analysis (Table 2) shows a slightly better performance for MIRA, when compared to CLC.

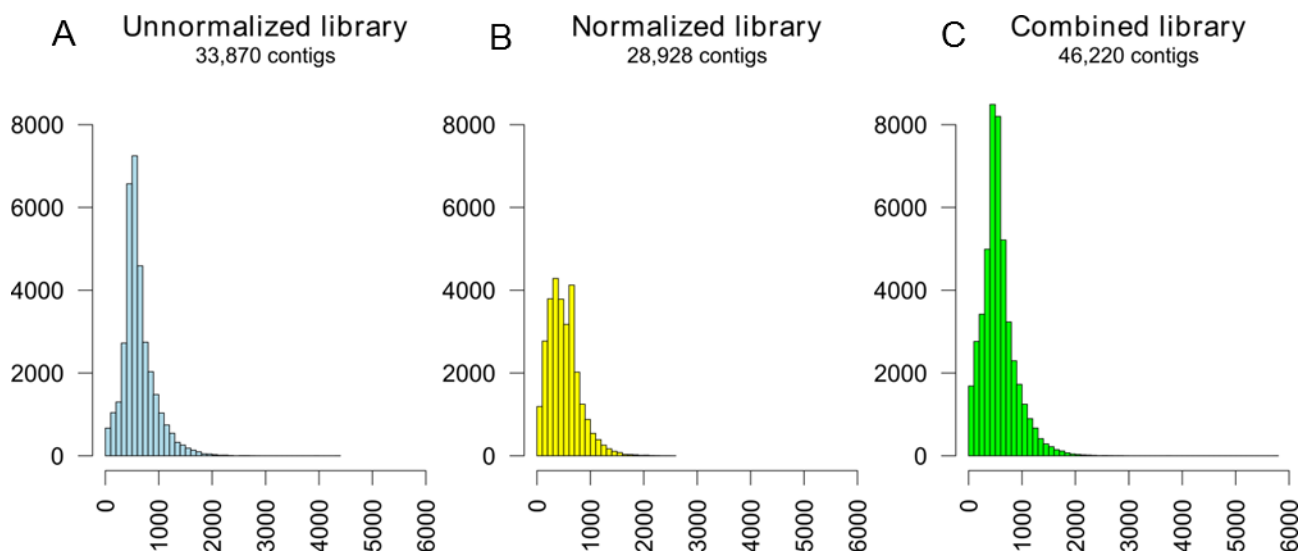


Figure 9 - Contig length distribution for the unnormalized library (A), normalized library (B) and a combination of the two (C). On the x axis, the length of the contigs is displayed, while on the y axis, the number of contigs is indicated.

After selecting the assembly method, in this case MIRA, further considerations can be obtained over the previously discussed read libraries, through the analysis of the assemblies obtained by the separate collections. Using the normalized library I achieved a similar number of contigs (normalized: 28,928 vs. unnormalized: 33,870), while keeping a slightly lower coverage of the *Arabidopsis* peptide-coding transcriptome (63.7% vs. 66.7%). Combining both libraries I obtained a high number of contigs (46,220) and an even higher coverage of the *Arabidopsis* protein-coding transcriptome (68.9%). The N50 was again comparable for both libraries. The contig length doesn't seem to be extremely affected by the normalized library read degradation described in Figure 8, although it shows a lower general length (Table 3 and Figure 9). While larger, as expected, than the raw reads, the population of contigs extracted by the assembly process is still showing a higher fragmentation than the one observed in the annotated collection of cDNAs for *Arabidopsis thaliana* (see Appendix, page 101). In order to obtain a transcript model quality assessment independent from *Arabidopsis* I checked the contigs for open reading frames (ORF) presence using the ORFpredictor with default parameters (Min et al., 2005). This pipeline marks as non-ORF containing those contigs falling in none of the 10 mRNA categories described in (Min et al., 2005) in any of the 6 reading frames, and are therefore to be considered not protein coding. All libraries yield a high density of potential ORF-containing transcript models (99.3% for the unnormalized one, 98.2% for the normalized one and 98.6% for the merged libraries).

Everything considered, with half the reads of the unnormalized library and a consistent read degradation, and consequently a much lower contig coverage, the normalized library yielded contigs of around the same quality and with a similar completeness as the unnormalized one.

| Library | Reads | Contaminant reads* | Assembled reads | Contigs | Contigs with ORFs§ | Largest contig |
|----------------------------------|-----------|--------------------|----------------------|---------|--------------------|----------------|
| Normalized | 400,631 | 2506 (0.6%) | 376,509 (84.6%) | 28,928 | 28416 (98.2%) | 2,537 |
| Unnormalized | 811,683 | 4496 (0.6%) | 712,262 (83.2%) | 33,870 | 33625 (99.3%) | 4,347 |
| Normalized + Unnormalized | 1,212,314 | 7002 (0.6%) | 1,060,666 (84.4%) | 46,220 | 45583 (98.6%) | 5,736 |

*Best hit E-value 10^{-10}, non *Streptophyta* sequences (NCBI nr nt 12-02-2010) (Pruitt et al., 2006)
§ Also partial ORFs (Min et al., 2005)

| Library | Average contig coverage | Average contig length | N50 | Contigs generated by 454 only | Arabidopsis peptides found# |
|----------------------------------|-------------------------|-----------------------|-----|-------------------------------|-----------------------------|
| Normalized | 6.74 | 502.4 | 632 | 16,949 (58.6%) | 21,270 (63.7%) |
| Unnormalized | 11.99 | 620.8 | 665 | 21,662 (64.0%) | 22,282 (66.7%) |
| Normalized + Unnormalized | 12.63 | 567.4 | 646 | 33,147 (71.7%) | 23,029 (68.9%) |

Any BLASTX hit E-value 10^{-10}, reference: TAIR9 (Swarbreck et al., 2008) peptides (33,410)

Table 3 - Transcript libraries and assembly statistics. The assemblies also include 44,551 EST sequences from (Wang et al., 2004) and (Wong et al., 2005)

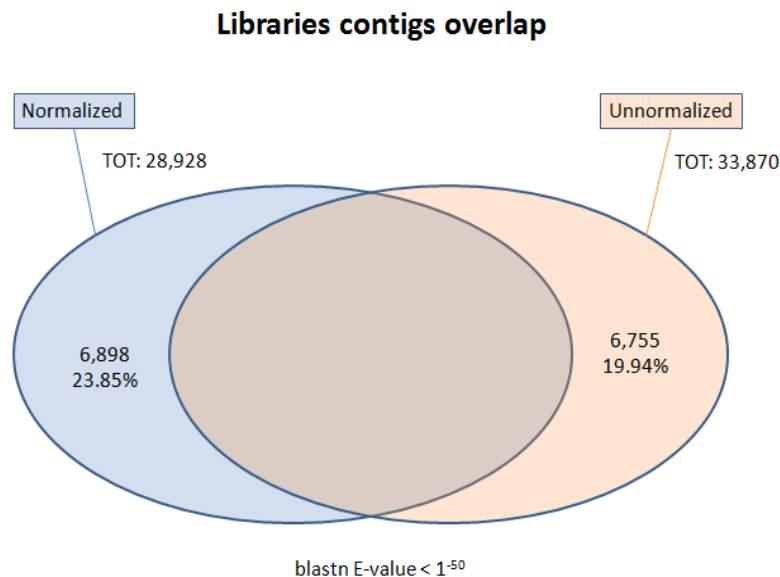


Figure 10 - Overlapping between transcript models generated with the unnormalized and normalized libraries

The libraries, surprisingly enough (since they derive from the same sample collection), show a certain number of unique transcripts (Figure 10), which justifies our decision to combine them for designing a comprehensive microarray platform). It can be noted that 454 reads are essential to build more than half of the final contigs: 71.7% of the final transcript representatives are assembled from this technology only in the combined library.

2.1.3 Transcriptome completion

The heterozygous status of the *Thellungiella* plants used for the library generation produced a high population of highly identical contigs in the final outputs of the assembly processes. For example the MIRA assembly based on all ESTs and 454 reads (Table 3) contains 4,020 contigs (out of 46,220) with at least another contig sharing more than 99% sequence identity and coverage. In order to reduce this nearly-identical transcript model group, and to fit the requirements for the Agilent oligoDNA microarray (www.agilent.com) based on these models (44,000 probes) I condensed these 4,020 contigs into 610 clusters of similarity. These clusters corresponded to multiple sequence alignments that were folded into partially degenerated IUPAC sequences and the microarray probes (~60mer) were specifically designed to avoid non-identical regions. In the end, the chip representative sequences will be composed by 42,220 unique target probes, and 610 multiple target probes, for a total of 42,810. Finally, since the orientation of the reads was not necessarily kept during 454 sequencing, the 42,810 probes were kept or reverse complemented given the most likely orientation (assessed through the orientation of the best hit to nr database (Pruitt et al., 2006) or, where not available, orientation with the longest predicted ORF (Min et al., 2005)). The pipeline is described in more detail in the Methods section of this thesis, Paragraph 4.1.

Thellungiella contigs – Arabidopsis transcripts overlap

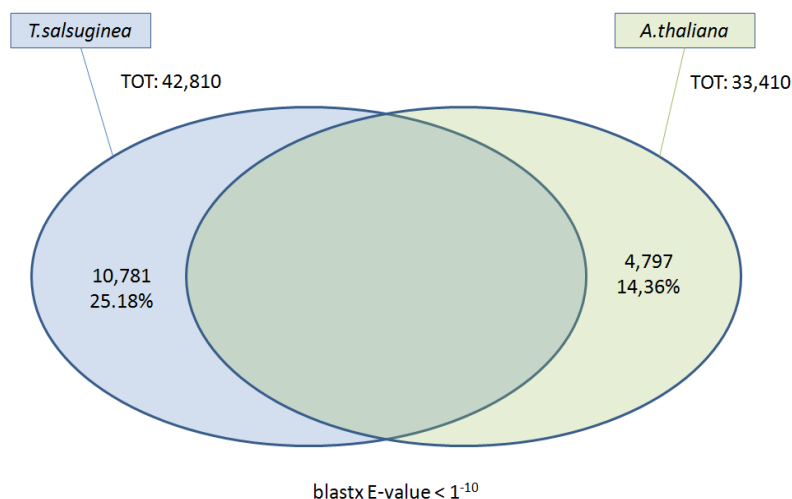


Figure 11 - Assessment of unique transcript models found in *Arabidopsis thaliana* (TAIR9, cds) and *Thellungiella salsuginea* (assembly made on the joint read library)

2.1.4 Comparative transcriptome considerations between *Thellungiella* and *Arabidopsis*

Comparing in more detail the transcripts of *Arabidopsis* with the *Thellungiella* MIRA models it can be observed that the two organisms possess apparently 14% and 25% of specific transcripts, respectively (Figure 11). While assessing this, it is important to bear in mind that some transcripts for *Thellungiella* could still be missing from our assembly, but at the same time that the Transcriptome for *Arabidopsis* is nearly fully characterized (Swarbreck et al., 2008). In order to understand the global biological characteristics of the *Thellungiella*

transcript model population, I performed a sequence-based annotation of the combined library contigs using Mercator (Lohse and Usadel, unpublished) and the MapMan ontology annotation (Usadel et al., 2009) (Figure 12). In comparison with *Arabidopsis thaliana* transcriptome, *Thellungiella* shows a certain functional similarity. Contrarily to what one could expect, *Thellungiella* (at least in what can be detected) possesses a lower fraction of stress-related transcript models (Figure 12), therefore it can be hypothesized that its resistance shouldn't be connected to a mere higher number of different stress-responding proteins. An exception are the "Late Embryogenesis Abundant" (LEA) proteins, a family of rather enigmatic proteins necessary for dehydration stress response in *Arabidopsis thaliana* (Hundertmark and Hincha, 2008). While *Arabidopsis thaliana* contains 51 expressed LEA genes, *Thellungiella salsuginea* seems to possess almost three times this amount, having 148 distinct LEA transcript models (BLASTX similarity threshold 10^{-10}). This high number, despite the merging of highly identical transcripts applied to the contig population, is however still partially an over-estimation deriving by different splicing variants considered as different genes. Finally, the relative higher fraction of "Not assigned" transcript models of *Thellungiella* in comparison with *Arabidopsis*, is partly expected due to a genetic drift between the two species (Inan et al., 2004), generating novel and rather dissimilar transcripts that will go beyond the detection capability of the Mercator pipeline (Lohse and Usadel, unpublished).

The pipeline described in the Methods section of this thesis (Paragraph 4.1), whose results have been discussed above, has been used to generate a novel Agilent 44k microarray (www.agilent.com) specific for *Thellungiella salsuginea*. At the time of writing this thesis, the expression pattern of several experimental stress conditions is being investigated using this microarray by Dr. Yang-Ping Lee (Max Planck Institute for Molecular Plant Physiology, Golm). It can be expected that the characterization of specific transcriptional regulation mechanisms in *Thellungiella* will give the scientific community the opportunity to better understand stress responses and provide insights into the capability of this plant to survive in extreme conditions.

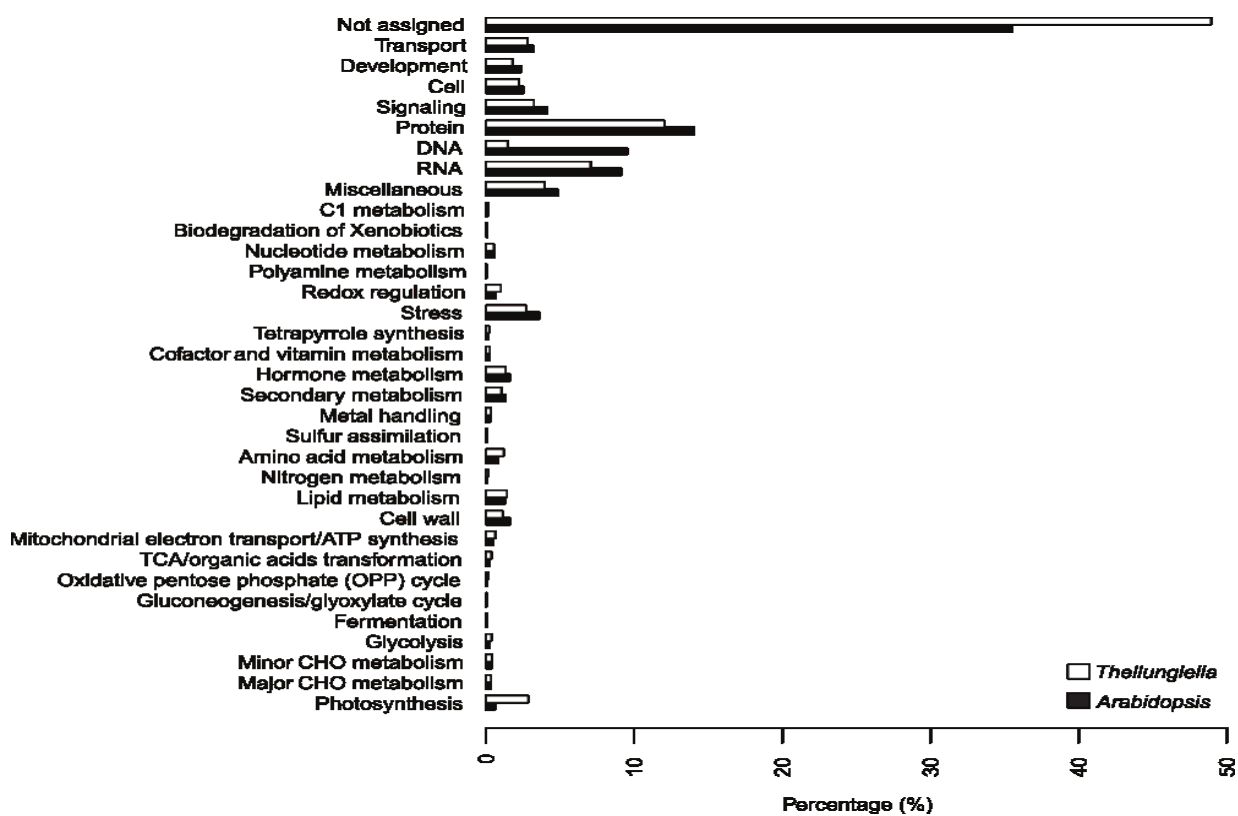


Figure 12 - MapMan functional classes in the *Arabidopsis* and *Thellungiella* transcriptomes

2.2 Algorithm-driven Artifacts in median polish summarization of microarray data: tRMA

The design of reliable probe sequence for a microarray platform still require the raw data to be measured and preprocessed in order to be analyzed by e.g. differential gene expression statistics, sample clustering, transcript correlation or global gene network reconstruction. In this section, I extend the analysis performed by (Lim et al., 2007), aiming to shed more light on the behavior of multi-array techniques specifically in the context of inter-array correlation. I will describe the characteristics of probesets which induce these artifacts and provide both a mathematical and a biological explanation for the phenomenon. Finally, I introduce a slightly changed version of the RMA code which massively reduces inter-array correlation artifacts, while retaining RMA features in the context of differential gene expression analysis.

2.2.1 Multi-array preprocessing effects

In order to compare the behavior of three of the most popular microarray preprocessing techniques (MAS5, RMA and GCRMA), Lim and colleagues (Lim et al., 2007), tested these on a single dataset of 10 microarrays hybridized with human samples. I extended this analysis on a considerably larger *Arabidopsis thaliana* dataset comprising 3707 microarrays, selecting different sample sizes (see Paragraph 2.2.2), according to the realistic size of a single experiment dataset (2 to 100 samples). First, I calculated inter-array correlations on randomly selected groups of these microarrays (Figure 13A). The plots show us a high correlation (>0.7) between samples, indicating that many genes' relative expression will remain constant across different treatments and genotypes, and therefore showing a certain robustness of *Arabidopsis*' genetic machinery in varying environmental conditions and other perturbations (e.g. gene knock-outs). The sample size doesn't seem to influence the high correlation between arrays, although some evident oscillations could be detected for RMA and GCRMA at lower sample sizes. The comparison of the three preprocessing methods shows that RMA and GCRMA yield somewhat more similar microarray expression values than the Affymetrix algorithm MAS5.

In order to compare real data with a null dataset, I analyzed the behavior of the three preprocessing techniques on permuted arrays (see Materials and Methods, Paragraph 4.2.3). Since permuted arrays are entirely shuffled and uninformative with respect to probeset expression, I expected them to be, on average, not correlated at all. However as previously reported (Lim et al., 2007), this is not the case for some of the techniques I used (Figure 13B). RMA and GCRMA show a high mean inter-array correlation, which is decreasing with the sample size and, interestingly much higher for odd sample sizes, and reminiscent of the oscillating behavior in real arrays (Figure 13A). Average values for Figure 13 are shown in Table 15 (Appendix, page 102). Although the higher (GC)RMA-derived inter-array correlation was known in literature (Lim et al., 2007), these results show for the first time an RMA/GCRMA effect related to sample size. In order to assess if these artifacts were due to the choice of correlation coefficient I repeated our analysis using Pearson's and Lin's correlation, but obtained nearly identical results (data not shown). MAS5 alone shows the expected no-correlation behavior. It must be noted that, unlike the other two techniques, MAS5 uses a single-array summarization technique (a robust Tukey-biweight average of the probe values) which treats each sample separately.

I will focus on the cause of this behavior observed when using RMA and GCRMA, trying to understand the mathematical and experimental scenarios that could introduce such a massive artificial inter-array correlation for these two methods.

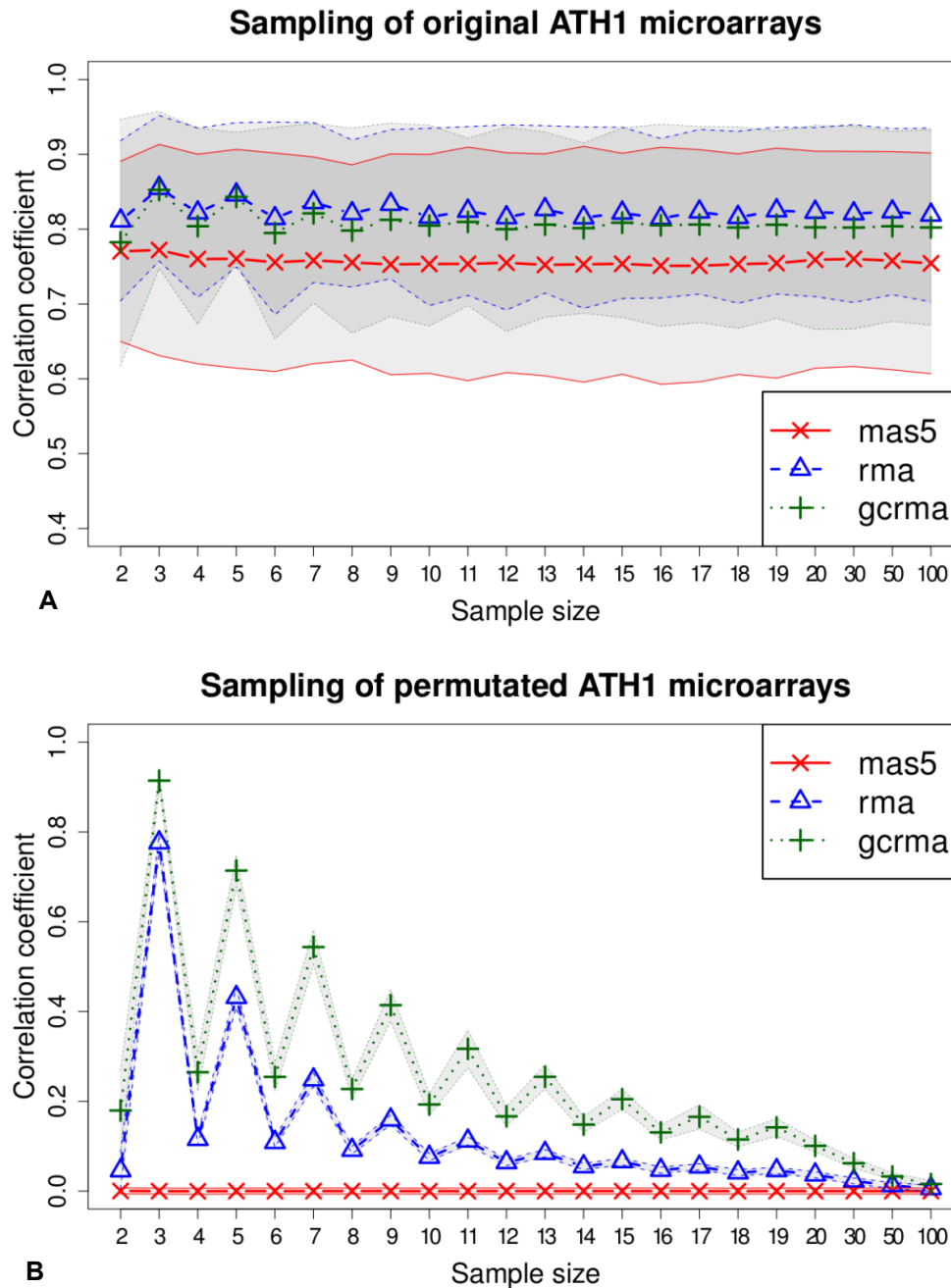


Figure 13 - Inter-array similarity calculated on the Arabidopsis dataset normalized by RMA, GCRMA and MAS5. 1000 groups of arrays for each sample size were selected, and then the averages and standard deviations of inter-array spearman correlation coefficients were calculated. The averages are reproduced as symbols which are connected by a broken line and averages plus minus one standard deviation are shown as shaded areas bordered by a solid line of the same color. Values for MAS5 are shown in red, RMA in blue and GCRMA in green. A) real samples. B) samples with their raw signal intensities internally permuted.

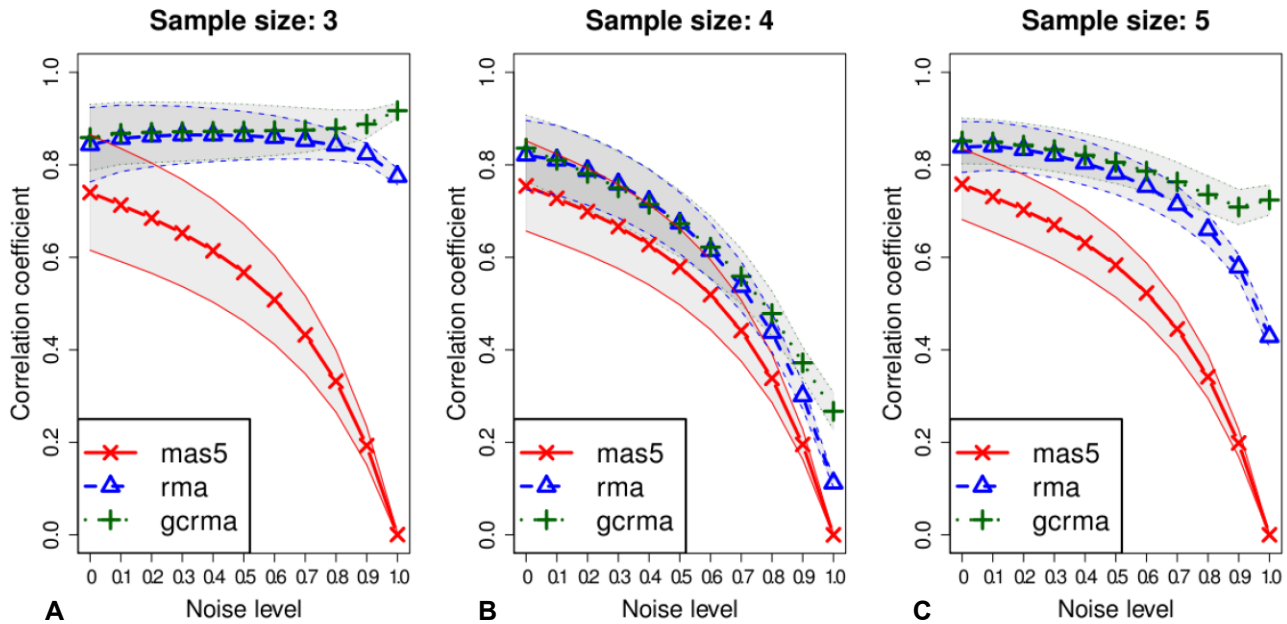


Figure 14 - MAS5, RMA and GCRMA behavior on inter-array correlation for samples of 3,4 and 5 arrays upon incremental noise addition by adding values from real arrays to values from permuted arrays before normalization. The “noise level” gives the fraction of the values that came from the permuted arrays (For more details see Paragraph 2.2.5). The figure shows means and standard deviations, as in figure one using three different sample sizes, an even number, 4 (panel B), and two odd numbers, 3 (panel A) and 5 (panel C).

2.2.2 Causes of RMA and GCRMA artifact generation

So far, it has been described that the introduction of artificial similarities between arrays by RMA and GCRMA is particularly strong for small and odd sample sizes. In Figure 14 I show how adding an increasing amount of noise to microarray samples in the Arabidopsis dataset (see Material and methods) results in the expected loss-of-correlation behavior for MAS5, GCRMA and RMA for an even sample size (Figure 14B). However, for sample size 3 (Figure 14A), RMA and GCRMA actually *add* inter-array correlation as noise is combined with the biological signal. The situation is still atypical for the next odd sample size (5 samples, Figure 14C).

Returning to our original Arabidopsis dataset, I observed that many probesets seem to yield completely identical values across different samples when processed by RMA or GCRMA. Datasets of three arrays normalized by RMA and GCRMA show, respectively, around 20% and 12% of the probesets population with identical values across all samples.

The effect will decrease with increasing sample size (see Figure 13B) as previously reported in (Usadel et al., 2009). I therefore measured the tendency to yield identical expression estimates for any particular probeset after RMA normalization (ID tendency) in 10,000 3-samples datasets of original microarrays and compared it to several probeset characteristics.

This ID tendency is inversely correlated (Spearman correlation coefficient = -0.624) to the probeset internal consistency (Figure 15), which I measured using the fit of the probeset to a linear model that measures concordance between probes.

Inconsistent probesets tend to yield identical results

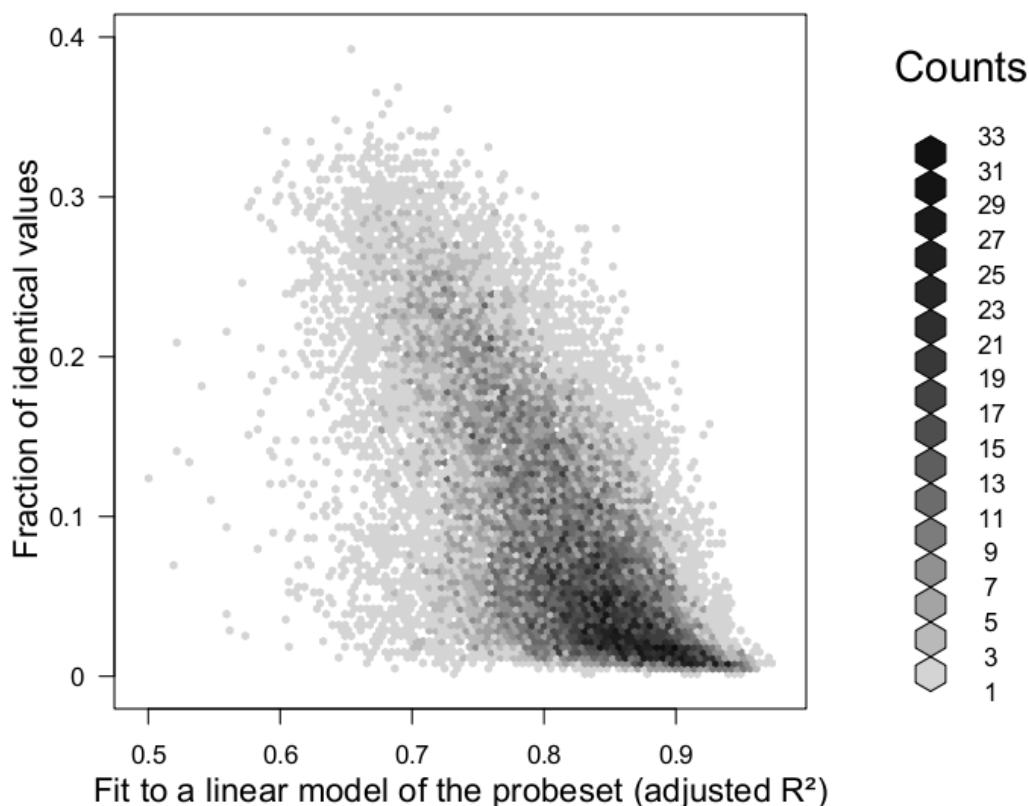


Figure 15 - Inverse correlation between probeset tendency to yield identical expression values and internal probeset consistency, measured as a probe linear model R^2 . The x axis shows a fit to a linear model of any given probeset across 3707 *Arabidopsis* microarrays, using probeset sample means as explanatory variable. On the y axis the fraction of 3 sample subsets yielding 3 identical arrays for a given probeset is shown (10000 randomly picked groups were selected).

This phenomenon is also particularly evident for lowly expressed probesets (Figure 16A) and those hybridizing to multiple targets (Figure 16C), especially if the different targets fall into different biological classes (Figure 16B). This hints that particularly "dirty" probesets, sharing different and diversified targets, are more affected by the normalization artifact effect than "clean" ones, having a strong, mono-target expression. As the problem of "dirty" probesets has been discussed before and been tackled by providing updated probeset definitions in the customCDF project (Dai et al., 2005), I assessed whether the oscillating behavior for real data was still observable when using such an updated definition. However, almost identical results were obtained using such an updated probeset annotation (data not shown).

In summary, RMA and GCRMA tend to yield identical values for probesets containing probes that yield grossly different measurements across samples, and therefore are either noise-driven or have multiple independent targets. Taken together, these results tell us that these normalization methods introduce artificial correlation across microarrays driven by lowly expressed, internally inconsistent, multi-target and/or multi-function probesets.

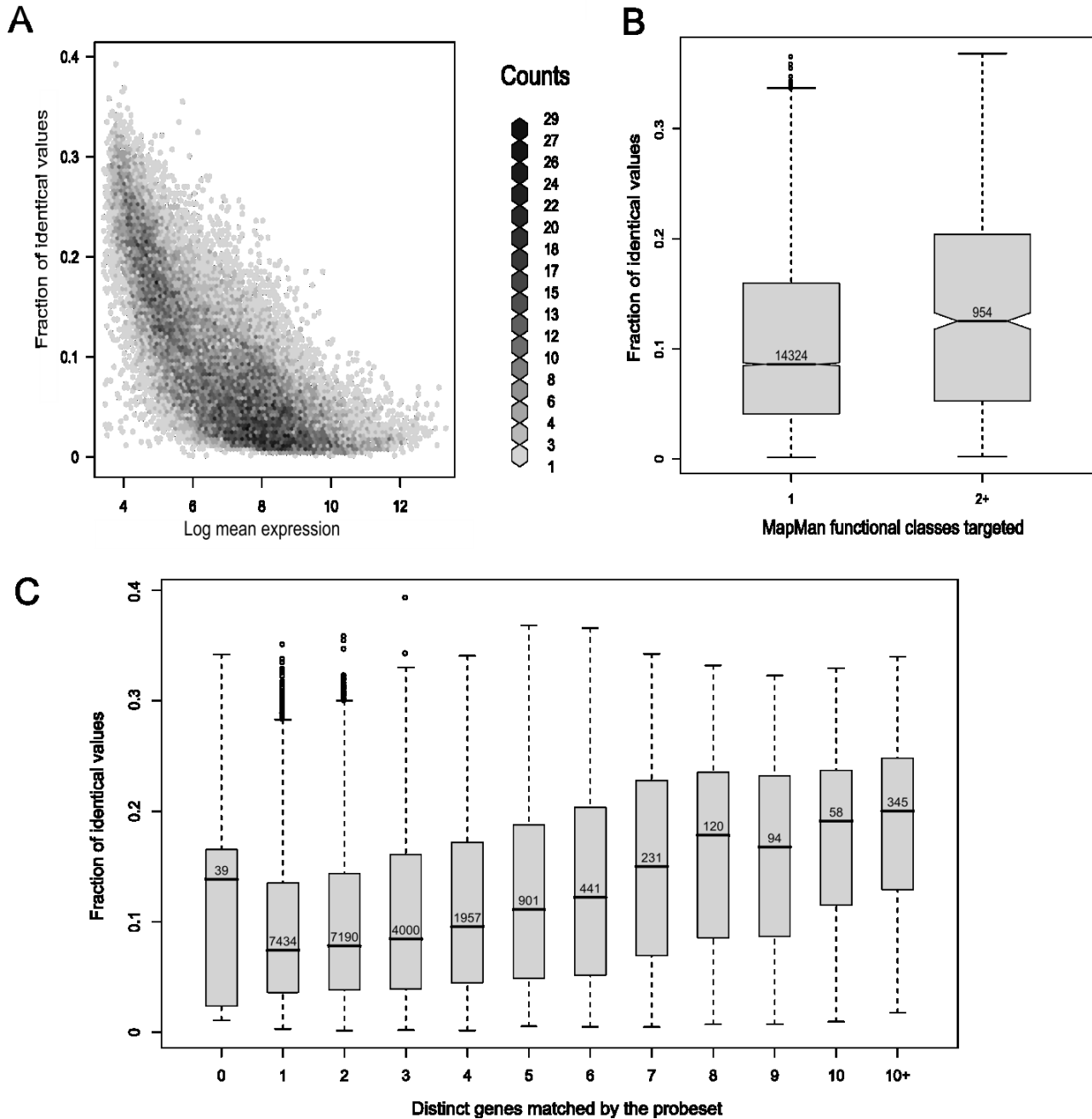


Figure 16 - Correlation between same-value artifactual behavior and probesets' features. In panel A, inverse correlation between probeset tendency to yield identical expression values and mean probeset expression. In panel B, identical arrays output for multi-target probesets matching only one (left) or multiple (right) MapMan functional classes (Usadel et al., 2009). In panel C, positive correlation between the number of distinct targets hybridized by a probeset and the tendency of a probeset to yield identical expression values across arrays. This tendency is calculated as the fraction of RMA normalized subsets of 3 arrays yielding 3 identical results for the given probeset. Within each boxplot the number of probesets in the category is indicated

2.2.3 Median polish inconsistency

RMA (Irizarry et al., 2003) and the closely related method, GCRMA (Wu and Irizarry, 2005), differ only in the initial background correction step. Since the same effect is present in both methods (Figure 13B), I reasoned

that the artifact generation should depend on either the shared normalization step (which is quantile normalization in both cases (Bolstad et al., 2003)), or on the probe summarization step (median polish (Tukey, 1977)). I concluded that the effect cannot arise from quantile normalization, since substituting it by scale normalization or removing it completely yields qualitatively identical plots whereas the inclusion of a median polish step always introduced the effect, regardless of background correction and normalization procedures (a full collection of this investigation is available in (Giorgi et al., 2010), Supplementary File S4). The shared artifact can therefore only be generated within the median polish summarization step. Indeed, substituting the median polish step with any other alternative summarization procedure available in the BioConductor RMA implementation eliminates the artificial inter-array correlation effect (Figure 17). In order to compare the RMA default summarization with another multiple-array summarization (see Paragraph 1.2), I substituted the median polish method with the robust least squares linear model summarization, described by (Irizarry et al., 2003), and showed that this procedure almost completely removes inter-array correlation (magenta dashed line in Figure 17). On the other hand, as an example of single-array summarization, I used RMA with an "average of log" summarization, which simply computes the average of the logarithms of probe intensities for every probeset. This single-array summarization yields a predictable 0 correlation among all arrays (orange dashed line in Figure 17).

Sampling of permuted ATH1 microarrays

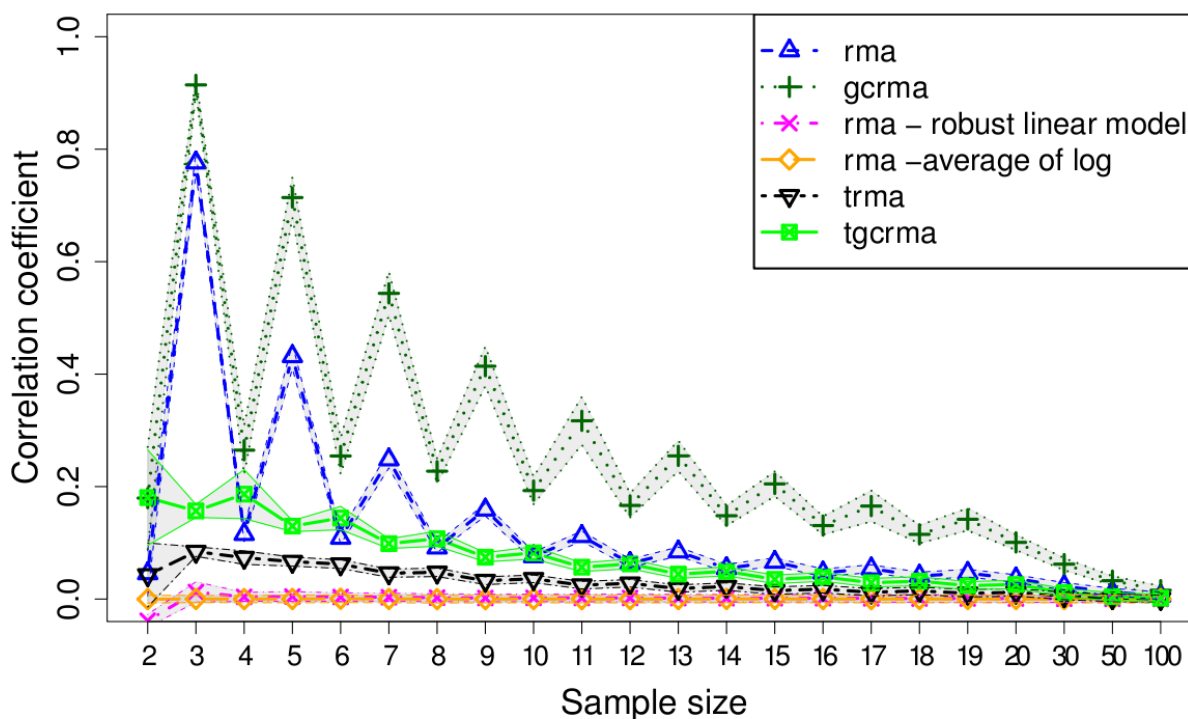


Figure 17 - Comparison of alternative RMA summarization steps on permuted datasets. The original median polish summarization step is plotted with BioConductor alternatives and the transposed median polish of the tRMA method. 1000 groups of arrays for each sample size were selected, and then the averages and standard deviations of inter-array spearman correlation coefficients were calculated and plotted as in Figure 13.

To identify why this artifact arises during the median polish procedure, I investigated the algorithm further. RMA and GCRMA apply median polish by creating a matrix from the measured values within each probeset, placing probes along each row, and samples along each column. The medians are subtracted from the intensities to cumulate residuals in each step and the grand effect (or median of medians) is subtracted from medians to cumulate probe effects in each step (Figure 18). This algorithm is more likely to introduce identical values with odd and small sample sizes, like the one depicted in Figure 18. In such a case, the row medians will fall on a specific value and be transformed to zero during the first row sweep (Figure 18B, top panel), this will increase the chance to have a zero as column median during the column sweep (Figure 18C, top panel).

Overall, the RMA implementation of the median polish algorithm shrinks all values in the probeset matrix to similar or identical values, with a stronger effect for samples (i.e. microarrays), since it starts subtracting probe (row) medians. In the example of Figure 18, the final sample values will be calculated by adding the grand effect to each column effect, and will therefore be equal to 8 for all samples.

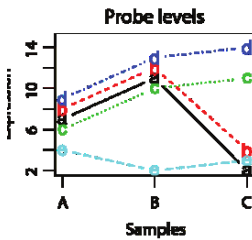
It could be argued that the median polish summarization step could be helpful in the context of Differential Gene Expression analysis, since it will flatten unclear probeset matrices and therefore highlight strong signals. However, the result of generating completely identical expression values across arrays is not always beneficial (Lim et al., 2007). Moreover, this effect can be dramatically reduced by swapping the order of row/column median subtraction within the median polish summarization step, or equivalently, by transposing the matrix created for each probeset, placing samples on rows and probes on columns. This alteration will introduce a presumably harmless similarity between probes within a probeset (which are assumed to be measuring the same quantity, and which don't form part of the output) while massively reducing the artificial sample identity.

To confirm this, I re-implemented the median polish summarization by inverting the order of the two sweep steps (Figure 18), in what I call "transposed RMA" or tRMA. As shown in Figure 17, the inversion of median subtraction steps alone reduces the median polish effect to a very small residual inter-array correlation. This can be explained by the fact that the likelihood for the sample effects to give a zero value in tRMA is very low during the first iteration, as it would require perfectly identical medians of raw probe values (Figure 18). Effectively, tRMA transfers the artifact of inter-correlation between sample to an inter-correlation between probes (in a common probeset), which might be more plausible biologically (as all probes in a probeset should measure the same target) and remains contained within the procedure and not yielded as output of the preprocessing method.

Figure 18 - Detail of the (t)RMA median polish procedures for a single probeset

A. Probeset Matrix.
Organize probes belonging to the same probeset in a matrix

| | | Samples | | |
|---|---|---------|----|--|
| | A | B | C | |
| a | 7 | 11 | 2 | |
| b | 8 | 12 | 4 | |
| c | 6 | 10 | 11 | |
| d | 9 | 13 | 14 | |
| e | 4 | 2 | 3 | |



RMA

B. Row sweep.
For each row, medians are calculated.

| | A | B | C | |
|---|---|----|----|----|
| a | 7 | 11 | 2 | 7 |
| b | 8 | 12 | 4 | 8 |
| c | 6 | 10 | 11 | 10 |
| d | 9 | 13 | 14 | 13 |
| e | 4 | 2 | 3 | 3 |

Then, medians are subtracted from each row.

| | A | B | C | |
|---|----|----|----|----|
| a | 0 | 4 | -5 | 7 |
| b | 0 | 4 | -4 | 8 |
| c | -4 | 0 | 1 | 10 |
| d | -4 | 0 | 1 | 13 |
| e | 1 | -1 | 0 | 3 |

The grand effect is calculated by adding to the previous grand effect (0 if it's the first iteration) the median of row medians.

$$\text{Grand effect} = 0 + 8 = 8$$

Median of row medians is then subtracted from row medians and added to previous row of effects (if present) to generate new row of effects.

| | |
|----|----|
| 7 | -1 |
| 8 | 0 |
| 10 | 2 |
| 13 | 5 |
| 3 | -5 |

Row of effects

C. Column sweep.
For each column, medians are calculated.

| | A | B | C |
|---|----|----|----|
| a | 0 | 4 | -5 |
| b | 0 | 4 | -4 |
| c | -4 | 0 | 1 |
| d | -4 | 0 | 1 |
| e | 1 | -1 | 0 |

Then, medians are subtracted from each column.

| | A | B | C |
|---|----|----|----|
| a | 0 | 4 | -5 |
| b | 0 | 4 | -4 |
| c | -4 | 0 | 1 |
| d | -4 | 0 | 1 |
| e | 1 | -1 | 0 |

The grand effect is calculated by adding to the previous grand effect the median of column medians.

$$\text{Grand effect} = 8 + 0 = 8$$

Median of column medians is then subtracted from column medians and added to previous column of effects (if present) to generate new column of effects.

| | | |
|---|---|---|
| 0 | 0 | 0 |
|---|---|---|

-0

| | | |
|---|---|---|
| 0 | 0 | 0 |
|---|---|---|

Column of effects

D. Iterate until convergence.

Steps B and C are repeated until all column medians fall to zero (in the present case, this happens during the first iteration). RMA will yield as output for each sample the final column of effects plus the grand effect.

Final results - RMA

Grand effect = 8
Row of effects = -1, 0, 2, 5, -5
Column of effects = 0, 0, 0
Sample values = 8, 8, 8

tRMA

B. Column sweep.
For each column, medians are calculated.

| | A | B | C |
|---|---|----|----|
| a | 7 | 11 | 2 |
| b | 8 | 12 | 4 |
| c | 6 | 10 | 11 |
| d | 9 | 13 | 14 |
| e | 4 | 2 | 3 |

Then, medians are subtracted from each column.

| | A | B | C |
|---|----|----|----|
| a | 0 | 0 | -2 |
| b | 1 | 1 | 0 |
| c | -1 | -1 | 7 |
| d | 2 | 2 | 10 |
| e | -3 | -9 | -1 |

The grand effect is calculated by adding to the previous grand effect (0 if it's the first iteration) the median of column medians.

$$\text{Grand effect} = 0 + 7 = 7$$

Median of column medians is then subtracted from column medians and added to previous column of effects (if present) to generate new column of effects.

| | | |
|---|----|---|
| 7 | 11 | 4 |
|---|----|---|

-7

| | | |
|---|---|----|
| 0 | 4 | -3 |
|---|---|----|

Column of effects

C. Row sweep.
For each row, medians are calculated.

| | A | B | C | |
|---|----|----|----|----|
| a | 0 | 0 | -2 | 0 |
| b | 1 | 1 | 0 | 1 |
| c | -1 | -1 | 7 | -1 |
| d | 2 | 2 | 10 | 2 |
| e | -3 | -9 | -1 | -3 |

Then, medians are subtracted from each row.

| | A | B | C | |
|---|---|----|----|----|
| a | 0 | 0 | -4 | 0 |
| b | 0 | 0 | -3 | 1 |
| c | 0 | 0 | 6 | -1 |
| d | 0 | 0 | 6 | 2 |
| e | 0 | -6 | 0 | -3 |

The grand effect is calculated by adding to the previous grand effect the median of row medians.

$$\text{Grand effect} = 7 + 0 = 7$$

Median of row medians is then subtracted from row medians and added to previous row of effects (if present) to generate new row of effects.

| | |
|----|----|
| 0 | 0 |
| 1 | 1 |
| -1 | -1 |
| 2 | 2 |
| -3 | -3 |

-0

| | |
|----|----|
| 0 | 0 |
| 1 | 1 |
| -1 | -1 |
| 2 | 2 |
| -3 | -3 |

Row of effects

D. Iterate until convergence.

Steps B and C are repeated until all row medians fall to zero (in the present case, this happens during the second iteration). tRMA will yield as output for each sample the final column of effects plus the grand effect.

Final results - tRMA

Grand effect = 7
Row of effects = 0, 1, -1, 2, -3
Column of effects = 0, 4, -1
Sample values = 7, 11, 6

The inter-array artificial correlation effect introduced by the median polish step is increased in GCRMA (Figure 17, dark green dotted line). As previously discussed by (Lim et al., 2007), GCRMA contains an independent problem in its background correction step, that adjusts probe intensity values through gene-specific binding. This introduces artificial inter-array correlations between probes with similar binding affinity, and therefore strengthens the effect of the following summarization step. However, substituting the median polish step with our transposed alternative “tGCRMA” (Figure 17, dark green line), massively minimizes the inter-array correlation between permuted samples.

| | MAS5 | RMA | tRMA | Best possible |
|-----------------------|-------------|-------------|-------------|----------------------|
| Signal detect slope | 0.71 | 0.63 | 0.63 | 1 |
| Signal detect R2 | 0.86 | 0.80 | 0.80 | 1 |
| Obs-intended-fc slope | 0.69 | 0.61 | 0.61 | 1 |
| Obs-(low)int-fc slope | 0.65 | 0.36 | 0.36 | 1 |
| null log-fc IQR | 0.85 | 0.19 | 0.20 | 0 |
| null log-fc 99.9% | 4.48 | 0.57 | 0.58 | 0 |
| low AUC | 0.07 | 0.40 | 0.39 | 1 |
| med AUC | 0.00 | 0.87 | 0.86 | 1 |
| high AUC | 0.00 | 0.46 | 0.44 | 1 |
| weighted avg AUC | 0.05 | 0.52 | 0.51 | 1 |
| Median SD | 0.63 | 0.11 | 0.12 | 0 |
| low.slope | 0.72 | 0.35 | 0.35 | 1 |
| med.slope | 0.80 | 0.76 | 0.76 | 1 |
| high.slope | 0.45 | 0.47 | 0.47 | 1 |

Table 4 - affycompII most indicative results (as in (Irizarry et al.)) for MAS5, RMA and tRMA, spike-in HGU95 dataset. Differences between RMA and tRMA are trivial, especially when compared to other methods (see also (Irizarry et al.))

2.2.4 Comparison between RMA and tRMA in biological contexts

In order to demonstrate that our tRMA procedure still performs nearly as well as the original RMA implementation, I used the latest implementation of the AffyComp (Cope et al., 2004; Irizarry et al.) benchmark (“AffycompII”) to compare the performance of the original RMA with our tRMA implementation. This benchmark is a well known tool to evaluate summaries of Affymetrix probe level data, based on known concentration of transcripts in the so-called “spike-in” experiments by Affymetrix (Affymetrix). In Table 4 I show some of the most relevant scores, as calculated for the HGU95 Affymetrix spike-in series. It is essentially a draw, with tRMA and RMA performing largely identically in all tests.

The differences between RMA and tRMA are minor when compared to the results for an independent method (MAS5): tRMA shares most of the qualities of RMA, without introducing inter-array correlations. It is interesting to highlight the fact that tRMA yields a higher median standard deviation (Median SD, in bold in Table 4) between spike-in replicates. This effect can be wrongly interpreted as tRMA’s lower sensitivity; however, this is due to the introduction, by the original RMA median polish implementation, of identical values across experiments, and therefore by the artificial reduction of the variance between spike-in replicates as well.

Since median polish alters inter-array correlation, sample classification is a common analysis that could be affected by this summarization step. Thus, I analyzed the AtGenExpress stress dataset for Arabidopsis (Kilian et al., 2007), and calculated the capability of both preprocessing techniques to separate root and shoot samples (see Paragraph 4.2.9).

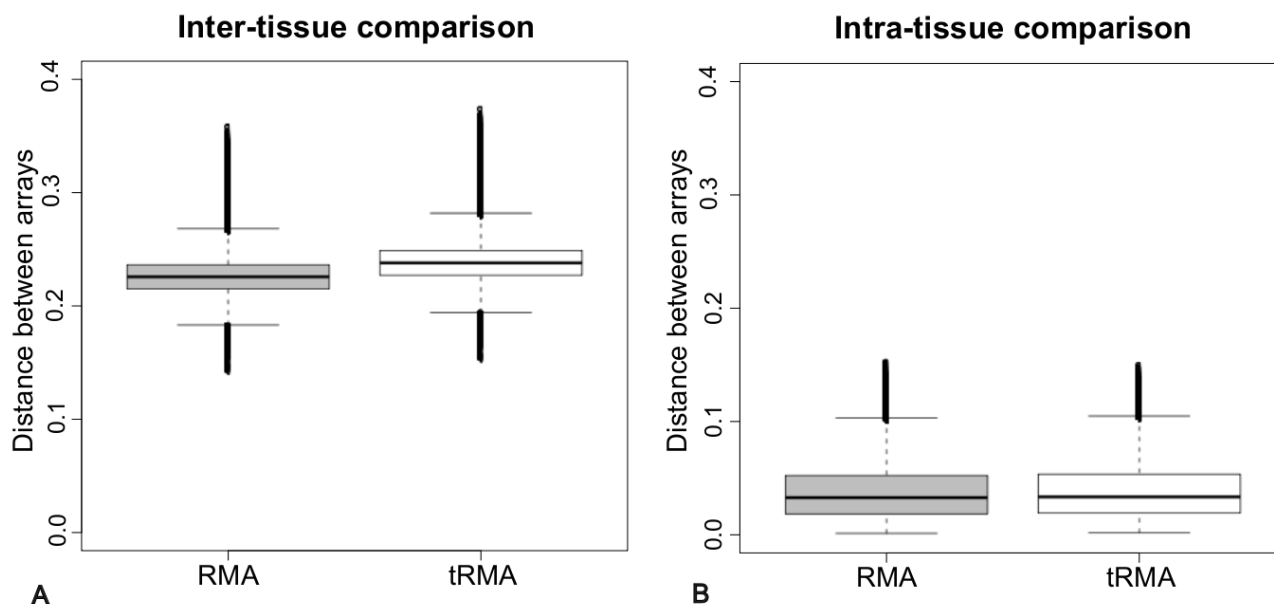


Figure 19 - Distances between Arabidopsis microarrays belonging to (A) different tissues (roots and shoots) and (B) the same tissue in 10000 5-samples subsets, calculated after RMA (left) preprocessing or tRMA (right) preprocessing. All probesets were used in this analysis. Distances are reported on the y axis and calculated as $(1 - \text{Spearman's correlation coefficient})$.

As can be seen in Figure 19A, tRMA outperforms RMA as it increases the distance between different tissue samples (Wilcoxon test: $p\text{-value} < 2.2 \times 10^{-16}$), while keeping similar low distances between samples coming from the same tissue (Figure 19B, Wilcoxon test: $p\text{-value} = 0.935$). As variance filtering is a common procedure for microarray clustering, I used only the 50% most varying genes in every subset and obtained similar results (inter-tissue distance $p\text{-value} < 2.2 \times 10^{-16}$, intra-tissue distance $p\text{-value} = 0.141$). It can be concluded that tRMA increases the capability to discern different array conditions, when only a small number of microarrays have been used.

In order to compare the relative performance of RMA and tRMA when filtering on differentially expressed genes, I used a dataset that was previously used by (Eklund and Szallasi, 2008), to tune classification where the provenience of the RNA in each sample was known. Choosing a sample size of 5 where 2 pairs of 2 samples each came from the same specimen and one sample came from a different specimen, tRMA yields better classification results for almost all FDR corrected $p\text{-value}$ thresholds (Figure 20).

However, when filtering out lowly expressed genes (Datta, 2003), RMA performed generally as well as tRMA when performing sample classification on this dataset (Figure 20).

Although objectively minor, these differences point out that tRMA may not necessarily be an improvement over RMA in all types of analyses.

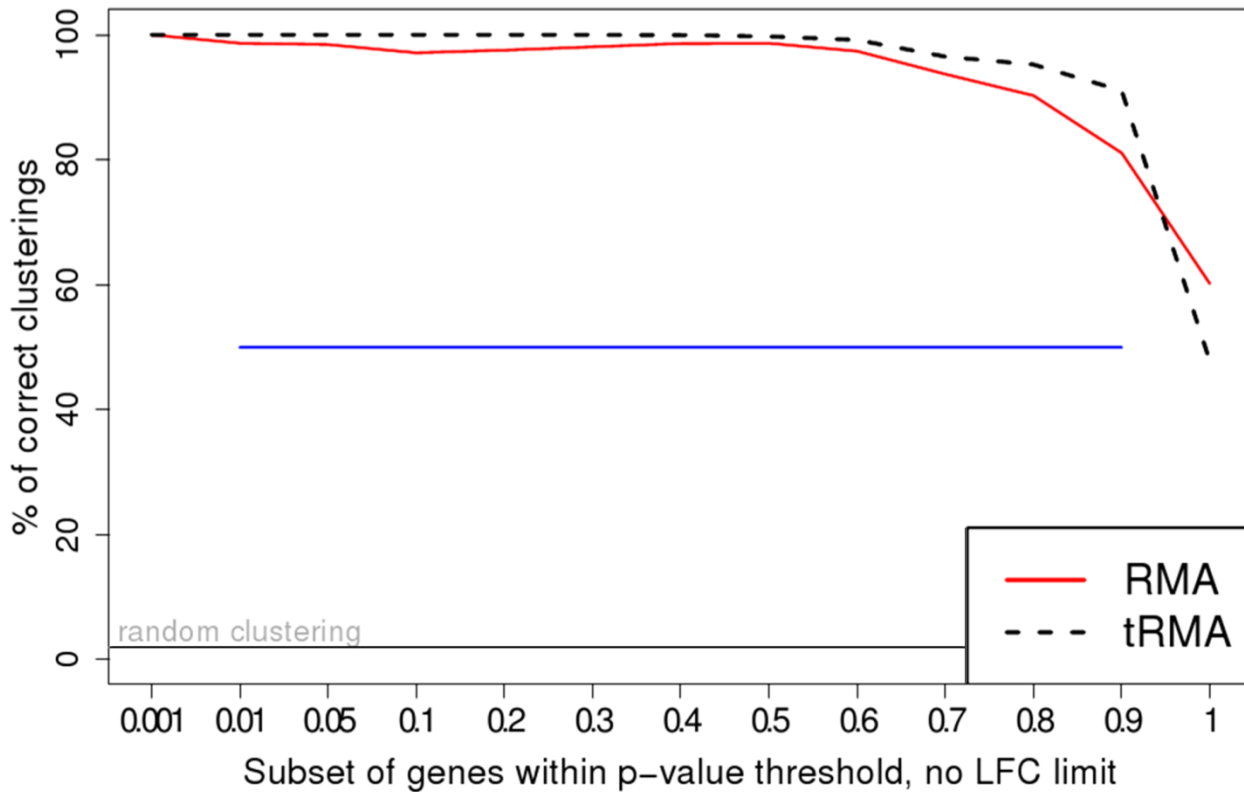


Figure 20 - Percentage of correctly clustered subsets of 1000 samples of 5 microarrays using a clinical dataset from (Eklund and Szallasi, 2008). The blue line indicates the range of significantly higher tRMA performance. The bottom grey line indicates the expected correct percentage from random clustering.

2.2.5 Conclusions on median polish based microarrays normalization methods

The use of GCRMA and RMA preprocessing algorithms for Affymetrix GeneChip technology has received a remarkably broad adoption in the community due to their low computation time and to their superiority with respect to other methods in previous benchmarks. However, one of the most relevant advantage of RMA and GCRMA in the AffycompII challenge (Cope et al., 2004), the low variance across replicates, seems to be partially the result of artificial inter-array correlation. Extending what was already noted by (Lim et al., 2007), I show that the artificially high similarity between samples given by RMA and GCRMA is caused by the shared median polish summarization step, step that could be corrected without losing any of the RMA/GCRMA positive properties. This artificial behavior is particularly strong in internally inconsistent, noise-driven and multi-hit probesets, and as a result identical results across arrays are generated. I analyzed this artifact effect for the Arabidopsis thaliana ATH1 Affymetrix GeneChip, but I found highly similar results in exploratory experiments on other organisms and platforms (specifically, human HG133 and E.coli Asv2 - data not shown).

2.3 Combining network centrality analysis and conditional correlation: application to essential gene prediction

In this Paragraph, I will show how the conditional correlation approach can be combined with the observation that essential genes tend to have a high network centrality in biological networks (Jeong et al., 2001).

2.3.1 Definition of Breaking Potential

Here, I propose a new concept of conditional centrality for networks extracted from data, which I specifically applied to coexpression networks derived from microarray data. An underlying gene regulatory network fragment, shown in the example in Figure 21a, will typically appear as a more densely connected undirected network, as in Figure 21b. With the use of conditional correlation it is possible to remove indirect connections and obtain a network as in Figure 21c. For instance, the edge (V1-V4) is removed upon conditioning on V3, which represents an intermediate in the pathway connecting them. This approach has recently been employed in network reverse engineering based on partial correlations (de la Fuente et al., 2004; Reverter and Chan, 2008).

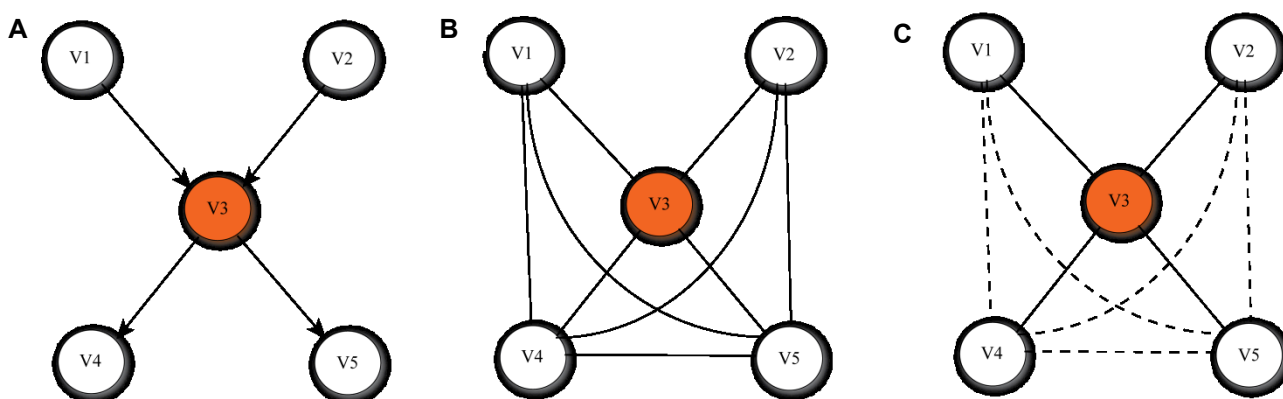


Figure 21 - Real network (A), correlation map between nodes (B) and removal of spurious connections upon conditioning on the V3 node (C)

Existing studies based on partial correlation do not consider the number of edges removed upon conditioning on a certain vertex in the calculation of the measure. During the conditional steps of network reconstruction, when first order partial correlation is applied, it is possible to count the number of edges a vertex can break (n_{brk}), and then normalize this number by the actual number of edges formed by its neighbors, yielding the term:

$$brk = \frac{n_{brk}}{n_{eff}}$$

The last expression gives the definition of the "Breaking Potential" index (brk) as a measurement of network causal importance (possibly, a "centrality") of a vertex. It allows the central regulator of information flow to be distinguished, where other commonly used centrality indices do not, as shown in the Appendix (page 103) with an artificial network example based on Figure 21. The main network centrality measures treated here are three: Degree, Betweenness and Clustering coefficient. The Degree of a gene is given by the number of its partners in the network; therefore, this centrality is commonly recognized as a measure of the overall activity

and inclusion of a gene in various cellular processes (Koschützki and Schreiber, 2008). The Betweenness is defined by the number of shortest paths crossing the gene, and it is considered as a measure of the control power the gene has over information transfer within the network (Koschützki and Schreiber, 2008). Finally, Clustering coefficient is calculated by the number of connections between a vertex's neighbours (n), divided by the maximum number of possible neighbors' connections ($n*(n-1)*0.5$); as its name implies, Clustering coefficient defines the rate of interconnection and "entanglement" of a subnetwork. The Breaking Potential index adds to these techniques the capability of analyzing the indirect correlation effects which can be detected by conditional techniques. It must be observed that an alternative to the Breaking Potential index could be simply the Degree of a Partial correlation network. However, in such a network the information on the capability of a node to dissolve specific edges in its neighborhood would be lost. Furthermore, simple partial correlation degree would be dependent on the initial zeroth order degree, and wouldn't account for the effective number of initial connections in the gene neighborhood.

2.3.2 Comparison between Breaking Potential and other centralities in *Arabidopsis thaliana* coexpression networks

We tested Breaking Potential on data obtained from a high quality microarray dataset from *Arabidopsis thaliana* (see Paragraph 4.3). As a threshold to generate the correlation network (r_0) we used $r_0=0.7$. This threshold is commonly used and accepted in literature for coexpression networks (Usadel et al., 2009) and yields realistically sized networks (1,457,367 edges, equaling to 0.020% of all possible edges). However I obtained similar results based on the same dataset and zeroth order correlation threshold $r_0=0.6$ (3,414,431, 0.046% of all possible edges) and $r_0=0.8$ (460,279, 0.006% of all possible edges). In Figure 22, it is shown how Breaking Potential relates to three other centralities for each gene analyzed. A positive correlation with Degree and Betweenness can be assessed both in non-logarithmic (Figure 22A-C) and logarithmic (Figure 22D-F) scale.

2.3.3 Breaking Potential is a positive predictor for gene essentiality in *Arabidopsis thaliana*

Breaking Potential can in principle be used to extract nodes central in information flow processes (see Figure 21), so we investigated how its capability to find particularly "central" and "fragile" nodes in gene regulation networks. We consider these nodes as genes the removal of which is sufficient to bring the whole organism to death, and to this aim we used a list of manually annotated essential genes for *Arabidopsis thaliana*. Homozygous mutations inactivating the function of these genes lead to embryonic lethality (Meinke et al., 2008). In fact, the Breaking Potential for essential genes is significantly higher than that for non-essential genes (Figure 23, panel A). This fact led us to the initial conclusion that Breaking Potential might be used as a major *in silico* marker for gene essentiality, as the same property was observed in preliminary studies in *Escherichia coli* and *Saccharomyces cerevisiae* studies (data not shown).

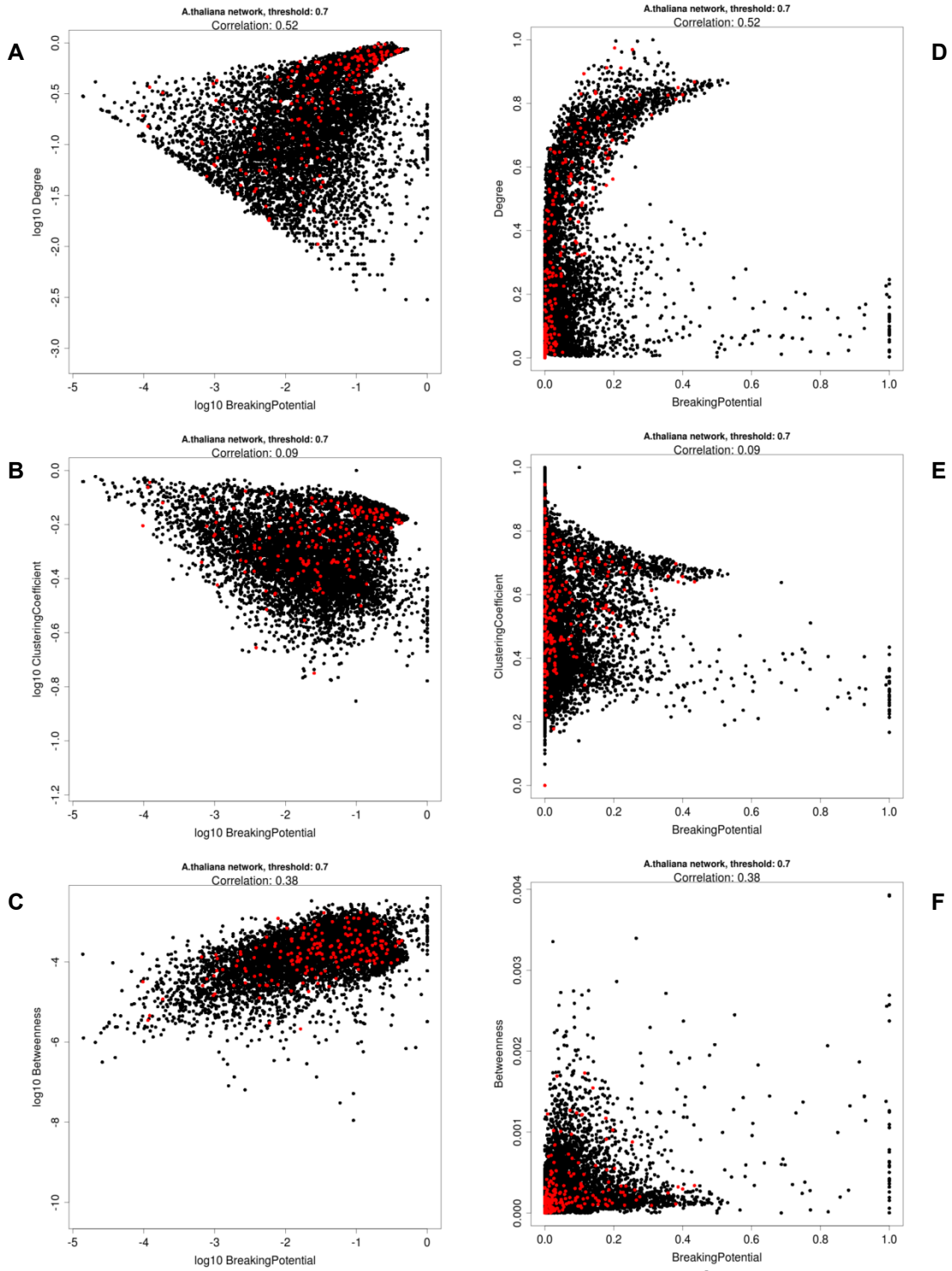


Figure 22 - Relationship between Breaking Potential and other centrality measures in the Arabidopsis thaliana coexpression network. Red points represent essential genes.

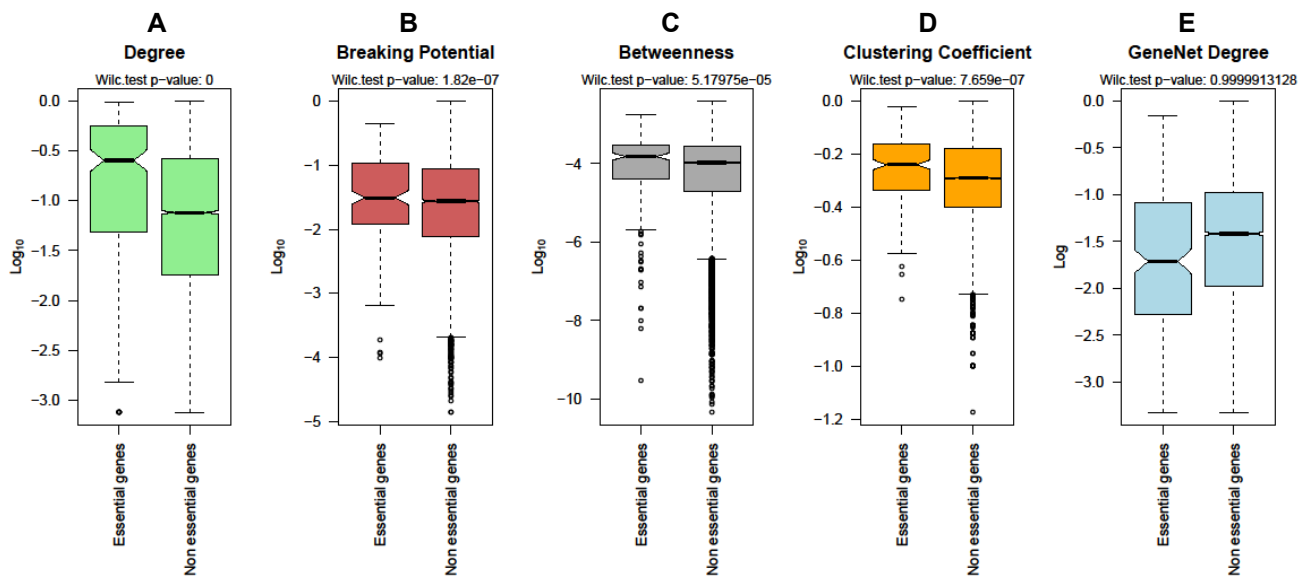


Figure 23 - Boxplots indicating distribution for essential (left) and non essential (right) genes calculated for different centrality measures in correlation networks, namely Degree (A), Breaking Potential (B), Betweenness (C), Clustering Coefficient (D) and GeneNet degree (E) (Opgen-Rhein and Strimmer, 2007). P-values for essential genes having a higher centrality were calculated using Wilcoxon test and shown under the centrality indices names

However, since essential genes are more "central" than non-essential ones also for other centrality measures (Figure 23, panels B-D), and since Breaking Potential is positively correlated with them (see Figure 22), especially with degree, it must be ruled out that what we see in Figure 23A is not a mere by-product of other centralities. To assess this, we calculated ROC curves for the essential genes prediction power of all centralities alone and when combined with Breaking Potential (results shown in Appendix, page 106). Interestingly enough, while Breaking Potential seems better than Clustering coefficient and Betweenness, and furthermore seems additive to them, in the task of separating essential genes from non essential ones, this is not the case for network Degree. Degree is constantly a better predictor than Breaking Potential in all scenarios investigated. The Areas Under the ROC curves (AUROC), generated by repetitive subset selection of the gene populations, give us the same message (data not shown). We tested also several hybrid centrality methods, among which a ranked combination of Breaking Potential and Degree indices for predicting essential genes (Appendix, page 106, Figure 49A, red dashed line), resulting in Degree always been the best predictor. Similar conclusions to our analysis could be drawn not only for correlation threshold 0.7, but also for 0.6 and 0.8.

Finally, we assessed the centrality of essential genes in networks extracted from full partial correlation matrices, e.g. via the GeneNet approach (Schäfer and Strimmer, 2005). In these networks the Degree seems,

surprisingly enough, a counter indicator for essentiality (Figure 23E and as a ROC curve in the Appendix, page 106, Figure 49D).

2.3.4 Conclusions on Breaking Potential as an essential gene predictor and future perspectives

Network analysis has provided major contributions to the understanding of the biological systems. Identifying *central* vertices, representing genes, proteins, metabolites, and other biological entities, has been the technique of choice for determining the key players in intracellular information flows. We used a conditional correlation-based approach to step further into the analysis of gene networks via coexpression, not by removing indirect edges, but by focusing on the conditional centrality of every gene. We defined the Breaking Potential index as the capability of a gene to dissolve correlations among its neighbors, by using a simple Pearson partial correlation approach.

We have investigated the properties of this novel conditional centrality in *Arabidopsis thaliana*, concluding that Breaking Potential index may be able to discern vertices associated to key pathways in the coexpression networks, and that it is partially complementary to other prominent centralities (Figure 21). However it seems to be substantially less predictive than network degree alone (Figure 23). We therefore think that Breaking Potential is an interesting device for analyzing causally central genes in correlation networks in *Arabidopsis thaliana*, but in the task of finding known "central" genes it is less performing than degree alone. For some of these, namely transcription factors, non-transcriptional control coupled with low expression makes finding relevant properties from microarray data a difficult task. We have demonstrated however the predictive power of all these centrality measures for finding essential genes, despite their ontological heterogeneity (Chun and Goebel, 2004) (Figure 23).

2.4 Expression-based gene network reverse engineering

2.4.1 Custom network reverse engineering and method comparison: the CorTo tool

As explained in the introduction, the "guilt-by-association" principle underlying the expression-based gene network reverse engineering has been useful for finding new candidates in a certain pathway or mechanism where certain "bait" genes were known. There are many co-occurrence tools and web services available, predominantly for plant species (Usadel et al., 2009), but also for mammals (Gurkan et al., 2005) (Zimmermann et al., 2004) (Obayashi and Kinoshita, 2011), and almost exclusively focusing on transcripts (hence the name term "co-expression tools"). However, there was hardly any user-friendly, light weight and fast stand alone tool being able to (1) cope with arbitrary and large data sets such as the ones in the following sections of this thesis, (2) implement advanced network reverse engineering methods and (3) combine network reconstruction approaches with functional analysis. Therefore, I developed CorTo ("Correlation Tool") as a lightweight and integrated analyzer of co-occurrence in multi-sample quantification datasets, designed with transcript co-expression and metabolite co-accumulation analyses in mind. CorTo is the answer to many of the issues I encountered during the comparative analysis of the gene network reverse engineering algorithms (see paragraphs 2.4.3, 2.5 and 2.6), namely: the necessity to have efficient and fast computational capabilities; the potential to analyze big (thousands of genes and samples) datasets; the convenience of method comparison through a visual representation of the connections. The biological community will take benefit from such a tool to the community, since many techniques (e.g. the LASSO) for gene network reconstruction are usually requiring strong bioinformatics skills to be performed correctly and with low computational time.

CorTo combines the possibility of providing a custom dataset with a range of direct co-occurrence techniques, namely Pearson correlation, Spearman correlation, Mutual information plus some "indirect" methods (Zampieri et al., 2008), i.e. Pearson/Spearman partial correlation (de la Fuente et al., 2004) and LASSO regression (Tibshirani, 1996). These two groups of techniques have been shown to be complementary in analyzing two classes of co-occurrence: direct methods, often more suited for co-present biological features (e.g. components of a protein complex) and indirect methods, more accurate at reconstructing causal processes (e.g. transcription factor-target relationships) (Zampieri et al., 2008). CorTo uses as input format tab-separated files with samples as columns and measured entities as rows and is preloaded with a number of *Arabidopsis thaliana* and *Solanum lycopersicum* datasets. In order to exclude noise-driven measures and to speed up calculations, CorTo implements an optional subset extraction step that analyzes only the most varying elements of the dataset, or alternatively focus on the gene/metabolite of interest. Analysis can then be carried out on one or more genes/metabolites, and a visualization performed providing all the elements that co-occur in the dataset as a network visualization. It is possible in any case to expand any analysis on-the-fly, by clicking on areas of the network where co-regulation should be calculated, up to (potentially) a global network reconstruction. Performance-wise, CorTo can calculate the co-occurrence for a single feature in a normalized microarray dataset with 22813 transcripts and 22 samples on a standard computer (e.g. Intel Core Duo E8400 3.00GHz) almost instantaneously.

Network representations of the co-occurrence analysis can then be defined and visualized by any combination of technique and threshold constraint, making it possible to compare different techniques (Figure 24). A color-based functional category map can be used for those species where a MapMan annotation is available (Usadel et al., 2009). Networks can be expanded at any time by right-clicking on the interactors of the already expanded genes/metabolites. Networks can be printed out as tables for further analysis; this feature is especially useful if the number of variables shown is too high for an interpretable visualization. The user can also explore any pairwise behavior in the dataset as a scatter plot, and show the co-occurrence behavior of several genes/metabolites as series plots.

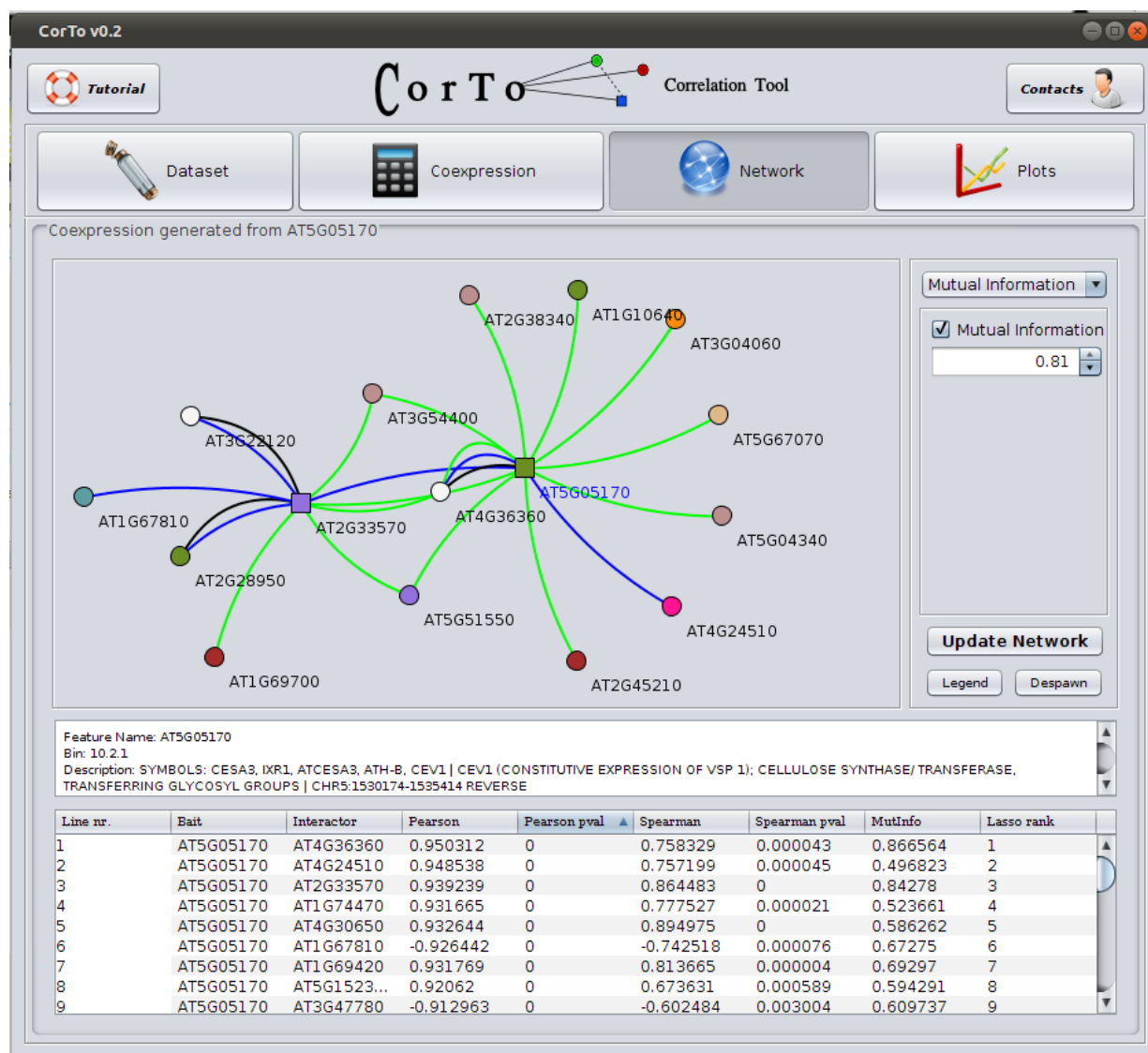


Figure 24 - Investigating the coexpressors of the *Arabidopsis thaliana* gene At5g05170 (Cellulose Synthase 3) in a flower and pollen dataset (Schmid et al., 2005). Several coexpression techniques have been used to infer potential connections: Mutual Information (green), the LASSO (blue) and Pearson Correlation (black).

2.4.2 Application of the LASSO to gene expression-based modeling

We briefly mentioned in the Introduction (Paragraph 1.3) a relatively new method has been proposed (Tibshirani, 1996) in the field of gene network reverse engineering called the LASSO (Least Absolute Shrinkage and Selection Operator). This method has been tested so far only on small networks (Lu et al., 2011) or on largely theoretical scenarios (Gustafsson et al., 2009) and is technically based on a modified Linear regression principle. Linear regression models a response (dependent) variable (y) through a list of predictor (independent) variables ($x_1, x_2, x_3 \dots x_n$). Linear models have been extensively used to provide an insightful description of how gene expression is influenced by several factors, e.g. carbon availability and the circadian clock (Usadel et al., 2008), and how genes behave in relation to each other (D'haeseleer et al., 2000).

Given a response y and a list of predictors x_i ($x_1, x_2, x_3 \dots x_n$), an ordinary linear model yields as a result a function $f(x_i)$:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \dots + b_nx_n$$

Where $b_1, b_2, b_3 \dots b_n$ are the weights assigned to every predictor, in order to minimize the unexplained behaviour of y (the residuals). When the number of predictor variables is higher than the number of measurements (samples), which is mostly the case when dealing with gene expression data, the underlying equation system is underdetermined and the model is overfitted, always explaining the response perfectly. To overcome this problem a class of linear regression methods, the so-called “shrinkage methods” has been developed. Shrinkage methods retain only a subset of the predictors and discard all the rest, providing a final model that is interpretable and possibly more accurate (Copas, 1983). The LASSO is a particular and recently proposed shrinkage technique (Tibshirani, 1996), which imposes a limit to the weights assigned to the predictor variables:

$$|b_1| + |b_2| + |b_3| + \dots + |b_n| \leq L1$$

Where $L1$ is a tuning parameter for the stringency of the model. Because of the nature of the constraint, making $L1$ sufficiently small will cause some of the coefficients to be exactly zero, so that several variables get discarded. This increases the *interpretability* of LASSO models, as relevant variables can be clearly separated from irrelevant ones.

The LASSO has been used in research to generate well performing models where a clear border between important and unimportant variables had to be discerned (Hastie et al., 2001), although with only a handful of biological applications so far (Shimamura et al., 2007; Gustafsson et al., 2009; Lu et al., 2011). The original algorithm to obtain the solution of LASSO at all possible sum-of-weights thresholds (referred to as $L1$ thresholds) is a computationally very demanding task (Efron et al., 2004) and is of nearly no practical interest for big datasets (more than hundreds of variables). However a more efficient algorithm to solve the full LASSO model has been recently developed, called **L**east **A**nge **R**egression for **L**ASSO, or simply LARS (Efron et al., 2004). In brief, LARS starts introducing an explanatory variable to the model and continues to increase its

weight in the model until a second variable reaches the same correlation with the model's residuals as the initial variable. Then, the model proceeds modifying the weights of the two variables in a direction that is *equiangular* to both. This process balances all variables in the model, while excluding indirect effects, since increasing the weight of one variable also reduces the chance to include variables from the same informational area, similarly to what happens for Partial correlation (see Paragraph 1.3). LARS will also discard variables that, during the iterative process, experience a sign change in their weight. In such a case, the variable is discarded from the model and all other variables' weights are subsequently re-calculated. The computation of LARS is therefore a loop involving linear algebra operations in the variable spaces, calculating residuals of the model at each step and then taking the decision of including/dropping variables (Efron et al., 2004).

The LASSO, although potentially interesting in gene network reverse engineering tasks, has never been used so far for large-scale biological investigations. We therefore developed a light Java implementation of the LARS algorithm (see Paragraph 4.6) optimized for global network reconstruction, which pre-calculates the steps shared by all models in a gene dataset and provides a parallel computation of the LARS solution. In the next paragraph, we will use this LASSO implementation for inferring expression-based relationships between genes in *Arabidopsis thaliana* and compare it with other common reverse engineering methods.

2.4.3 Comparative analysis of expression-based methods for gene network reverse engineering

In order to compare different network reverse-engineering techniques, it is necessary to adopt several methods for assessing network quality. I adopted the following ones, inspired from various literature sources, mainly (Lim et al., 2007; Usadel et al., 2009) (see also Materials and Methods, Paragraph 4.5). In Table 5, the methods properties are summarized.

First, I adopted the so-called "Ontology Agreement" method. The idea that genes sharing similar functions are also co-regulated is well-established (Stuart et al., 2003; Yu et al., 2003), and it is possible to assume that an expression-based gene network will contain several clusters of genes sharing identical functions (see also later in this Paragraph, Figure 31B) (Peng and Weselake, 2011). This is true for at least two reasons: genes are co-regulated to produce stoichiometrically balanced quantities of subunits interacting in the same complex (Tanya and Ben, 2008), or they are co-regulated to carry out parallel or sequential activities in the same pathway (Thimm et al., 2004). This method is highly dependent on the completion of gene annotation for a given organism; in our case, I used the 2010 MapMan annotation for *Arabidopsis thaliana*, which covers more than 60% of this organism gene population (Usadel et al., 2009).

Second, I adopted the theoretical assumption that a biological network degree distribution should be scale-free, *i.e.* follow a power law (Barabási and Albert, 1999). This method is assessing the structural quality of the network, assuming the presence of a reduced amount of central gene regulators (or hubs) and a high amount of genes with a reduced number of connections (see also Figure 2). However, it's been debated in literature that biological networks are not necessarily "scale-free" (Khanin and Wit, 2006).

| Network Quality Assessment | Advantages | Disadvantages |
|---|--|--|
| Ontology Agreement | <ol style="list-style-type: none"> 1. Based on sensible biological assumption: a network will contain several connections between genes having similar functions 2. In the organism investigated (<i>Arabidopsis thaliana</i>) the MapMan ontological annotation is highly curated (Usadel et al., 2009) | <ol style="list-style-type: none"> 1. Highly dependant on the amount of ontological information available on the organism 2. Over-assumption bias: it assumes as "wrong" connections between different ontologies |
| Degree distribution fit to power law | "Pure" method, doesn't require any <i>a priori</i> knowledge on the organism | It assumes that expression networks a scale-free distribution (Barabási and Albert, 1999) of the degree, which is not necessarily the case (Khanin and Wit, 2006) |
| Overlap with experimentally verified protein-protein interaction (PPI) networks | Direct co-regulation and protein-protein interaction have been positively associated in several studies (Zampieri et al., 2008) (Persson et al., 2005) (Yu et al., 2003) | <ol style="list-style-type: none"> 1. Dependant on available experimental data, which for the protein-protein interaction networks is complete only for a few organisms. <i>Arabidopsis thaliana</i> AtPin contains only around 6000 interactions, and therefore it's lacking around 80% of the complete picture 2. Protein-Protein interaction isn't necessarily associated with gene co-regulation |
| Overlap with manually curated genetic interactions | The "real" network to be reverse-engineered from expression data is the genetic interaction network | <ol style="list-style-type: none"> 1. Nearly-complete gene-gene interaction maps are available only for simpler organisms (Gama-Castro et al., 2008) 2. Ideally, network motifs like feedback loops are nearly undetectable in condition-independent datasets (Shipley, 2002; Usadel et al., 2009) |
| Experimental validation | Based on real evidence | <ol style="list-style-type: none"> 1. Time-consuming 2. Costly |

Table 5 - Summary table for Network Quality Assessment methods

I then adopted two methods, namely overlapping the obtained reverse-engineered networks with known experimentally validated and publicly available gene-gene connections, specifically (for *Arabidopsis thaliana*) protein-protein interactions (PPI)(Brandão et al., 2009) and genetic interactions (Palaniswamy et al., 2006). For *Arabidopsis thaliana*, this information is far from being complete and therefore some *caveats* have to be applied to these quality assessments (only 6784 protein experimental interactions have been manually collected in AtPin (Brandão et al., 2009), and 10640 genetic interactions have been annotated in AtRegNet (Palaniswamy et al., 2006), out of 27416 protein coding genes confirmed in this organism - TAIR10 (Swarbreck et al., 2008)). Furthermore, PPI, although it is often accompanied by gene co-regulation and it requires at least basal expression of the interactors (Yu et al., 2003; Persson et al., 2005; Zampieri et al., 2008), is not necessarily requiring co-expression to occur.

Finally, as will be extensively discussed in Paragraphs 2.5 and 2.6, I decided to compare the different gene network reverse engineering methods, particularly LASSO and Correlation, in experimental scenarios, in order to validate their conclusions in "real" biological pathways.

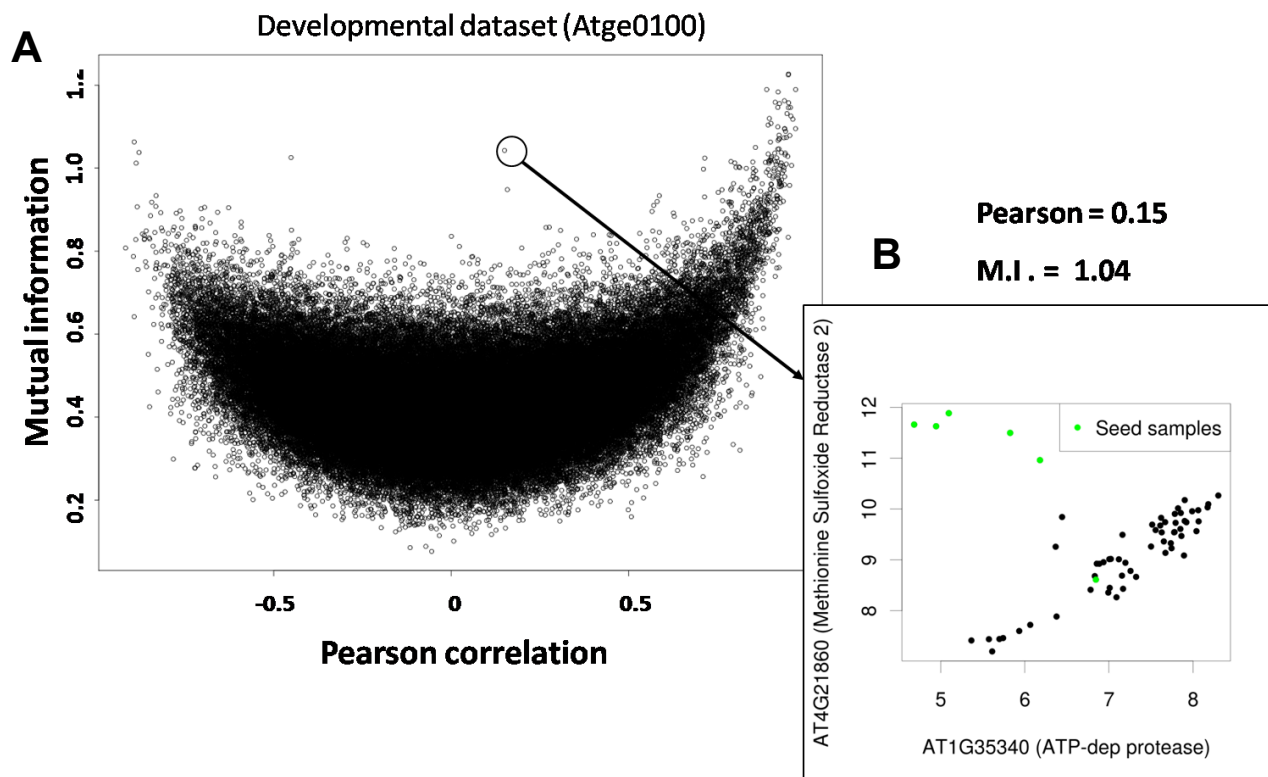


Figure 25 - Mutual Information (unnormalized index) vs. Pearson correlation coefficient (panel A). In panel B, a particular case of high M.I. index and low Pearson correlation scenario is shown. Here, a positive trend between two genes in all samples except in seed tissues can be observed

Once a list of methods to assess network quality had been characterized, I assembled a list of the most used and arguably the most powerful expression-based network reconstruction techniques (Zampieri et al., 2008): Pearson Correlation, Partial Pearson Correlation and Mutual Information. In particular, I investigated the capability of Mutual Information to discern non-linear relationships between gene expression patterns (Daub et al., 2004). One example of this capability of Mutual Information is shown in Figure 25, where several gene-gene relationships are quantified based on their expression in the *Arabidopsis thaliana* AtGenExpress developmental dataset (*atge0100* (Steinhauser et al., 2004), see Paragraph 4.3). The example shows how the gene At1g35340 (an ATP-dependent protease of unknown function) positively correlates with At4g21860 (Methionine Sulfoxide Reductase 2) in all samples, except in seed tissues (Figure 25B, green points), where they appear to have a slightly negative correlation. Although interesting, the lack of more real examples of such differing behaviors between Pearson Correlation and Mutual Information make such finding more an interesting exception than a general expression phenomenon (Daub et al., 2004).

In order to comparatively assess the capability of these techniques to reverse engineer biological networks, I applied them on a large *Arabidopsis thaliana* dataset collecting all publicly available Affymetrix microarrays (Paragraph 4.5 and (Mutwil et al., 2011)). I used several threshold combinations for direct and partial Pearson Correlation, and several bin number/relevance index for Mutual Information, in order to assess the power of these techniques as widely as possible.

For the LASSO, I generated models for every gene and selected for each of them one single lowest-error (based on cross-validation model) L1. In the best-L1 models we drew edges between the dependent variable and the independent ones with a non-zero weight in the model. Finally, I merged all models' edges into a final global network.

While the LASSO reconstruction yields a single lowest-error network, there is no such default assessment for the other methods, which therefore need to be assessed at various stringency levels. I therefore generated a collection of Correlation and Mutual Information networks, obtaining a total representation of the network characteristics judged by several quality approaches. This study not only allows us to compare the LASSO with other methods, but it is also the first comprehensive threshold-independent assessment of reverse engineered expression-based networks with a biological perspective.

The first results, obtained through the Ontology Agreement method (based on MapMan annotations of *Arabidopsis thaliana* (Usadel et al., 2009)), are shown in Figure 26 for Correlation techniques. As expected, an increase of threshold stringency increases the quality of the networks, while simultaneously reducing the size of the networks (Figure 27). The highest stringency networks ($r_0 > 0.6$ and/or $r_1 > 0.2$) are composed by a few hundred connections, largely describing within-pathway relationships (up to 71% edges formed by genes sharing a functional annotation). The performance of partial correlation yields, in comparison to standard correlation, largely different results, with apparent lower quality; for example compare in Figure 26 and Figure 27 the Ontology Agreement for correlation networks at $r_0 = 0.8$ (2766 edges, 68.44% Ontology score) with the similar size but lower score partial correlation network at $r_0 = 0$ to 0.5 and $r_1 = 0.4$ (2333 edges, 25.12% Ontology score). The quality of Correlation networks with and without first order conditioning (i.e. Partial correlation) has also been assessed with other quality methods, with results shown in the Appendix (pages 112-119), and similar conclusions can be drawn: simple correlation ranks high in the network it yields. Partial correlation is able to remove, even at relatively low threshold, a great fraction of the possible edges (Figure 27), which implicates that the initial network structures are very dense and intercorrelated (partial correlation wouldn't have such a massive effect in completely unrelated expression patterns).

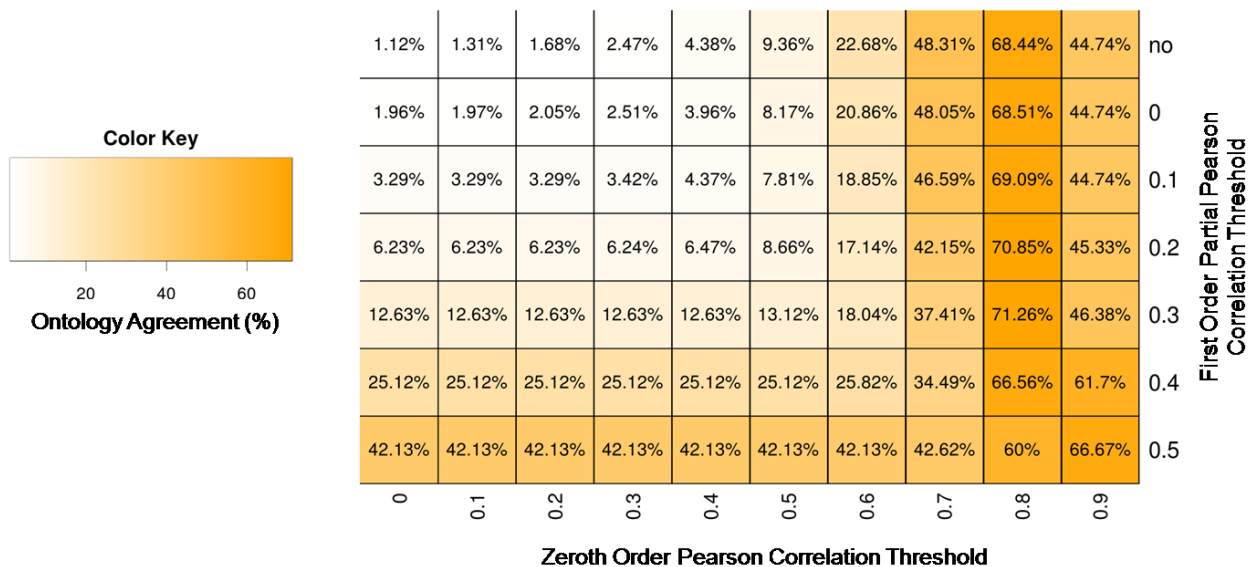


Figure 26 - Ontology Agreement percentage score for *Arabidopsis thaliana* expression based Correlation and Partial Correlation networks at different thresholds. Absolute correlation coefficients were considered. A first order threshold of 0 means that edges suffering a sign change were excluded from the resulting networks, while a first order threshold marked with "no" corresponds to the standard zeroth order Pearson Correlation network



Figure 27 - Network size (number of edges) for *Arabidopsis thaliana* expression based Correlation and Partial Correlation networks at different thresholds (see Figure 26)

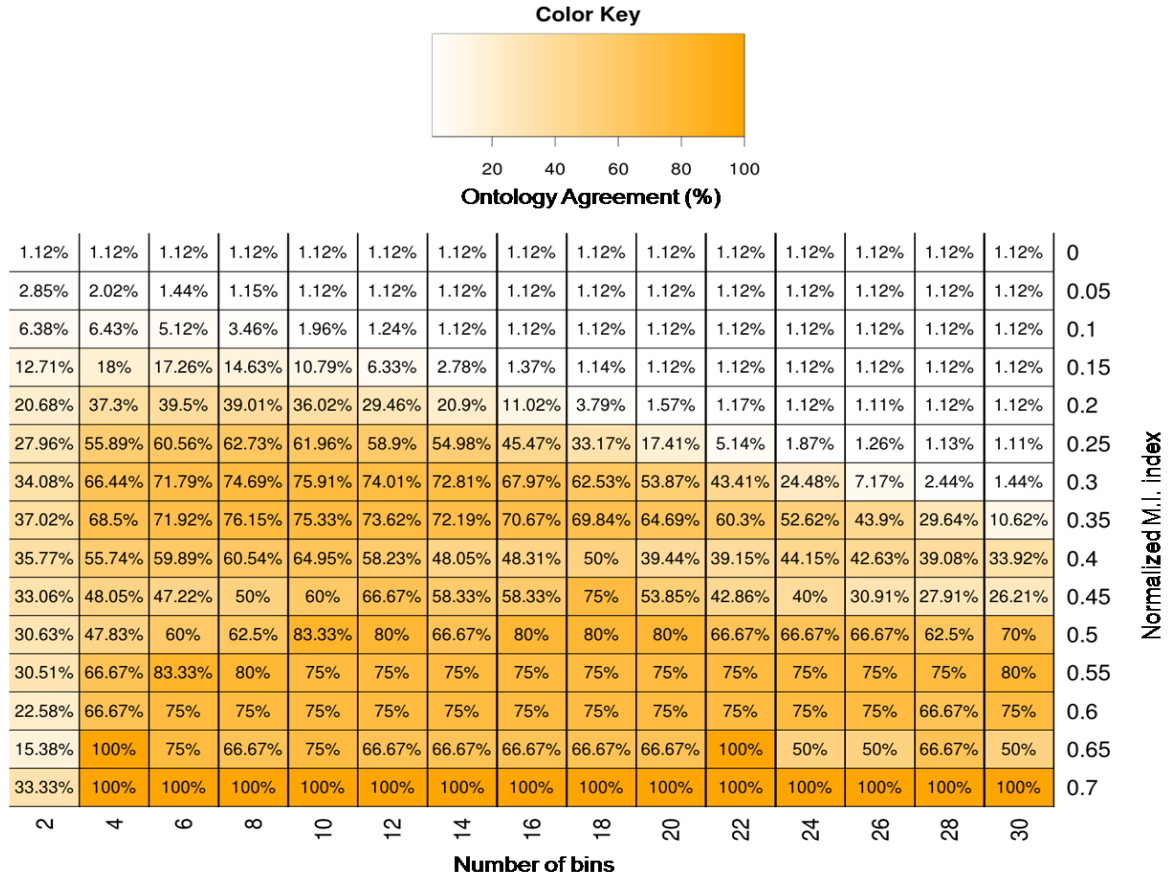


Figure 28 - Ontology Agreement percentage score for *Arabidopsis thaliana* expression based Mutual Information networks at different combinations of significance thresholds and bin numbers

We therefore analyzed the performance of Mutual Information at different thresholds and changing the number of bins, i.e. the discrete groups in which to subdivide the expression behavior of the genes prior to calculating the entropies. A bin number of 2, for example, signifies that both genes expression values were subdivided in two groups, either "high" or "low", while higher bin numbers arguably allow for a more finely tuned assessment of gene behavior. The results for Mutual Information are shown in Figure 28 for the Ontology agreement score and in Figure 29 for the size of these reverse-engineered networks. In general, it can be observed how the bin number changes the M.I. normalized index distribution and therefore the effect of the applied threshold used. This, although expected, has never been fully treated before in Mutual Information-reconstructed gene networks, and should be taken into consideration when choosing the index threshold in these kinds of inferences. Furthermore, the rapid shift of network topologies from fully connected to sparse structures is an indication of the very narrow M.I. index distribution, which is not so dissimilar between real datasets and what expected from a null distribution (see Appendix, page 120, Figure 60B). On the other hand, Pearson correlation shows a real distinction from what is observed and what would be expected from a completely uninformative dataset (see Appendix, page 120, Figure 60A).

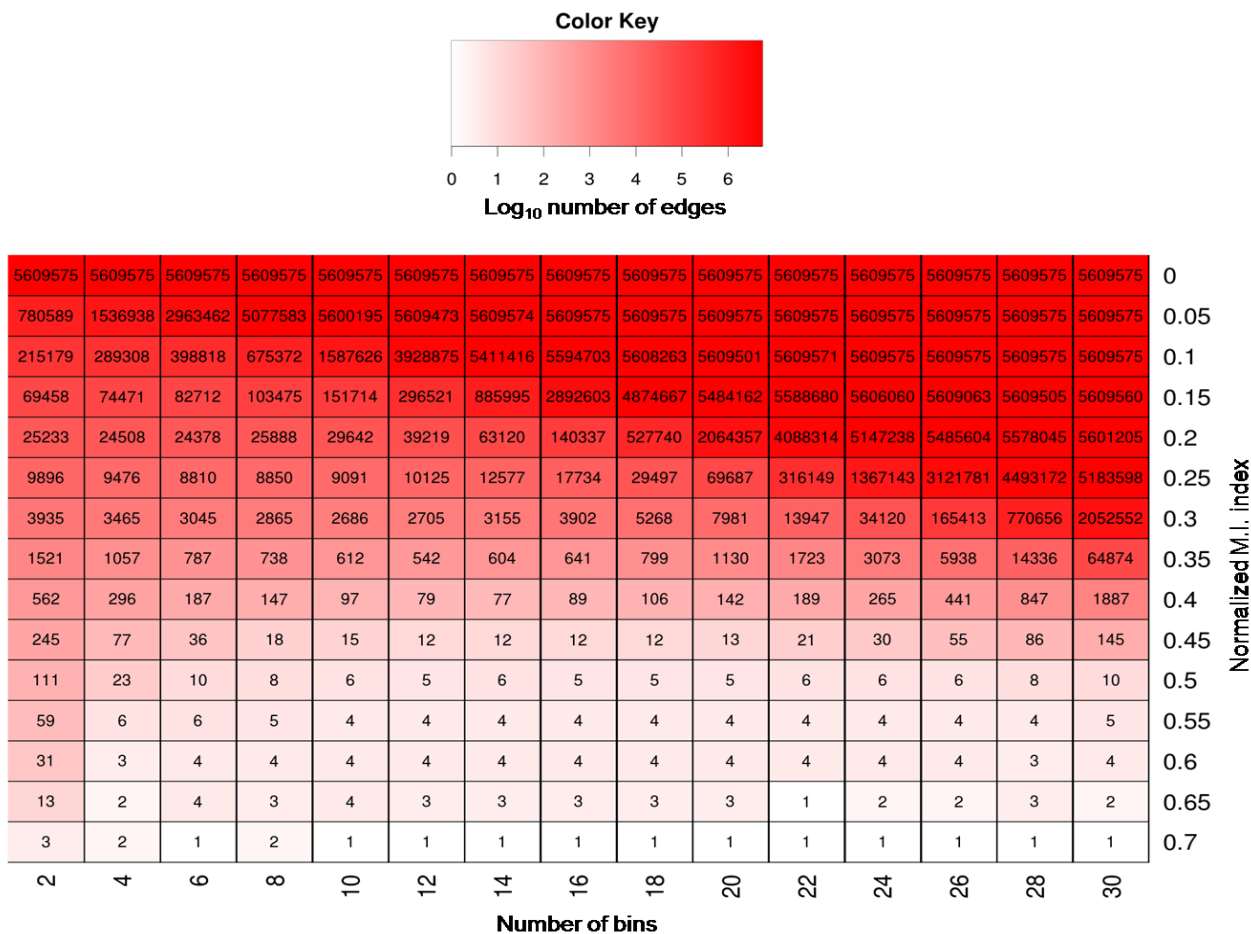


Figure 29 - Network size (number of edges) for *Arabidopsis thaliana* expression based Mutual Information networks at different combinations of significance thresholds and bin numbers

| Network Property | Pearson Correlation $r=0.4$ | Partial Correlation $r_0=0.2$ $r_1=0.1$ | Mutual Information M.I. norm. index=0.2 bins=18 | LASSO |
|---|--------------------------------|--|---|---------|
| Number of connections | 481,300 | 370,297 | 527,740 | 557,976 |
| Ontology Agreement | 4.38% | 3.29% | 3.79% | 1.38% |
| PPI overlap accuracy (Brandão et al., 2009) | 91.42% | 93.39% | 90.59% | 90.05% |
| PPI overlap Matthews coefficient (Brandão et al., 2009) | 0.0065 | 0.0064 | 0.0036 | 0.0069 |
| R^2 fit to a power law for degree distribution | 0.4386 | 0.0207 | 0.5046 | 0.1324 |

Table 6 - Network Quality Scores for the *Arabidopsis thaliana* LASSO network created by merging the best individual gene models generated on expression data and other networks with sizes of similar order of magnitude.

At this point, we generated LASSO models for every gene in the *Arabidopsis* dataset and among these we extracted the lowest-error cross validated model. Then, a connection was drawn between the bait used to generate the model and every gene used as explanatory variable, without consideration on the magnitude or

size of the weight, so the resulting connections could be merged and treated as a Boolean network. In almost all quality assessments, LASSO doesn't perform better than the other methods, reaching, with its lowest-error network composed by 557,976 edges, an Ontology Agreement percentage score of 1.38%, whereas Pearson Correlation networks with a similar size can achieve a score three times higher (Table 6).

With respect with the golden standard AtPin Protein-Protein interaction (PPI) network (Brandão et al., 2009), we calculated the amount of overlap between this and our inferred networks. Given the characteristics of the overlap, the number of True Positive hits can be calculated (gene-gene pairs present in both the PPI network and in the expression-based network), and consequently the True Negatives (gene pairs absent in both), the False Negatives (gene interactions present in the AtPin database but not found via reverse engineering) and the False Positives (gene pairs inferred by the computational predictions but not annotated as interacting in AtPin). From these numbers, we could assess the accuracy and Matthews coefficient (Baldi et al., 2000) describing the agreement of the inferred networks to the protein interaction data. All techniques quickly reach a high accuracy (>90%) signifying that the large part of the *Arabidopsis* genes whose protein products interact are also co-regulated, and that this co-regulation is detected by almost all methods. The other calculated index of agreement with the PPI network, the Matthews coefficient, comprehensively indicates the capability of a method (in this case, the network investigated) to successfully separate positive hits (i.e. the presence of a PPI) from negative ones (lack of PPI). The formula can be calculated given the number of True Positives (TP), True Negative (TN), False Positives (FP) and False Negatives (FN):

$$\text{Matthews coefficient} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP) \cdot (TP+FN) \cdot (TN+FP) \cdot (TN+FN)}}$$

The Matthews coefficient formula ranges from -1 (perfectly wrong binary classification) to +1 (perfect classification), with 0 being equivalent to a random assignment of PPI hits. In our networks (Table 5 for LASSO and similarly sized networks, Appendix, pages 118-119, for all networks) the Matthews coefficients are always positive (successful prediction of PPI golden set) but also very close to zero, which is given by the high amount of spurious interactions (i.e. not PPI connections) yielded by the network reverse-engineering. This is partly expected, due to the fact that only a few co-regulated gene pairs will actually be also interacting, and due to the incompleteness of *Arabidopsis* PPI experimental data. Another approach for assessing overlap to PPI would have been the use of ROC curves; however these curves, which show the sensitivity vs. specificity trend, provide very little information due to the high number of True Negatives in expression-based networks (Reverter and Chan, 2008)). It must also be noted that the network accuracy assessment through overlap with known genetic interactions (Table 5) cannot be robustly applied to this network, which contains only a handful of reliably expressed genes annotated in the collection of interactions publicly available (Palaniswamy et al., 2006). The problem of re-building genetic networks is also affected by the fact that genes involved in Transcriptional Regulation (e.g. Transcription Factors) are significantly less expressed than other genes, and therefore more likely to be affected by experimental noise and to pose a problem for network reverse engineering approach. This is true both for the whole population of *Arabidopsis thaliana* transcripts measured

on the ATH1 microarray (Figure 30A) and for a subset constituted by the most consistently expressed transcripts (Figure 30B).

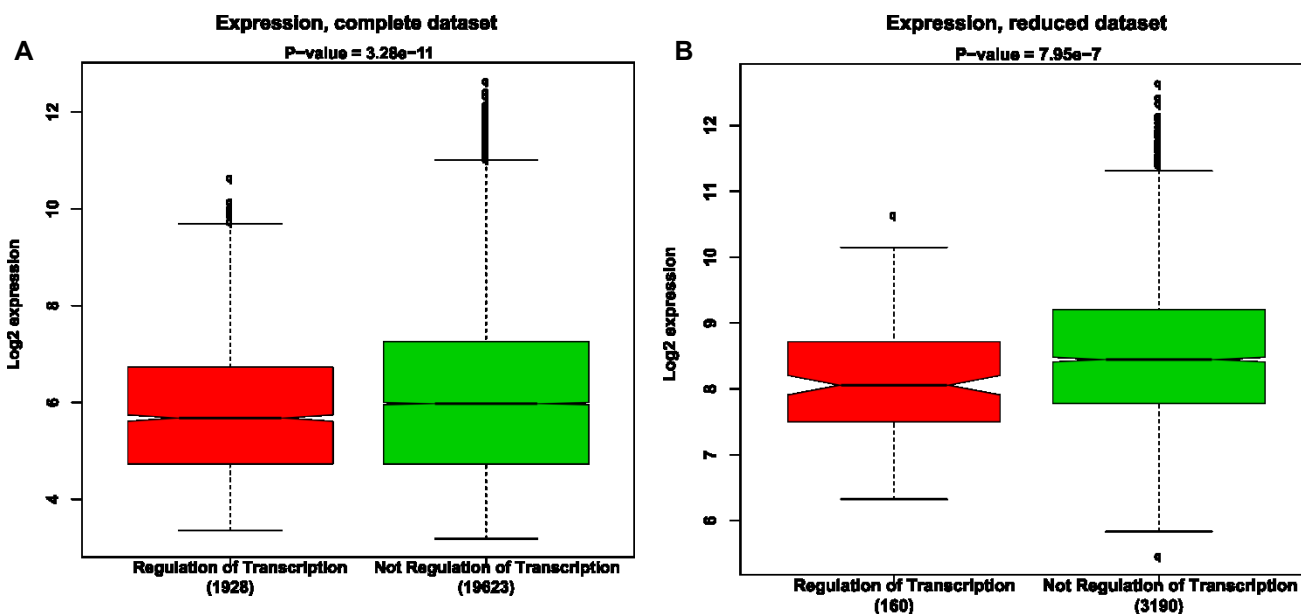
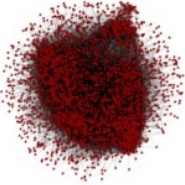
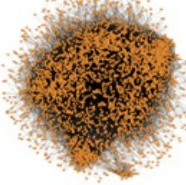
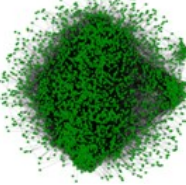
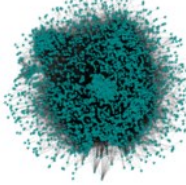


Figure 30 - Average expression intensity for Transcriptional genes (according to MapMan) in red, and not Transcriptional genes in green, in the *Arabidopsis thaliana* Affymetrix dataset discussed in this paragraph considering all transcripts (A) or only the most constantly expressed (B) according to PA 90% rule, see Paragraph 4.5

Another test for network quality investigated here is the fitting to a power law distribution of the network degree, which assesses the structure of the networks and their similarity to an expected scale-free topology. As shown in the Appendix (pages 112-114) the application of strict Partial correlation thresholds transforms the networks into highly scale-free structures, with a few hubs connected to several low-degree isolated genes. This is interesting, since it shows that Partial correlation removes edges based on a low number of gene hubs. As we showed before (Paragraph 2.3) this "conditional centrality" is a property shared by almost all essential genes. A functional enrichment analysis of these partial correlation hubs however doesn't show a clear pattern: what is deduced is that every pathway tend to have its own hub, and this hub is driving the removal of several gene connections around it through partial correlation (see Figure 21 in the previous paragraph). Furthermore, the scale-free architecture is increased together with threshold stringency in both simple Correlation and Mutual Information, therefore gene hubs intrinsically present in expression data. The fit to a power law distribution of the network degree can also be used to assess the collapse of a network structure due to a too high stringency (see for example the Mutual Information networks described in the appendix, page 113, with M.I. score threshold higher than 0.5). In these architectural considerations LASSO still shows a scale-free topology, however lower than similarly sized networks obtained with the competitor methods ($R^2=0.13$, Table 6).

| |  Pearson Correlation thr=0.4 |  Partial Correlation thr0=0.2 thr1=0.1 |  Mutual Information thr=0.2 bins=18 |  LASSO |
|--|--|--|--|--|
| Pearson Correlation thr=0.4 | | 40.9% | 33.3% | 6.9% |
| Partial Correlation thr0=0.2 thr1=0.1 | | | 20.1% | 9.8% |
| Mutual Information thr=0.2 bins=18 | | | | 5.8% |
| Table 7 - Percentage overlap between similarly sized expression-based <i>Arabidopsis thaliana</i> gene networks (described in Table 6) | | | | |

It is therefore interesting to investigate the true nature of the LASSO network, despite its lower "quality" (as calculated by the methods in Table 5), and therefore we decided to keep this as it is, also given the preliminary positive results obtained with a particular LASSO gene model, using the gene RHM2 as bait (Usadel et al., 2004), discussed in the next Paragraph. The collection of the lowest-error LASSO individual gene models was therefore compared to networks provided by the other methods, having similar overall sizes. The thresholds applied on the similarly sized networks are not strict (e.g. for Pearson correlation $r_0=0.4$), which allows for a comprehensive assessment of the overlap between all methods, although at the cost of a higher chance for "false" gene connections. An intersection between the four methods shows an unexpected low overlap between the LASSO and the others (Table 7), even lower than between Correlation and Mutual Information (despite the fact that both Pearson Correlation and LASSO are intrinsically based on linear relationships).

It is interesting to note, though, that adding a LASSO filtering to an intersection of all the networks in Table 7 significantly increased the quality of the resulting intersection when compared to random edge removal (p-value <0.001 by permutation test, valid for Ontology Agreement, overlap to Protein-Protein interactions and Degree Distribution tests). This observation hints that the LASSO can be additive and complementary to the ordinary reverse engineering methods. In order to understand what's the nature of the LASSO uniqueness, it is mandatory to summarize the information contained in large gene-gene networks.

The intersection network between the networks shown in Table 7 (including the LASSO), shown in Figure 31A was therefore annotated to functional categories (Figure 31B) and from this a network of significantly interacting categories was extracted (Figure 31C). This network can be considered as a condensed version of the original gene-gene expression based network, where co-regulation can be seen at work between pathways and groups of genes rather than between individual transcript behaviors. In Figure 31D, a diagram of the most interacting categories is shown as an example, these interactions, described in clearer detail in Figure 32A. The ribosomal proteins show an extremely strong co-regulatory behavior and, through chloroplastidic ribosomes, they are strongly connected to genes involved into the light reactions of

photosynthesis. Ribosomal genes are in turn strongly co-regulating with the Tricarboxylic acid Cycle and with the basal mechanism for RNA processing and transcription. Impressively enough, these backbone category connections are so evident and conserved that they are found by all reverse engineering methods, also in *Oryza sativa* (data not shown).

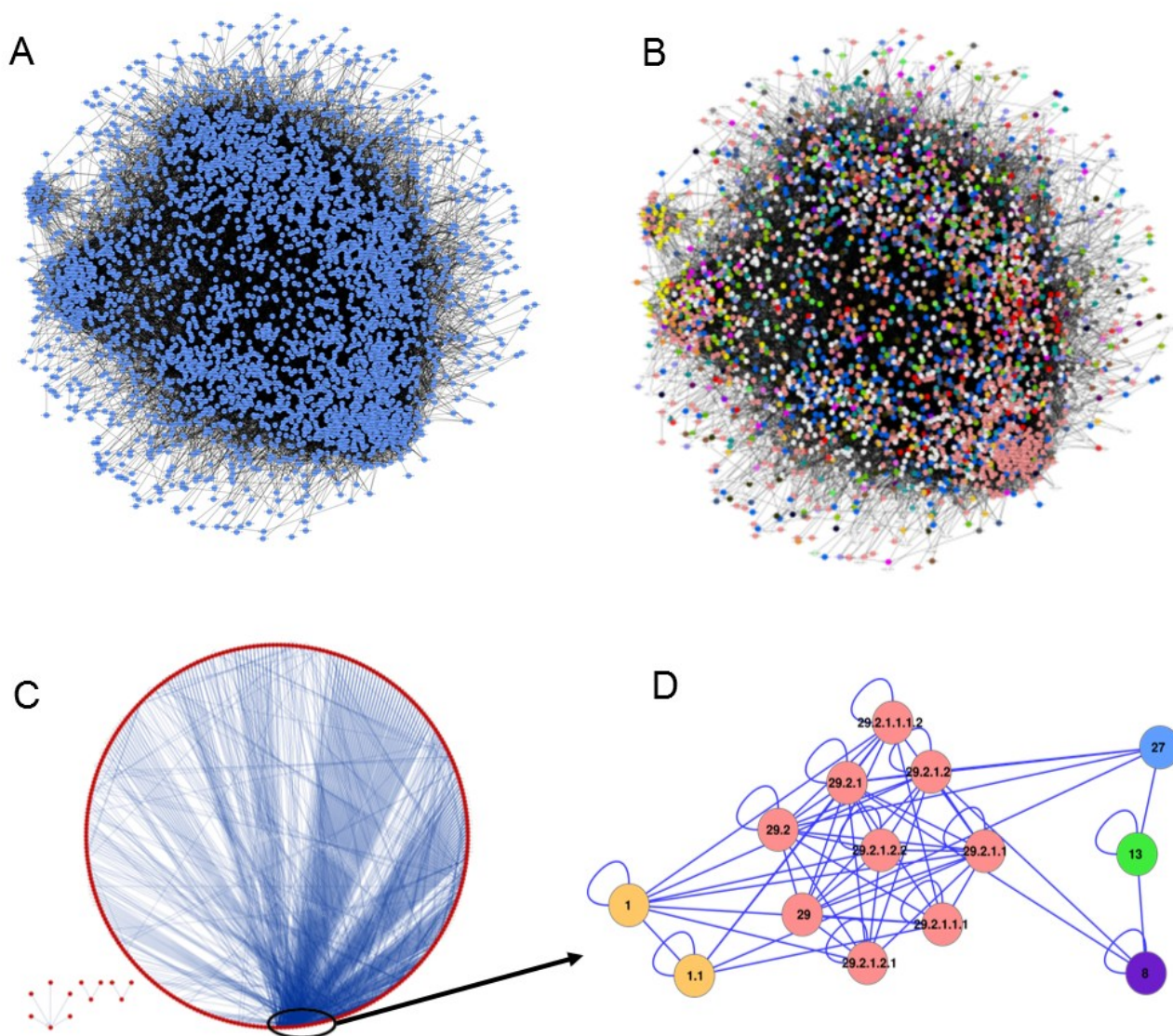
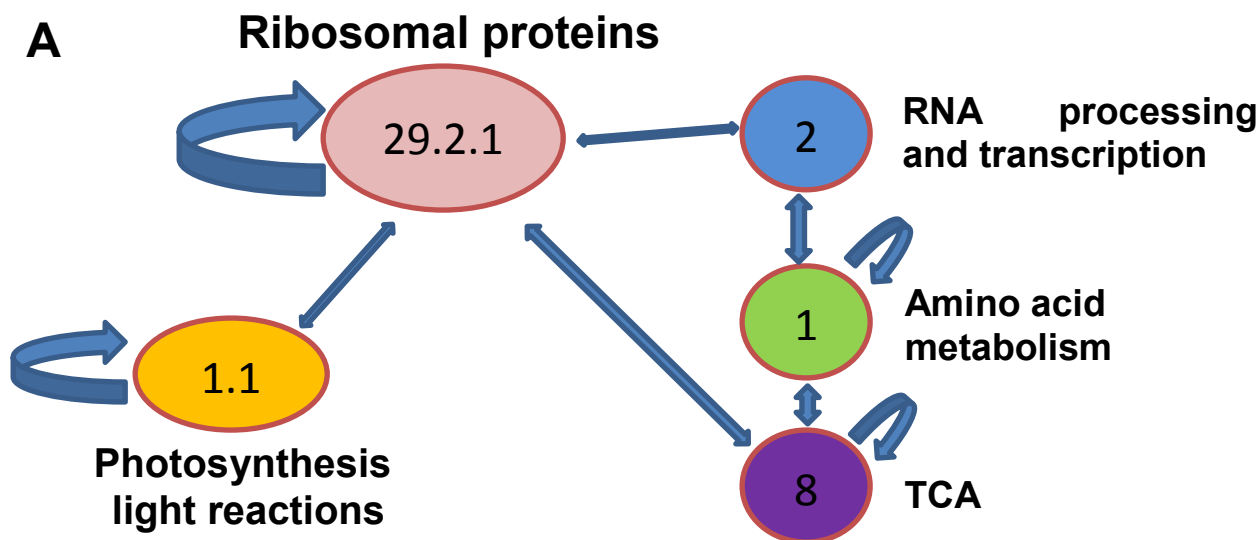


Figure 31 - (A) Intersection of Pearson Correlation ($r=0.4$), Mutual Information (normalized coefficient=0.2, bins=18) and LASSO expression-based gene networks for *Arabidopsis thaliana*. (B) ontology mapping of the network in panel A, associating every gene to one MapMan bin (Usadel et al., 2009). (C) Significantly enriched connections between MapMan bins as mapped in panel B (p -value <0.05). (D) Simplified diagrams of the most connected MapMan bins in panel C.

It is puzzling, however stimulating, to notice that in fact every method is able to extract, even independently from the threshold, a backbone of "easy" gene-gene relationships, accompanied by a set of particular

connections. For example the LASSO, in its own area, is able to find several significant category associations that are not found by the other methods analyzed here (in Figure 32B the most connected categories are shown). Specifically, it seems to be able to discern the mechanisms in charge of protein post-translational modification and transport. Ubiquitination, which is principally a mechanism to drive protein degradation (Scheffner et al., 1995), is deemed by the LASSO to be significantly co-regulated with post-translational modification mechanisms, such as phosphorylation; it is in fact well known that phosphorylation and ubiquitination are acting parallelly in determining specific protein activity and turnover (Karin and Ben-Neriah, 2000), such as in the cell cycle (Lodish et al., 2003), and this, as also detected by the LASSO, is requiring an active relocalization of the proteins themselves to the proteasome degradation complex (Glickman and Ciechanover, 2002). It must be reminded that the diagram illustrated in Figure 32 is only depicting the most connected and significant interactions, and that a full characterization of the specificity of these methods is still in progress.

Found by all algorithms



LASSO-specific

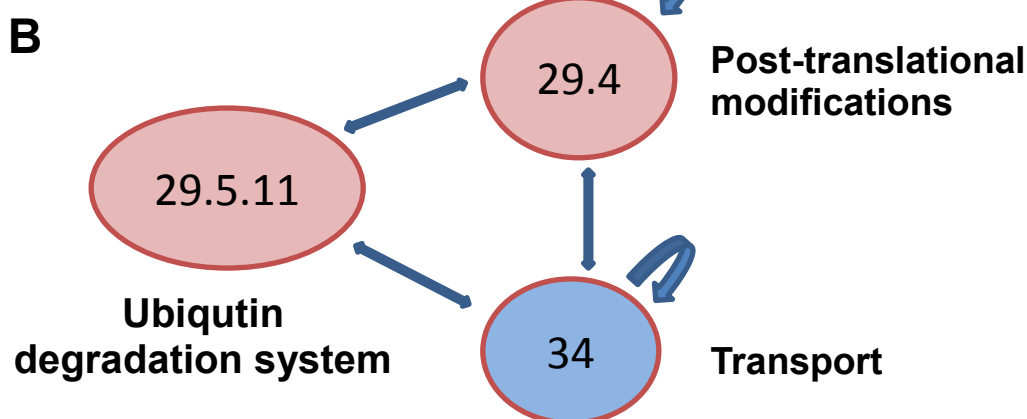


Figure 32 - Schematic representation of the most enriched ontology connections (see also Figure 31D) in *Arabidopsis thaliana* expression-based networks found by all expression-based network reconstruction methods (A) and by LASSO only (B).

In conclusion, the messages that can be drawn from this Bioinformatics analysis are two, at least for what concerns the LASSO. First, that this method is able to grasp significant and meaningful interactions between genes based on expression data, although with general lower performance than Pearson correlation. Second, that the LASSO is largely complementary to other network reverse engineering methods, yielding a complete overlap in evident co-regulation structures, such as between ribosomal genes, but providing particular conclusions on other pathways, such as protein ubiquitination. However in science, and especially in Bioinformatics, it is always unwise to draw pure theoretical hypotheses without testing them. Therefore I decided to test the LASSO in real biological scenarios, described in the following two paragraphs, together with other Correlation-based approaches, to test if this "complementarity" were real or artifactual.

2.5 LASSO and correlation for reverse engineering the seed coat mucilage pathway in *Arabidopsis thaliana*

2.5.1 RHM2 expression network analysis

The Rhamnose Biosynthesis 2 gene (*RHM2*, also known as *MUM4*) is one of the few genes with a defined biosynthetic molecular function (UDP-L-rhamnose synthase) in the mucilage synthesis gene pathway(s) (Figure 6) (Usadel et al., 2004; Oka et al., 2007). The direct regulators of *RHM2* however have not been characterized yet, and several missing links are present in this molecular process (Arsovski et al., 2009; Huang et al., 2011). Therefore, we decided to rely on a guilt-by-association *in silico* screening using *RHM2* as gene bait, and all genes measured by the Affymetrix ATH1 as explanatory variables (see Methods, Paragraph 4.6.1). In order to be as comprehensive as possible, we decided to use all *Arabidopsis thaliana* samples available in the public repositories Gene Expression Omnibus (Edgar et al., 2002) and ArrayExpress (Parkinson et al., 2007), obtaining a final dataset composed of 5750 samples and 21000 genes (Paragraph 4.6.1). In such a scenario, fitting a complete LASSO model is not computationally feasible. We therefore decided to simplify the procedure, by including only the 2000 genes with highest absolute Pearson correlation to *RHM2* in our model. As a modeling solution we used the LARS algorithm (Efron et al., 2004), which allows for a relatively quick LASSO model generation and therefore makes a model of this kind computationally tractable.

The result is a rather complex collection of weights for several gene variables, obtained with varying L1 constraints to the sum of variable weights. The reader can appreciate this complexity in the Appendix (page 121, Figure 61), where a plot is depicted indicating the weight trend of the explanatory genes at different L1 thresholds for the *RHM2* model. The path of the LASSO model generation, which is developed starting from L1 equaling zero, is quite complex, and at the right end (no constraint) it ends up assigning a weight to every gene. However, by focusing on stringent constraint regions we see that only a few variables are included in the model. For example, at L1=1% (percentage calculated over the sum of weights with no constraint), only 10 predictor variables were selected. This list is partially overlapping with the top 10 correlators (using Pearson correlation), but three genes are already unique to LASSO. As we will see below in greater detail for a reduced *RHM2* model, the candidates selected by Pearson correlation and the LASSO tend to diverge exponentially as the L1 is increased.

In order to have the widest analysis scope, we generated a list of all genes included by the LASSO algorithm at L1=1%, 2%, 3%, 4% and 5%. To these, we added for technique comparison the top 30 candidates obtained by Pearson Correlation, first order Partial Pearson correlation, full-order Shrunken Partial Correlation (using the approach by (Schäfer and Strimmer, 2005)) and top 30 weighted variables from a LASSO model with no constraint (L1=100%) (see Methods, Paragraph 4.6.1). We selected genes having readily available knock-out insertion lines (Alonso et al., 2003), and ended up with a short list of 38 candidates. In order to experimentally confirm the involvement of these genes in the polysaccharide synthesis of the pectinaceous mucilage, a sugar analysis of the knock-out lines was performed, as described in Paragraph 4.7, in order to assess differences at

the monosaccharide level (I thank Mr. Aleksandar Vasilevski - Max Planck Institute of Molecular Plant Physiology - for conducting these measurements). The content of Rhamnose, Arabinose, Galactose, Glucose, Xylose, Mannose and Galacturonic Acid was quantified in the seed coat mucilage of the lines corresponding containing the 38 gene knock outs. Especially for Rhamnose and Galacturonic Acid (Figure 33) a significant divergence from the Columbia-0 wild-type content often was associated with peculiar staining pattern of the seed coat mucilage. As an example, the low Galacturonic acid mutant *GH9C2* (knocked out in the gene glycosyl hydrolase 9C2, corresponding to the *Arabidopsis* locus At1g64390, found in our approach by Pearson correlation), resulted in a tight, shell-shaped mucilage structure, rather different than the flocculating wild-type pattern and shows less than half Galacturonic Acid content when compared to the wild-type (Figure 33). Interestingly, this gene has been characterized and patented as a novel modulator of cellulose cristallinity in 2010.

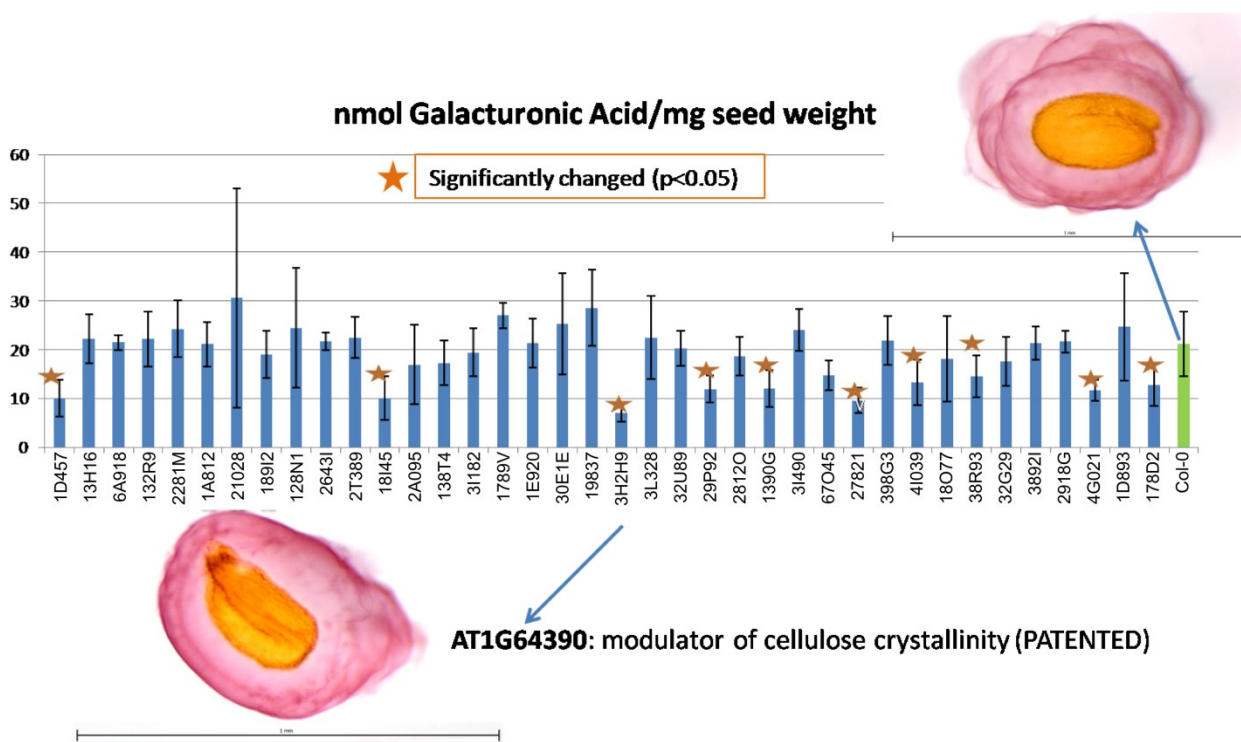


Figure 33 - Galacturonic acid content in the seed coat mucilage of part of the knock out lines extracted via the LASSO and Correlation approaches using *RHM2* as a gene bait. The picture of a wild-type Ruthenium red seed is showed (top right) and compared to the At1g64390 KO line (bottom left). Lines varying significantly more than the wild-type (T-test, p -value < 0.05) are marked with a star. Sugar measurements performed by Aleksandar Vasilevski (MPIMP)

The analysis of the KO lines for the 38 genes, obtained by the four partial overlapping methods, showed us a total of 11 genes with a potential sugar phenotype (Figure 34). The complementarity of these techniques is evident, for example the LASSO provides three candidates that are outside the top100 ranked lists for any other Correlation method. This fact strengthens the need for a complementary approach we took for our Bioinformatics candidate selection. The genes found by these methods are in part unknown and uncharacterized, in part possessing functions already known to be involved in polysaccharide processes (such

as the already cited At1g64390). The nucleotide-rhamnose synthase/epimerase-reductase (NRS/ER, locus At1g63000), found by LASSO and Partial methods, but not by Pearson correlation, is of particular interest, since this enzyme is directly involved in the synthesis of UDP-L-Rhamnose (Watt et al., 2004), a principal building block of mucilage polysaccharides. A in-depth investigation of the functions of these eleven putative novel members of the mucilage pathway is currently under progress. Preliminary data confirm also a lack of mucilage synthesis and/or release for a number of the genes in Figure 34. However we will omit in this dissertation a detailed description of these genes, especially the unknown ones, in order to preserve the novelty of the discovery prior to publication.

| Candidates selected using RHM2 – full dataset | | | | |
|---|----------------|----------------|-----------------------------|----------------------|
| | Genes selected | Genes screened | Genes with sugar alteration | Ratio pheno/screened |
| Pearson Correlation | 30 | 8 | 3 | 0.375 |
| Partial Correlation | 30 | 11 | 7 | 0.636363 |
| Shrunk Correlation | 30 | 10 | 7 | 0.7 |
| LASSO | 85 | 24 | 9 | 0.375 |

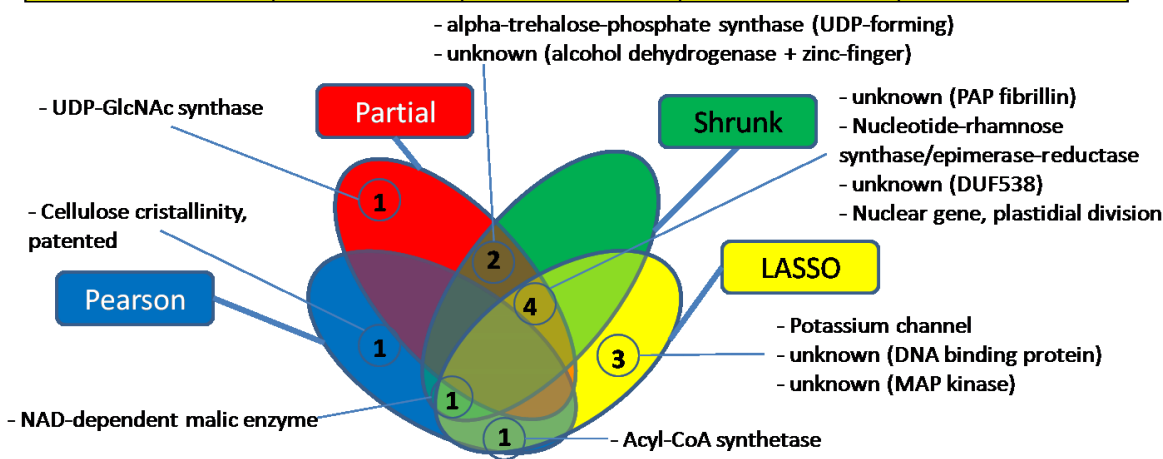


Figure 34 - Summary of the genes having confirmed and significant monosaccharide (Rhamnose, Arabinose, Galactose, Glucose, Xylose, Mannose or Galacturonic Acid) levels alterations in the seed coat mucilage when knocked out

2.5.2 Network reconstruction based on several mucilage genes

The *RHM2* LASSO models, although rich in information, mask two of the most important features of LASSO: they are not easy to interpret (a lot of variables are included, see Figure 61) and they don't exploit LASSO's ability to infer information when the number of samples is much lower than the number of variables.

Therefore, we decided to use a smaller dataset, focusing on samples where seed coat mucilage is effectively synthesized and released, and where *RHM2* is known to be expressed, together (hopefully) with its functional partners. We selected the Affymetrix AtGenExpress seed and silique developmental series samples (GEO accession: GSE5634, composed by 24 samples)(Schmid et al., 2005). By reducing the number of samples,

we could include all 20000 genes measured by the Affymetrix array (not, as before, only the "top correlators") and keep the modeling computationally tractable. The models generated by a number of samples so small are also including less variables and are therefore simpler to interpret. This happens because the model will stop exploring the variable space when a number of predictor variables equalling the number of samples minus 2 has been included. At this stage, the model cannot proceed without becoming underdetermined.

In order to increase the range of our analysis, we generated a model for each of the 11 genes known to be involved in mucilage production and/or release by knock-out experiments (Table 8). Half of this list is composed of transcription factors, *i.e.* ideally upstream regulators, and half of enzymes, therefore putative mucilage synthesis effectors (see also Figure 6 in the Introduction for a map of known mucilage pathway genetic interactions). It must be noted that, when we started this analysis, the analysis confirming the transcription factor *LUH* as a mucilage-deficiency gene had been not published yet (Huang et al., 2011), and therefore *LUH* and its targets were not included in our analysis.

| Gene | Function |
|---------------|--|
| RHM2 | UDP-L-Rhamnose synthase |
| MUM2 | β -Galactosidase |
| ARA12 | Subtilase |
| Myb61 | Transcription factor |
| AtBXL1 | β -D-Xylosidase/ α -L-Arabinofuranosidase |
| GAUT1 | α -1,4-galacturonosyltransferase |
| Myb5 | Transcription factor |
| TTG1 | Transcription factor |
| AP2 | Transcription factor |
| GL2 | Transcription factor |



Table 8 - List of genes known in literature to have experimentally confirmed roles in seed coat mucilage synthesis and/or release. The *Myb5* mutant seed (impaired in mucilage release) is shown after Ruthenium Red staining (see Paragraph 4.7.1). The release of the mucilage for this mutant can be triggered by mechanical stress (see Appendix, page 122)(Arsovski et al., 2009)(Arsovski et al., 2009)[137]

Initially, I explored different ways to extract "gene candidates" from the variables used by the LASSO model generation.

1) The first is a simple R^2 threshold, which contemplates stopping the LASSO modeling when the sum of residual errors of the model reaches a certain threshold. However, in this way initial variables may be excluded for good at a certain LASSO point (e.g. black line Figure 37 for *AP2*, which corresponds to At1g45474, a component of the light harvesting complex of photosystem I; this gene is the top correlator of *AP2*, but it is excluded as not significant after a few LASSO steps).

2) The second model selection method is constituted by a L1 bound threshold. The same method used for the MUM4 LASSO model of the previous paragraph. As in using an R^2 threshold, it is affected by a certain arbitrariness and may discard variables included only at certain early L1 ranges.

3) A very popular model assessment concept is the Akaike's Information Criterion (AIC) commonly used to measure the goodness of fit when adding an explanatory variable to a model. It is calculated by the formula:

$$AIC=2k+n[\ln(RSS)]$$

Where k is the number of variables, n the number of observations (samples) and $\ln(RSS)$ is the natural logarithm of the residual sum of squares of the model (Akaike, 2002). When comparing multiple models, the one having the lowest AIC is considered as the best one, meaning the one with the most tuned balancing between simplicity and goodness of fit. However in low-number of samples scenarios, the models including all variables are, in all cases explored here, the ones with the lowest AIC: inclusion of a variable, since only a few are allowed before reaching an undetermined system, is always beneficial for the model. We explored a high-number of samples dataset, and saw that for up to 500 explanatory variables, every subsequent LASSO model has a better AIC than the previous ones (Figure 35A and Figure 35B). These models, used here as explanatory, show also that the LASSO tends to include more variables rather than readjusting the weights of precedently included ones. It is interesting to compare Figure 35C with the model described in the Appendix (page 121), where the growth of the model, driven by the increase of the L1, massively increases the number of genes included.

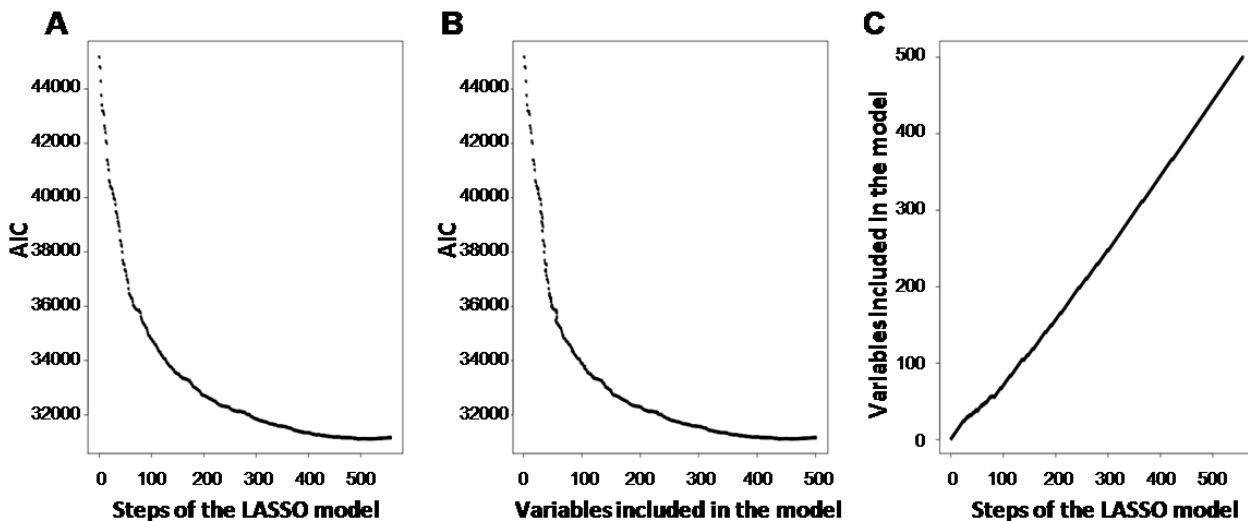


Figure 35 - Panel **A**: AIC plotted against the number of steps taken by the LASSO *RHM2* model in an *Arabidopsis thaliana* dataset composed of 5750 samples (see Paragraph 4.6.1). Panel **B**: scatterplot of AIC score versus the number of variables included at any model. Panel **C**: number of variables in the model at any LASSO step. The tendency of LASSO in these conditions is to include variables, although occasionally dropping steps are taken. Conceptually identical results as in Panel A and B can be obtained using Cross-Validated model error instead of AIC (not shown)

4) A best cross-validated model selection, already discussed and applied in Paragraph 2.4.3, is arguably the best option, as it would provide the least unbiased assessment of robustness for each L1 and provide the comparative best model (Efron and Tibshirani, 1995). The only drawback of this method is that it could theoretically exclude several steps taken by the LASSO modeling and exclude genes that have been deemed highly relevant at a non-best cross-validated L1. Furthermore, blindly accepting the lowest error model doesn't account for the possibility of local *minimi* in the cross-validate error during LASSO progression. Although ignoring these genes would be, statistically, the correct way to proceed, with models as small as the ones described in this paragraph it is possible to extend this gene selection a bit further.

5) An approach including the best crossvalidated model selection allows the inclusion of all variables that participated in the LASSO modeling, and were at some point included in the model before reaching the final solution. Likely, all the variables included will show a certain degree of relationship with the response variable, which makes them some of the best predictors during the modelization. In our case of 24 samples, the LASSO will stop at a model with 22 variables (plus the intercept), and has no way to improve without falling in an underdetermined situation (basically, a system of equations with multiple solutions). We decided to include these variables plus all the variables which were included during the LASSO steps. This variable inclusion approach is evidently unfeasible when the number of variables included is too high, such as in a dataset composed of thousands of samples (Figure 61). In a holistic perspective, experimental understanding of the mucilage signal transduction and synthesis pathway will help us understanding what the best option for variable selection would have been, and if effectively the lowest-error cross-validated model (point 4) is the "best" one.

It is clear from the weight plots that these models are much more interpretable than the one generated in the previous paragraph (Figure 61). Only a few genes are included in the *RHM2* model now, and the progression of their importance in the models at different L1s is analyzable in a less noisy situation (Figure 36). For *RHM2*, three genes only appear to have high importance in early models, namely At5g63800 (MUM2), At1g10760 (SEX1, an α -glucan, water dikinase required for starch degradation) and At2g37090 (IRX9, a putative xylosyl transferase involved in xylan biosynthesis in secondary cell wall) - these are represented respectively by the black, red and cyan leftmost lines in Figure 36. This is encouraging, since all these genes have been connected to polysaccharide synthesis (Gómez et al., 2006) (Bauer et al., 2006) (Dean et al., 2007).

The situation for a multi-purpose transcription factor, *AP2*, is quite different, and the weight plot looks more entangled. Maybe this derives from the fact that *AP2* has several different functions, being a hub in the Arabidopsis developmental network (Okamuro et al., 1997) and with a lower number of clear coexpressor, having for example 2 neighbors with Pearson correlation coefficient higher than 0.9, compared to the only 622 of *RHM2* (Figure 37).

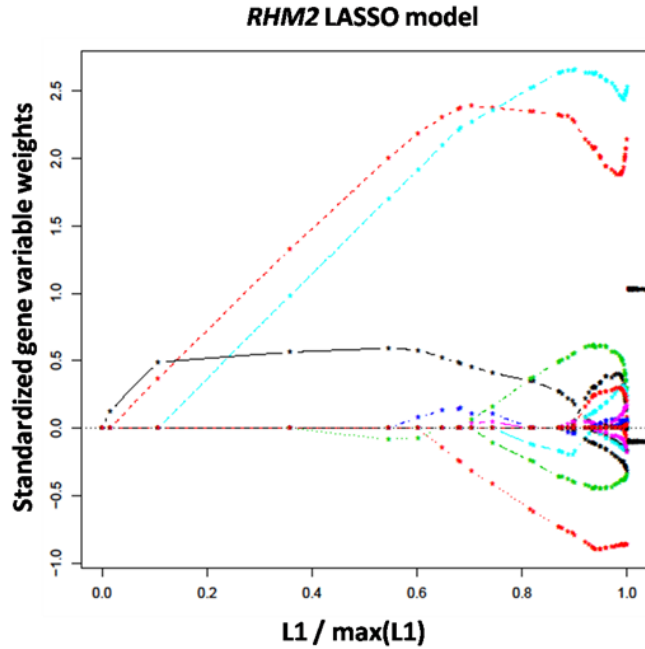


Figure 36 - LASSO model weight plot for the *RHM2* models based on the GSE5634 seeds/siliques *Arabidopsis thaliana* microarray dataset (Schmid et al., 2005). On the x axis, varying sum of variable weight constraints, on the y axis, the weights for every gene predictor. Every line corresponds to a gene included as prediction variable for *RHM2*

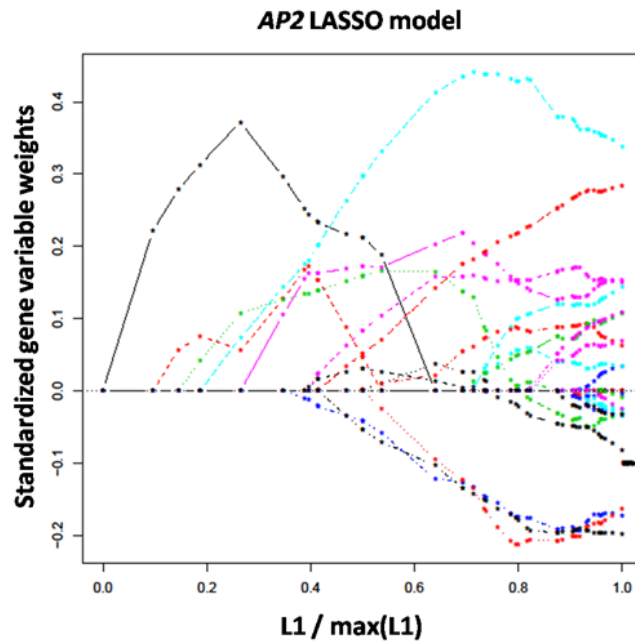


Figure 37 - LASSO model weight plot for the *AP2* models based on the GSE5634 seeds/siliques *Arabidopsis thaliana* microarray dataset (Schmid et al., 2005). On the x axis, varying sum of variable weight constraints, on the y axis, the weights for every gene predictor. Every line corresponds to a gene included as prediction variable for *AP2*

It is interesting to notice that also in this case Pearson Correlation and LASSO differ on candidate selection, while giving the same results for the most correlated variables. Here's for example a list of all variables appearing in the *RHM2* model, and their Pearson correlation rank (Table 9).

| Rank of appearance in LASSO | Pearson rank |
|-----------------------------|--------------|
| 1 | 1 |
| 2 | 2 |
| 3 | 3 |
| 4 | 5 |
| 5 | 6 |
| 6 | 11 |
| 7 | 12 |
| 8 | 20 |
| 9 | 27 |
| 10 | 33 |
| 11 | 34 |
| 12 | 36 |
| 13 | 46 |
| 14 | 65 |
| 15 | 85 |
| 16 | 107 |
| 17 | 116 |
| 18 | 138 |
| 19 | 172 |
| 20 | 193 |
| 21 | 202 |
| 22 | 205 |
| 23 | 231 |
| 24 | 267 |
| 25 | 310 |
| 26 | 367 |
| 27 | 455 |
| 28 | 915 |
| 29 | 928 |
| 30 | 1372 |
| 31 | 1464 |
| 32 | 1829 |

Table 9 - Rank order of appearance for genes in the LASSO *RHM2* modeling (left) and absolute Pearson correlation coefficient rank for *RHM2*.

We merged the models obtained by the eleven "bait" genes using our candidate selection approach discussed before, and obtained a final "mucilage network" specific for seed and silique tissues, shown in Figure 38. Four of our "bait genes" (*AP2*, *GAUT1*, *GAUT11*, *MYB5*) appear disconnected from the main network component. This doesn't necessarily mean that they are involved in different pathways, but it may be that they are not transcriptionally regulated and/or (such in the case of *AP2*) they have too many interactors to be found by our approach. The giant component of Figure 38 shows an interesting network of common predictors. Most importantly, *MUM2* and *RHM2* appear in each other's LASSO short-lists, together with three other shared genes. These three genes are a phosphoglyceride transfer family protein), an ATPase) and a putative Galacturonosyltransferase (*GAUT*), specifically a predicted polygalacturonate 4-alpha-galacturonosyltransferase. It is tempting to consider this *GAUT* in particular as a potential "missing link"

between these two pathways: this enzyme has been only hypothetically linked to polysaccharide synthesis by sequence homology, but its specific pathway is still unknown.

Another interesting gene is the linker between *GL2* and *RHM2*: At1g05230 (Encodes a homeobox-leucine zipper family protein belonging to the HD-ZIP IV family). In general, among the *RHM2* and *MUM2* neighbors, we had the impression of a population split between proteins functionally active in sugar metabolism or involved in signal transduction.

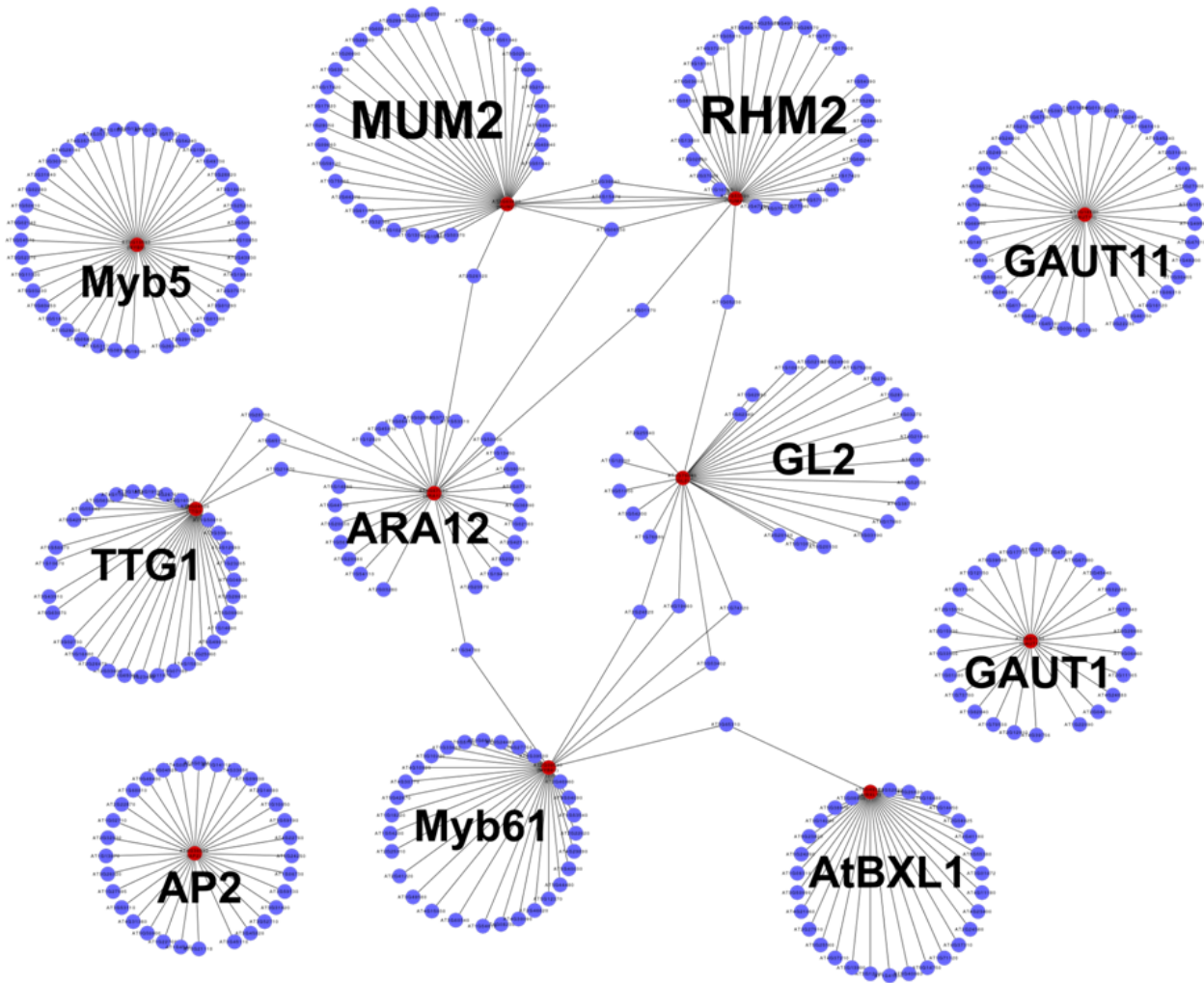


Figure 38 - Joint network representations of the eleven LASSO models based on gene baits known to be involved in the seed coat mucilage synthesis and release pathway. The models generated using the GSE5634 seeds/siliques *Arabidopsis thaliana* microarray dataset (Schmid et al., 2005)

Preliminary screening of the LASSO networks show promising results. In this case, there are not only many mutants having less extractable Galacturonic Acid and Rhamnose, but lack of mucilage phenotypes are clearly visible. For example, in Figure 39, we show the appearance of two Knockout *Arabidopsis* mutants after coloring their seeds with Ruthenium Red (see Methods, Paragraph 4.7.1). The first is found in the neighborhood of *GL2* and it shows lack of mucilage in two independent KO lines (Figure 39B and Figure 39D).

The second is introduced in the LASSO models for *Myb61* and also shows a complete lack of mucilage phenotype (Figure 39C). These seeds are not releasing the mucilage even after mechanical stress (Figure 39E for the *Myb61* neighbor and Figure 39F for the *GL2* neighbor), and therefore it can be hypothesized that the very synthesis of mucilage pectins is impaired in these mutants. In the Appendix (page 122), we show how mechanical stress can indeed induce the release of mucilage in some mutants, for example the *Myb5* knockout (where also EDTA treatment can trigger mucilage release (Arsovski et al., 2010)). This hints at the existence of two separate pathways: one for mucilage synthesis, one for mucilage release upon seed hydration. It is also interesting to add that all these knockout lines, as far as we could see, seem to show a wild type-like phenotype during their adult life (data not shown). Surely, the detailed characterization of these knockout lines will be helpful in understanding the true function of these novel mucilage genes, and to possibly place them in the still incomplete picture of the seed coat mucilage synthesis and release processes (Figure 6).

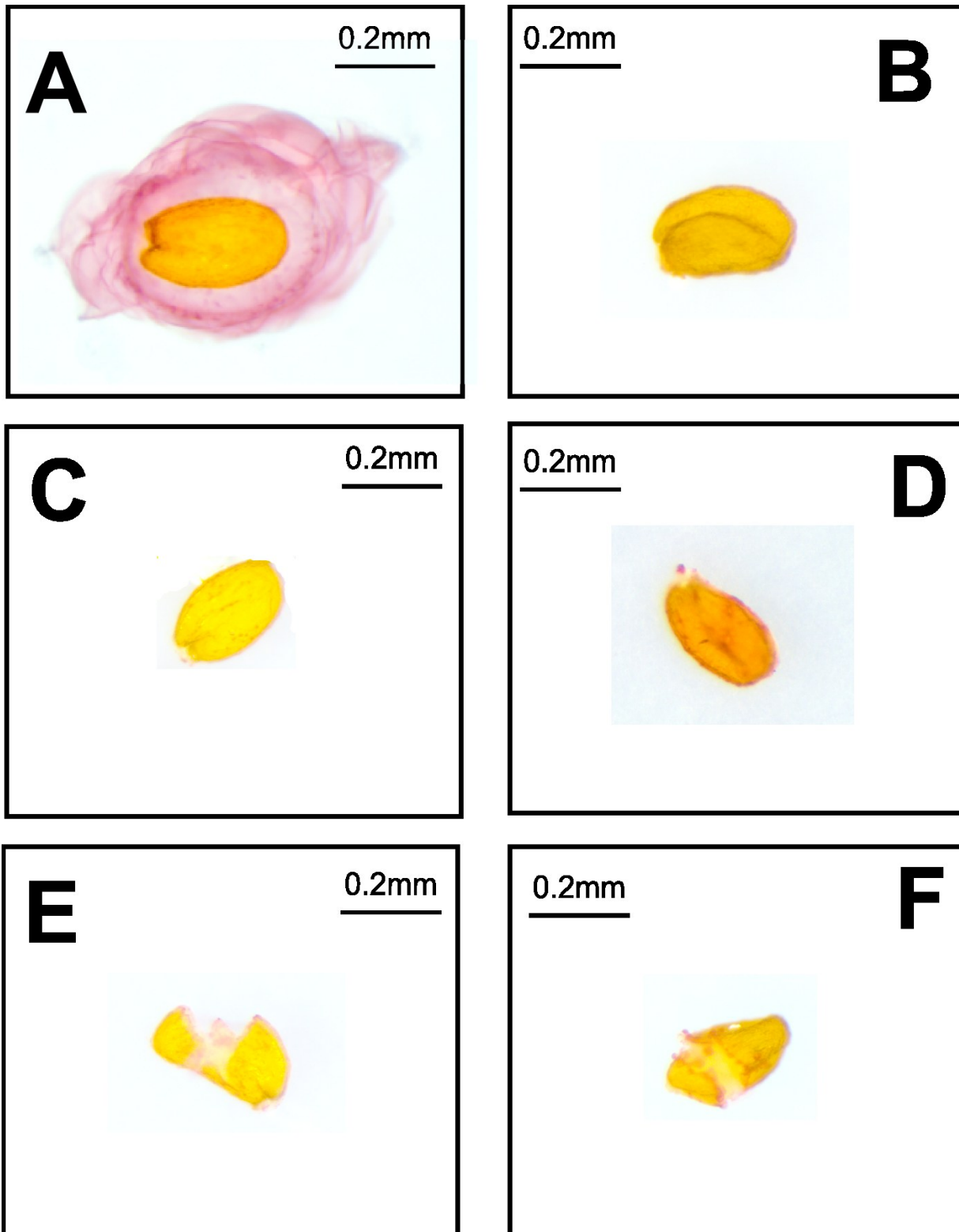


Figure 39 - Ruthenium Red staining of several *Arabidopsis thaliana* genotypes. (A) wild-type Columbia 0. (B) and (D) two independent knockout lines for a *GL2* network neighbor. (C) knockout line for a *Myb61* network neighbor. (E) *Myb61* network neighbor knockout line after mechanical stress. (F) the same *GL2* neighbor knockout line as panel B, after mechanical stress.

2.6 LASSO and correlation for reverse engineering the hypoxia-regulated tuber development pathway in *Solanum tuberosum*

In the previous paragraph it was shown how, based on one or more gene "baits", gene network reconstruction can be carried out following a simple guilt-by-association approach, identifying genes co-regulated in similar ways. However, in the scenario tackled here, the *Solanum tuberosum* hypoxia-regulated tuber development, none of the genes involved has been characterized yet. Therefore, in order to compare LASSO and Correlation, it was first necessary to *find* the bait genes needed for expression-based analyses. In the following paragraphs (2.6.1, and 2.6.2) I will describe how three transcripts involved in hypoxia and tuber development were found in *Solanum tuberosum* via knowledge transfer from *Arabidopsis thaliana*, and characterize their expression pattern. In the subsequent paragraphs (2.6.3 and 2.6.4) I will focus on a comparative LASSO-Correlation expression-based candidate selections using these transcripts as baits.

2.6.1 Identification of hypoxia responsive ERFs in *Solanum tuberosum*

To isolate ERF-encoding genes that are responsive to low oxygen conditions, cDNA from leaf and root tissues excised from hypoxic and aerobic *Solanum tuberosum* cv. Désirée plants were amplified using degenerated primers designed to specifically anneal to members of ERF transcription factors group VII (Table 14). A single band was obtained (Figure 40) that corresponded to three different fragments putatively encoding ERF VII proteins (Figure 41).

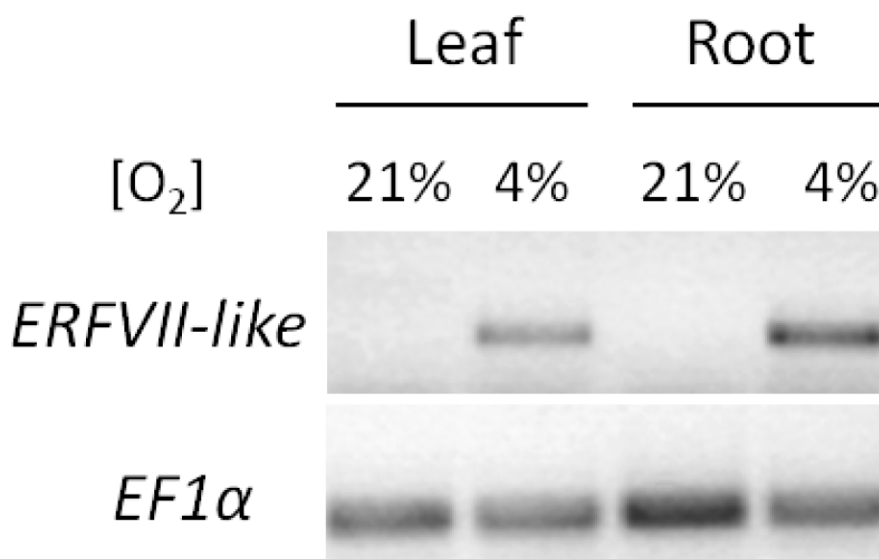


Figure 40 - Semi-quantitative Reverse Transcription PCR on potato leaves and roots excised from plants treated at 21% or 4% (v/v) oxygen for 3 h using ERFVII-specific degenerated primers. EF1 α was used as housekeeping gene to normalize loading

Two of the fragments shared over 90% identity, differing only in a few codons located in the N-terminal part. These two sequences map uniquely to the same locus on the tomato (*Solanum lycopersicum*) genome (release SL2.40, chromosome 9 62625165:62625529, E-value 1×10^{-63}), and correspond to a single Unigene tomato model (Gene Bank entry AY192368, E-value 3×10^{-85}). Also in the recently released genome of

Solanum tuberosum (Xu et al., 2011) both sequences map the same locus (scaffold PGSC0003DMS000001201, genome assembly v3, E-value 4×10^{-59}). Therefore, these two sequences correspond most likely to different splice variants. BLAST-analysis of the deduced amino acid sequences against the UniProtKB database (Schneider et al., 2009) revealed that one fragment corresponds to a *CIP353*, an ERF gene already identified as moderately cold inducible (Mine et al., 2003), and the two remaining are identical to the DNA binding protein STWAEIRD (Campbell et al., 1998), which is 298 aminoacids long. Full sequencing of the amplicons provided the full length of these three genes (see Appendix, page 107), which were renamed in accordance to their *Arabidopsis* homologues: *Hypoxia Responsive ERF1* (*StHRE1*) and *Hypoxia Responsive ERF2* (*StHRE2a* and *StHRE2b*).

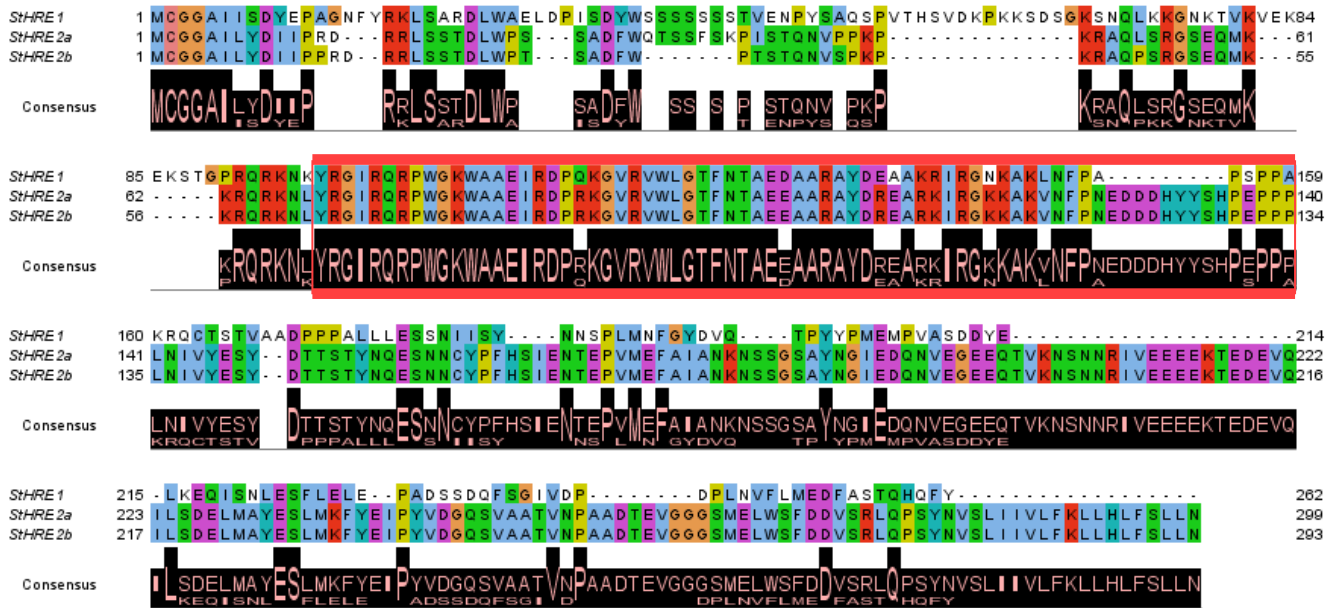


Figure 41 - Alignment of the protein sequences of the potato ERFVII-like proteins identified. Amino acidic sequences were aligned using MUSCLE (Edgar, 2004). The multiple alignment is visualized using JalView (Waterhouse et al., 2009). In the red box, the AP2/ERF domain is highlighted, as automatically annotated by SMART (Letunic et al., 2009)

Analysis of the deduced amino acid sequences indicated that all three HREs contain a highly conserved AP2/ERF domain (Figure 41, red box). Using the WoLF PSORT web-tool (Horton et al., 2007) all HREs proteins were predicted to be nuclear localized. In addition, two serine/threonine-rich regions that may function as activation domains were found to reside in the N-terminal and the C-terminal regions of *StHRE1* and *StHRE2* proteins, respectively. *StHRE2b* is missing a 7 amino acid sequence SFSKPI, which is present in the N-terminus of *StHRE2a* (Figure 41).

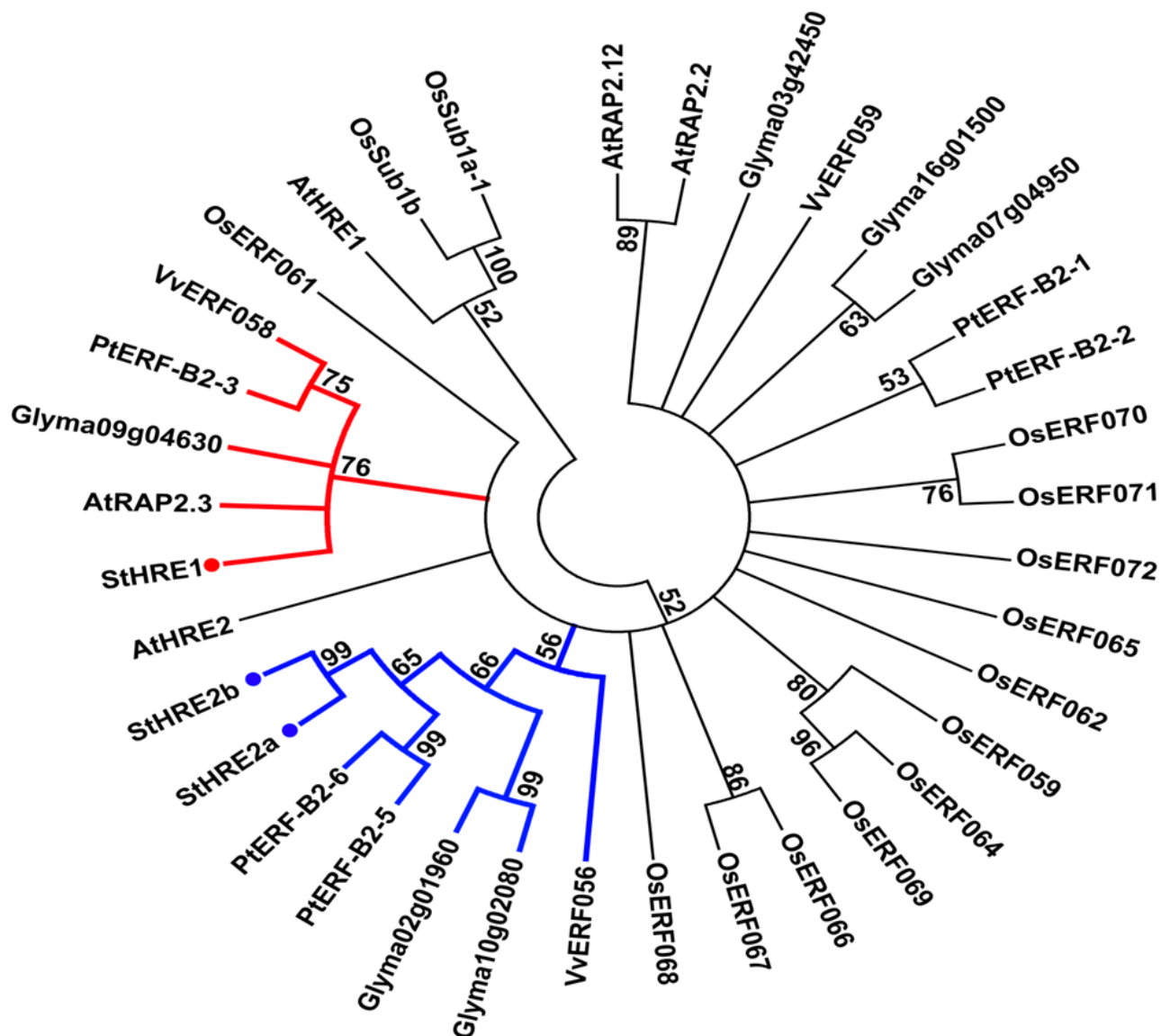


Figure 42 - Phylogenetic consensus tree of group ERF-VII proteins from different plant species whose genome has been sequenced. A maximum-likelihood tree was created with MEGA5 (Tamura et al., 2007) based on a multiple sequence alignment calculated by MUSCLE (Edgar, 2004). The distance bar is shown on the top of the tree. Numbers at each node are the percentage bootstrap value of 100 replicates. The tree is collapsed where bootstrap support is lower than 50%.

To investigate the sequence similarity of the identified HRE sequences compared with the ERF-VII proteins encoded by other representative plant species, I performed a sequence-based phylogenetic analysis. The resulting phylogenetic tree (Figure 42) shows a loose distinction of several ERF subgroups, and allows us to place the newly found potato sequences in defined subclades. In particular, StHRE1 clusters together with *Arabidopsis thaliana* RAP2.3 and HRE2 proteins, while StHRE2a and StHRE2b belong to a subclade including *Vitis vinifera* VvERF056 (Licausi et al., 2010), poplar and soybean proteins but, interestingly enough, no *Arabidopsis* protein.

2.6.2 *StHREs* expression during tuber development

Tuber size is assumed to play a major role as a constraint to the oxygen availability for the inner tuber cells (Geigenberger et al., 2000). Oxygen concentrations and mRNA levels were measured in developing tubers at different stages defined in an age-related manner according to (Kloosterman et al., 2008). The internal O₂ concentrations progressively decreased with tuber development and size-increase, to around 10% of the ambient oxygen concentration in developed, 25-day old, tubers (Figure 43). Concomitantly, the potato Sucrose Synthase 4 (*StSus4*) mRNA increased, achieving very high expression levels when tubers reached a volume of about 3-4 cm³ (Figure 43). Sucrose synthase is used as an hypoxia marker since the enzyme encoded by this gene, synthesizing UDP-glucose, is fueling glycolysis at the expense of starch synthesis and storage, and therefore it is one of the most important protein for anoxic ATP production (Taiz and Zeiger, 2006). *StHRE1* displayed a modest increase at a tuber age of 7 days, during the transition from stolon to juvenile tuber (Figure 43) after which expression decreased until it ultimately reached a value of about half of the expression level in a stolon. *StHRE2a* and *StHRE2b* displayed a transient upregulation at 7 days and toward the 14 days-age the expression level fell back to levels similar to those at the stolon-stage. However, both *StHRE2a* and *StHRE2b* were subsequently markedly up-regulated when oxygen levels decreased to 10% of the environmental concentration, in 25-day-old tubers (Figure 43B) congruent to the sudden increase in gene expression of *StADH* and *StSUS4*.

2.6.3 Characterization of *StHREs* co-regulators in tuber development by Spearman Correlation and the LASSO

Previous studies in *Arabidopsis thaliana*, *Oryza sativa* and *Vitis vinifera* have led to the conclusion that some genes belonging to group ERF-VII could play a role in the hypoxic induction of low-oxygen responsive genes (Fukao et al., 2006; Hinz et al., 2010; Licausi et al., 2010). However, the physiological hypoxia established during the development of the tuber is unlikely to pose a threat for the plant itself and therefore it is possible that in this context the ERF-VII genes play additional or different functions.

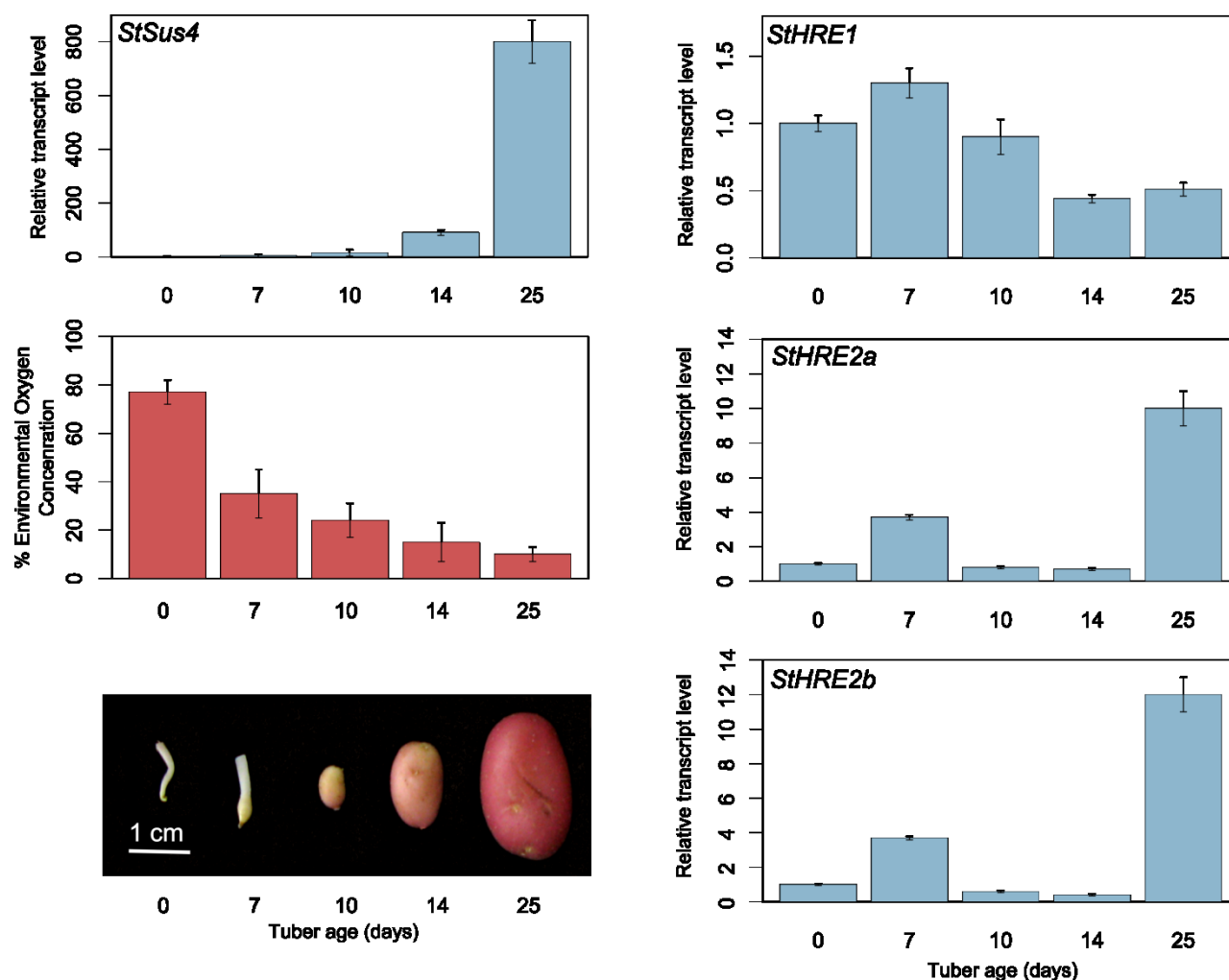


Figure 43 - Relationships between tuber age, oxygen concentrations and transcript levels of *StHRE1*, *StHRE2a*, *StHRE2b* and *StSUS4* in developing potato tubers. Internal oxygen concentration is measured in the core of tubers at different developmental stages. Transcript levels are obtained in developing potato tubers at the same stages used for the internal oxygen measurements using RT-qPCR. The fold change was calculated according to the $\Delta\Delta C_t$ method (stolon = 1). Error bars represent the standard deviation calculated for 4 biological replicates. A picture depicting the tuber growth stage and an indicative age is shown in the bottom. Note: the oxygen concentration measurements were conducted by Dr. Francesco Licausi (Scuola Superiore Sant'Anna, Pisa)

In order to detect genes that can be associated to the behavior of *StHREs*, and therefore shed light on the functions and mechanisms *StHREs* may be involved with, we carried out an expression-based gene network reverse engineering analysis. Without assumptions on causal relationships, co-regulation in this case can grasp the existence of mechanisms of regulation present in the studied samples that are altering the rates of transcription. It can therefore offer an unbiased approach for finding new genes sharing pathways and functions with *StHREs*, possibly other than the typical anaerobic adaptive response. In order to perform the co-regulation analysis we chose two distinct datasets that describe the transcriptomic changes occurring during potato tuber development. Both are based on the Potato Oligo Chip Initiative (POCI) microarray, which was based on the most recent representation of the *Solanum tuberosum* transcriptome at the time of analysis

(Kloosterman et al., 2008; Ferreira et al., 2010) (see Paragraph 4.6.3). At this point, we compared the LASSO (Tibshirani, 1996) with a distinct co-regulation based method for gene network reverse engineering, Spearman correlation. Spearman correlation was used instead of Pearson correlation, in this project, in order to further minimize the overlap of candidates selected by this technique and the LASSO (as by definition the first LASSO candidate is also the first Pearson candidate, see Paragraph 4.6.3). Furthermore, given the paucity of samples and the high correlation coefficients showed by the *StHREs* but also in general by the POCI dataset (Figure 44), Spearman correlation was chosen given to its robustness to outliers and noise artifacts (Usadel et al., 2009). Also the LASSO can be considered as a theoretically ideal technique in this case: since it prevents overfitting by keeping the number of explanatory variables low, it is potentially very robust even in scenarios where the number of samples is much lower than the number of genes, like in our case (14 samples, 31293 gene probes in the analyzed dataset).

Correlation Coefficient distribution for POCI array

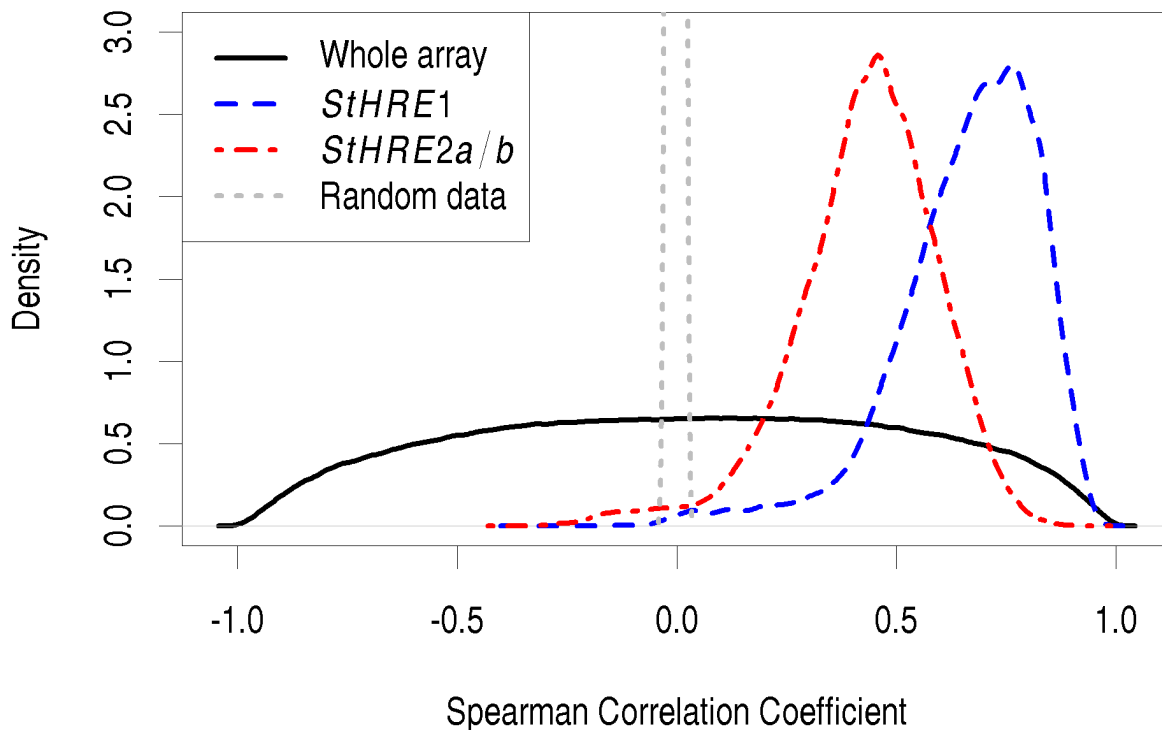


Figure 44 - Spearman correlation coefficients distributions for *StHRE1* (blue line), *StHre2a/b* (red line), all the probesets in the POCI array (black line) and the expected distribution from random Gaussian data (grey line).

As gene "baits" for our coexpression analysis using Spearman Correlation and LASSO we used the probes matching *StHRE1* and *StHRE2a/ StHRE2b*. Both genes are transcriptionally regulated during tuber

development, with a transcriptional variation higher than the population average (Figure 45). All other probes in the normalized datasets were taken as potential guilt-by-association candidates (see Paragraph 4.6.3).

Twenty-three genes were obtained by analysis of the LASSO model using *StHRE1* as response variable, and sixteen were found for *StHRE2*. Using Spearman correlation we found a considerable number of positively correlated genes, specifically 14758 for *StHRE1* and 927 for *StHRE2a/b* with an absolute correlation coefficient higher than 0.7, a commonly accepted correlation threshold (Usadel et al., 2009). These groups of coexpressors are in both cases significantly enriched for genes belonging to major and minor CHO metabolism, among other ontology groups (see Appendix, page 108) (Usadel et al., 2009). Interestingly enough, the vast majority of correlations for both *StHREs* are positive (Figure 44). Since the number of correlating genes was too high for an in-depth analysis, we decided to limit our candidates here to the top-10 correlators for both *StHREs*. This brought us to a total of thirty-three candidate coexpressors for *StHRE1* and twenty-five (as one gene is found in both the LASSO and the top-10 Spearman list) for *StHRE2/b*. Almost all these candidates are positively correlated to *StHRE1* and *StHRE2a/b* (Table 10 and Table 11) or with positive weights in the LASSO models (Figure 46).

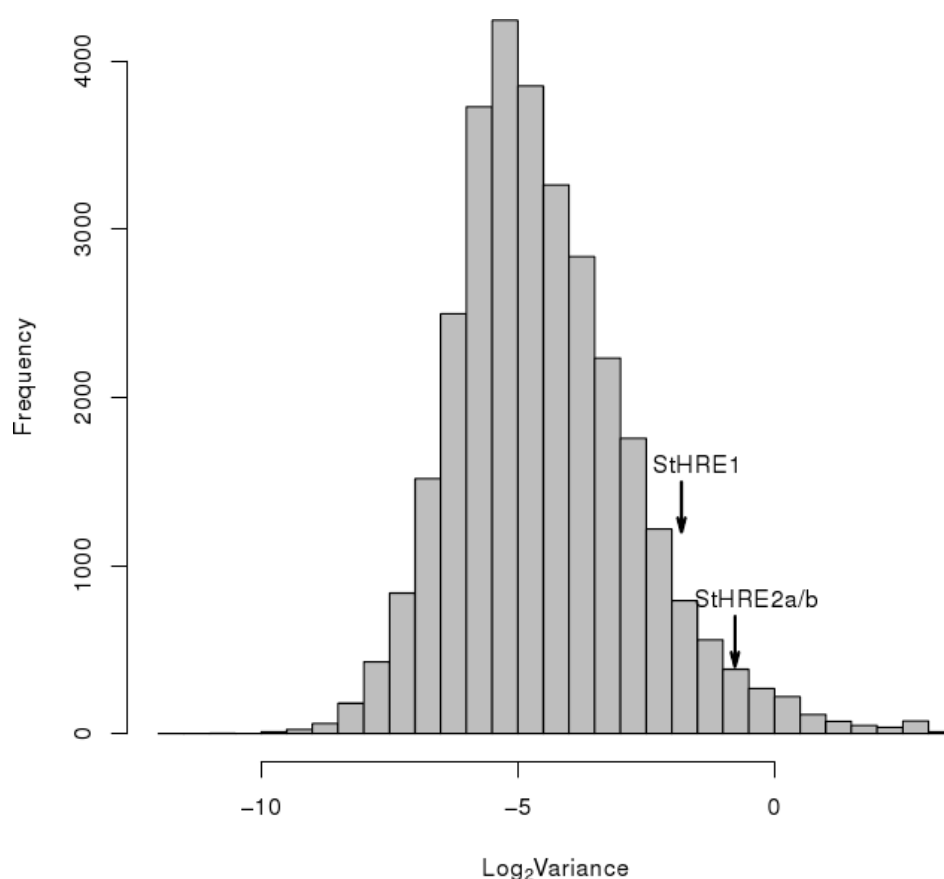


Figure 45 - \log_2 Variance distribution of probe intensities in the joint potato dataset (mean: 0.128; median: 0.038). Variances for *StHRE1* (0.285) and *StHRE2a/b* (0.587) are indicated.

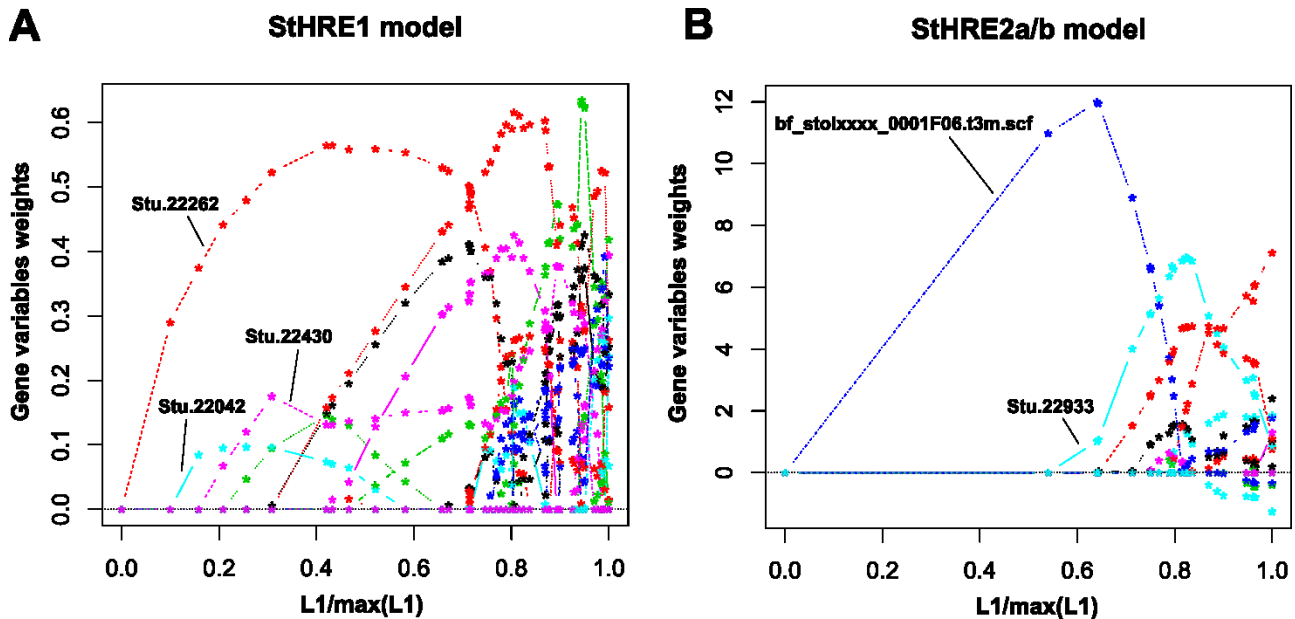


Figure 46 - Model weights for genes included in the *StHRE1* (panel A) and *StHRE2* (panel B) LASSO calculation. On the X-axis, the L1 threshold imposed to the models as a fraction of the maximum threshold. On the Y-axis, the variable weight. Each line corresponds to the model coefficients at different LASSO steps for a particular variable (gene).

The LASSO, although finding genes that are in most cases well correlated to the *StHREs* baits (Table 10 and Table 11), is providing candidates almost completely distinct from the top Spearman ranking ones. This behavior is not unexpected: as it could be seen in Table 9 in the previous Paragraph, LASSO and Correlation candidate selection tends to separate quite markedly. The first variable introduced by LASSO is always the top Pearson correlator (not the Spearman one). After the first variable has been included, the LASSO exploration proceeds by finding the genes most correlated with the “residual” behavior of the bait gene, yielding a quick divergence between LASSO and Correlation.

The group of genes whose expression relates with *StHRE1* (both top Spearman correlators and variables included by the LASSO models, see Paragraph 4.6.3) includes genes coding for anabolic enzymes involved in starch and sucrose biosynthesis such as a sucrose phosphate synthase (SPS1) and a phosphoglucumutase (PGM), as well as two enzymes involved in sugar catabolism [disproportionating enzyme (DPE) and a subunit of the cytosolic glyceraldehyde 3-phosphate dehydrogenase enzyme]. Four putative transcription factors-encoding genes also correlated with *StHRE1* during tuber development and three of them belonged to the Zinc-Finger family. The group of genes correlating with *StHRE2a* and *StHRE2b* instead did not include any transcriptional regulator but contained several genes involved in sugar metabolism and sugar-starch interconversion: phosphoglucan water dikinase (PWD), fructokinase (FRK), two glucose 6-phosphate transporter-like proteins (G6P-transporter) and *StSUS4*. Among the genes known to be involved in the anaerobic response (Mustroph et al., 2010) only *StSUS4* was identified as *StHRE2a/b* among the 58 candidates inferred by our coexpression analysis.

| Unigene ID | Coexpression method | Putative Function | Spearman Correlation Coefficient |
|------------|---------------------|--|----------------------------------|
| Stu.22262 | LASSO | Aspartyl protease family protein | 0.9297 |
| Stu.22430 | LASSO | Xanthoxin dehydrogenase | 0.8330 |
| Stu.7176 | LASSO | Amino acid transporter-like | 0.9121 |
| Stu.17280 | LASSO | Phytochrome b | 0.8725 |
| Stu.18175 | LASSO | Copine-related | 0.8769 |
| Stu.7147 | LASSO | Embryo-specific protein like | 0.8075 |
| Stu.449 | LASSO | Heat shock protein-like protein | 0.7978 |
| Stu.18907 | LASSO | Sucrose phosphate synthase (SPS1) | 0.9341 |
| Stu.16271 | LASSO | Basic helix-loop-helix (bHLH) | 0.7978 |
| Stu.6717 | LASSO | Zinc finger (B-box type) | 0.8418 |
| Stu.8764 | LASSO | Metal ion binding protein | 0.7714 |
| Stu.4930 | LASSO | Zinc finger (CCCH-type) | 0.7495 |
| Stu.5337 | LASSO | Zinc finger (B-box type) | 0.6879 |
| Stu.2227 | LASSO | Unknown protein | 0.7714 |
| Stu.18198 | LASSO | Phosphoglucomutase (PGM) | 0.7538 |
| Stu.5413 | LASSO | Homogentisate 1,2-dioxygenase | 0.7143 |
| Stu.3076 | LASSO | Pectinesterase | 0.8769 |
| Stu.22042 | LASSO | Unknown protein | 0.8374 |
| Stu.7459 | LASSO | Unknown protein | 0.7934 |
| Stu.399 | LASSO | Unknown protein | 0.8769 |
| Stu.9491 | LASSO | Unknown protein | 0.8505 |
| Stu.10230 | LASSO | Unknown protein | 0.7582 |
| Stu.4200 | LASSO | Unknown protein | 0.8330 |
| Stu.9263 | Spearman | Peptidyl-prolyl cis-trans isomerase | 0.9692 |
| Stu.4665 | Spearman | Mannose-1-phosphate guanylyltransferase | 0.9648 |
| Stu.18546 | Spearman | Disproportionating enzyme (DPE) | 0.9604 |
| Stu.290 | Spearman | Ubiquitin-conjugating enzyme, putative | 0.9604 |
| Stu.4096 | Spearman | Protease-related | 0.9604 |
| Stu.17424 | Spearman | Short-chain dehydrogenase | 0.9560 |
| Stu.22641 | Spearman | Glyceraldehyde-3-phosphate dehydrogenase C subunit (GAPDH-C) | 0.9560 |
| Stu.6095 | Spearman | Regulator of chromosome condensation related | 0.9560 |
| Stu.15767 | Spearman | RAS-related protein | 0.9560 |
| Stu.3773 | Spearman | Unknown protein | 0.9516 |

Table 10 - *Solanum tuberosum* sequences co-regulated with *StHRE1* during tuber development

| Unigene ID | Coexpression method | Putative Function | Spearman Correlation Coefficient |
|------------|---------------------|--|----------------------------------|
| Stu.21019 | LASSO | 60S ribosomal protein | 0.6000 |
| Stu.20114 | LASSO | Sucrose synthase 4 (SUS4) | 0.8593 |
| Stu.21913 | LASSO | DEAD box RNA helicase, putative | 0.5780 |
| Stu.10146 | LASSO | Aldehyde oxidase | 0.6615 |
| Stu.3439 | LASSO | Unknown protein | 0.6527 |
| Stu.4779 | LASSO | Prohibitin-like | 0.6747 |
| Stu.22464 | LASSO | Mitochondrial 18S ribosomal RNA | 0.0813 |
| Stu.8788 | LASSO | Light-harvesting chlorophyll b-binding protein | -0.2176 |
| Stu.20202 | LASSO | Rubber elongation factor (REF) | 0.3143 |
| Stu.9387 | LASSO | Phosphoglucan, water dikinase (GWD) | 0.8418 |
| Stu.22933 | LASSO/Spearman | Unknown protein | 0.8769 |
| Stu.22888 | LASSO | Unknown protein | 0.5165 |
| Stu.13477 | LASSO | Unknown protein | 0.7495 |
| Stu.23374 | LASSO | Unknown protein | 0.4462 |
| Stu.22993 | LASSO | Unknown protein | 0.7407 |
| Stu.20380 | LASSO | Unknown protein | 0.7011 |
| Stu.2176 | Spearman | Fructokinase (FRK) | 0.9385 |
| Stu.22678 | Spearman | Glucose-6-phosphate transmembrane transporter | 0.9253 |
| Stu.8847 | Spearman | Apolipoprotein D-related | 0.9209 |
| Stu.14473 | Spearman | Electron carrier | 0.9077 |
| Stu.15975 | Spearman | Remorin family protein | 0.8901 |
| Stu.4348 | Spearman | Glucose-6-phosphate transmembrane transporter | 0.8857 |
| Stu.6993 | Spearman | Unknown protein | 0.8813 |
| Stu.3944 | Spearman | GILT family protein | 0.8725 |
| Stu.2262 | Spearman | Unknown protein | 0.8637 |

Table 11 - *Solanum tuberosum* sequences co-regulated with *StHRE2a/b* during tuber development

As expression of the *StHREs* was modified by hypoxia (Figure 43), it is possible that the newly identified genes co-expressed with *HREs* during tuber development are also affected by decreased oxygen conditions. We checked whether the *Arabidopsis* and rice best BLAST hits of these genes were indeed induced or repressed by using publically available microarray datasets. Interestingly, among the homologues of the *StHRE1*-correlating genes in *Arabidopsis* only few genes exhibited hypoxia-responsiveness whereas the expression of the rice homologues was affected by anoxia in the coleoptiles (Lasanthi-Kudahettige et al., 2007). More specifically, 17% of the rice *StHRE1*-coregulated genes homologues were down-regulated (≤ 1.5 logFC) and 25% upregulated (≥ 1.5 logFC). Rice homologues of 50% of the *StHRE2* co-regulated genes were induced (≥ 1.5 logFC) and only 10% repressed (≤ 1.5 logFC). This is not significantly differing (tested via χ^2 test) from the expected range of this dataset, where 24.91% of the total genes are up-regulated and 36.30% are down-regulated by hypoxia.

The genes in the same co-regulation network with either *StHRE1* or *StHRE2a* and *StHRE2b* during tuber development (Table 10 and Table 11) can therefore be assumed to be alternatively regulated by (1) tuber development, (2) oxygen availability, or (3) both. Since the potato tuber is already an hypoxic tissue (Figure 43), it is difficult to separate hypoxic effects from developmental effects. However, it is possible to treat the tubers in hyperoxic (40% O₂) conditions, in order to distinguish genes that are effectively affected by oxygen

concentrations. The expression of the same set of genes was therefore analyzed in fully developed tubers of soil-grown potato plants treated under hyperoxic (40% O₂) and normoxic (21% O₂) conditions. Nine out of twenty-one genes exhibited significantly reduced mRNA levels when the pots in which they had grown were submitted to an oxygen-enriched atmosphere for 12h (Table 12). Six of these significantly repressed genes were found by LASSO modeling, while three were obtained through Spearman correlation, showing an interesting complementary behavior of these two techniques. This group of genes included the *StHRE1*-coexpressed genes *SPS1*, *Xhantonine dehydrogenase*, two Zinc Finger transcription factor genes, a *RAS-related protein* and *DPE*. Among the *StHRE2*-coregulated genes *StSUS4* was down-regulated by hyperoxia, together with *FRK* and *REF* (Table 12). Our analysis clearly showed a correlation between ERF-VII and other genes that goes beyond the co-regulation due to low oxygen stress (Youm et al., 2008; Hinz et al., 2010; Licausi et al., 2010). Instead, in the case of growing potato tuber, many *StHRE*-coexpressed genes code for enzymatic reaction involved in sugar catabolism and starch synthesis (Appeldoorn et al., 1997).

| Gene | Bait | Method | Putative function | Relative mRNA level | S.D. | P-value |
|------------------|--------|----------|--|---------------------|-------|----------------|
| Stu.16271 | StHRE1 | LASSO | Basic helix-loop-helix (bHLH) | 2.100 | 1.742 | 8.8E-02 |
| Stu.18198 | StHRE1 | LASSO | Phosphoglucosmutase (PGM) | 0.961 | 0.943 | 4.7E-01 |
| Stu.18907 | StHRE1 | LASSO | Sucrose phosphate synthase (SPS1) | 0.366 | 0.190 | 1.6E-02 |
| Stu.22430 | StHRE1 | LASSO | Xantonin dehydrogenase | 0.277 | 0.172 | 9.0E-04 |
| Stu.3076 | StHRE1 | LASSO | Pectinesterase | 0.590 | 0.656 | 1.4E-01 |
| Stu.4930 | StHRE1 | LASSO | Zinc finger (CCCH-type) | 0.901 | 0.818 | 4.2E-01 |
| Stu.5337 | StHRE1 | LASSO | Zinc finger (B-box type) | 0.314 | 0.306 | 4.9E-02 |
| Stu.6717 | StHRE1 | LASSO | Zinc finger (B-box type) | 0.427 | 0.062 | 1.1E-05 |
| Stu.7147 | StHRE1 | LASSO | Embryo-specific protein like | 3.112 | 3.755 | 8.5E-02 |
| Stu.7176 | StHRE1 | LASSO | Amino acid transporter family protein | 2.815 | 2.451 | 9.0E-02 |
| Stu.22641 | StHRE1 | Spearman | Glyceraldehyde-3-phosphate dehydrogenase C subunit (GAPDH-C) | 0.554 | 0.107 | 2.0E-01 |
| Stu.4665 | StHRE1 | Spearman | Mannose-1-phosphate guanylyltransferase | 5.391 | 8.710 | 1.6E-01 |
| Stu.15767 | StHRE1 | Spearman | RAS related protein | 0.387 | 0.321 | 1.5E-02 |
| Stu.18546 | StHRE1 | Spearman | Disproportionating enzyme (DPE) | 0.073 | 0.015 | 1.3E-03 |
| Stu.20114 | StHRE2 | LASSO | Sucrose synthase 4 (SUS4) | 0.208 | 0.120 | 8.1E-03 |
| Stu.20202 | StHRE2 | LASSO | Rubber elongation factor (REF) | 0.272 | 0.181 | 2.6E-05 |
| Stu.4779 | StHRE2 | LASSO | Prohibitin-like | 0.843 | NA | NA |
| Stu.9387 | StHRE2 | LASSO | Phosphoglucan, water dikinase (GWD) | 0.532 | 0.540 | 2.0E-01 |
| Stu.2176 | StHRE2 | Spearman | Fructokinase (FRK) | 0.211 | 0.216 | 3.0E-03 |
| Stu.22678 | StHRE2 | Spearman | Glucose-6-phosphate transmembrane transporter | NA | NA | NA |
| Stu.4348 | StHRE2 | Spearman | Glucose-6-phosphate transmembrane transporter | 0.907 | 0.413 | 3.7E-01 |

Table 12 - Effect of an hyperoxic (40% O₂) atmosphere on the expression of HRE1 and HRE2a/b co-expressed genes in fully developed tubers. In bold, significant deviations from control conditions (21% O₂)

2.6.4 Conclusions on *StHRE* characterization and co-regulation analysis

In summary, in the present study we have identified three ERF-coding genes, named *StHRE1*, *StHRE2a* and *StHRE2b* belonging to the group VII and displaying a differential responsiveness to low oxygen in potato. They appear not only to have a role in the response to the hypoxic stress decreased oxygen availability in the surrounding environment, as observed in *Arabidopsis* and rice, but also when developmental and growth programs pose constraints to oxygen diffusion. A simple expression based gene network reverse engineering analysis suggested a possible role of these TFs in the regulation of sucrose and starch metabolism during tuber development. Specifically, LASSO coupled with Spearman Correlation indeed helped us finding genes whose expression behavior is actually affected by what we expect to be a major regulator to *StHREs*, oxygen availability. As shown in Paragraphs 2.4.3 and 2.5, LASSO and Correlation are complementary in this task, since they can find non-overlapping groups of positive genetic hits (Table 13). Our screening shows us that some of our candidates are not only simply varying during development, but are possibly regulated by the oxygen availability in the tuber. Further genetic screenings will help us elucidate if *StHREs* are indeed the transcription factors regulating our candidates. The identification of potential regulators of adaptive mechanisms to low oxygen conditions in potato will be of great agronomical interests as oxygen availability affects tuber yield and quality (Holder and Cary, 1984).

| | Gene bait | | Total |
|-----------------------------|---------------|------------------|-------------|
| | <i>StHRE1</i> | <i>StHRE2a/b</i> | |
| LASSO | 4/10 | 2/4 | 6/14 |
| Spearman Correlation | 2/4 | 1/3 | 3/7 |

Table 13 - Fraction of co-regulators repressed by hyperoxia in *Solanum tuberosum* tubers, over total number of genes screened by RT-PCR.

3. Discussion

3.1 Transcript model characterization for microarray generation

It has recently been shown how DNA microarrays and next generation sequencing transcript quantification (RNA-Seq) yield largely equivalent results (Malone and Oliver, 2011). However, microarrays still have the not negligible advantage of over a decade of perfecting, both on the technical and on the statistical side. Furthermore, their relative lower cost (about 10-fold less than next-generation sequencing) make them still an exceptional platform for quantitative and fast transcriptome measurements. Finally, the enormous amount of data collected in different species and experimental conditions by microarrays is able to support comparative tasks and co-regulation studies that rely on large datasets of co-measured transcripts, such as gene network reverse engineering.

In this dissertation, I showed how a probe population for microarray design can be obtained by drafting a transcriptome through a combination of publicly available ESTs and *ad hoc* next generation sequences in *Theillungiella salsuginea* (salt cress, Paragraph 2.1). The advantage for the scientific community of such an approach is three-fold, since it provides a pipeline for transcriptome characterization, a comprehensive microarray for understanding gene expression in *Theillungiella*, and allows for comparative considerations between this extremophile cress and its close non-extremophile relative *Arabidopsis thaliana*. The transcriptome assembly pipeline shows how several technical and conceptual issues must be considered in this task, for example the necessity to remove contaminant reads and to merge highly-similar transcript models arising from a heterozygous sample population. Furthermore, it was shown that normalizing a transcript library, i.e. reducing the highly abundant transcripts, is indeed providing a partially not overlapping set of transcripts to an unnormalized library. This phenomenon will need further investigation, however it can be hypothesized that the normalized library highlights lowly abundant transcript, but at the same time fragments the mRNAs making some transcripts harder to assemble. The measurements obtained by the *Theillungiella* Agilent 44k microarray platform generated through this study, which are being conducted at the time this thesis is written, will possibly give answers to these topics. This prototype microarray will then act as both a product and a validation instrument for our transcriptome characterization pipeline. If deemed valid, this pipeline could be automated and used as a model framework for fast generation of screening microarrays on any nonmodel plant species, which will be based on publicly available sequence data joined with targeted next-generation sequencing measurements encompassing several experimental conditions. On the other hand, the availability of the transcriptome itself is bound to shed light on the salt- and drought- resistance features of *Theillungiella salsuginea*. Our preliminary screening shows that, when comparing *Theillungiella* to its relative *Arabidopsis* (which is not as much salt and drought resistant), there is no evidently higher number of transcript types with an abiotic stress ontology. This can be due to several reasons, one being the incompleteness of our transcriptome, although several abiotic stress conditions (all those known to be involved with the stress-related LEA proteins activity for example (Hundertmark and Hinch, 2008)) were applied to the *Theillungiella* samples used for sequencing and therefore should have triggered the transcription almost the

complete picture of stress-related genes. Another reason is that *Thellungiella* can have evolved a stress-response mechanism that doesn't depend as much on transcript variety, therefore not a "one transcript for one stress" principle, but rather a general system for dealing with extreme environmental conditions, based also on cellular and tissue structure and chemical properties to deal with a constant "stress threat" (Wong et al., 2005), and a highly refined intertwining with other transcriptional pathways, such as growth and hormonal regulation (Wong et al., 2006). A third reason could be that *Thellungiella* developed stress-specific transcripts that are so dissimilar from *Arabidopsis* and other model species sequences that our automatic annotation pipeline (Lohse and Usadel, unpublished) couldn't annotate them in the abiotic stress MapMan bin; although this is improbable, given the close evolutionary proximity between *Arabidopsis* and *Thellungiella* (Wang et al., 2004), it may be a justification for the high number of transcripts categorized as "unknown" in the *Thellungiella* transcriptome.

3.2 Caveats in microarray data normalization

After discussing the necessity of a detailed transcript model generation for microarray design, it was shown how commonly used normalization procedures for microarrays, despite the "age" of these platforms and a well-established statistical methodology (Bolstad, 2008), can still contain spurious effects and generate undesired artifacts. I showed how, in Affymetrix chip normalization procedures implementing a median polish probeset summarization (e.g. the commonly used RMA and GCRMA methods), the correlation between samples can be largely overestimated in small and odd-sized datasets (Paragraph 2.2). This effect could be significantly associated to certain properties of the *Arabidopsis thaliana* Affymetrix ATH1 microarray, although similar conclusions can be drawn also with other Affymetrix platforms, such as the *Escherichia coli* ASv2. Specifically, probesets with low expression and a high number of targets, especially if these are highly diversified in their function, yield a more severe over-estimation of the inter-array correlation. The problem is not to be underestimated, since the number of multiple-hit probesets is relevant, at least in the ATH1 microarray (Figure 47), which shows a 13.4% of "promiscuous" probesets.

This particular artifact correlation between samples, depending on "noisy" probesets, doesn't pose a particular problem in differential gene expression analyses, because, on the contrary, it could enhance the differences of changing transcripts by shrinking most unclear probesets to identical values across experiments. In any case, the small underlying change in gene expression of such an unclear probeset would generally be below the cut-off value to be considered an 'interesting' gene. However, this artificial correlation can't be ignored in contexts where unbiased measurements are needed, like transcript clustering (Golub et al., 1999), genetic network reverse-engineering (Basso et al., 2005), sample classification (Nielsen et al., 2007) (Eisen et al., 1998) or global transcript models (Usadel et al., 2008), and I show in the dissertation how RMA can artificially decrease the gap between samples coming from different tissue types. Furthermore, median polish has already been shown to work poorly when compared to MAS5 summarization in correlation between *E. coli* operon members (Harr and Schlotterer, 2006). Thus, these results should be taken as caveats on the validity of many studies obtained on the basis of correlation measures after these normalization procedures were applied, especially when small sample sizes are used. We therefore propose an easy fix to the median polish summarization

step, which remove almost completely the over-correlation between samples, while keeping all the positive features of RMA and GCRMA. It will be interesting to see how this modified RMA method, called tRMA ("transposed RMA", due to a step in the probeset matrix summarization), will behave in a wide range of applications. In particular, we assessed (data not shown) that its performance in preprocessing data for gene network reverse engineering is largely equivalent to that of RMA; RMA is known to perform relatively well with this respect, avoiding the background correction issues of GCRMA, which massively decreases the quality of inferred gene-gene associations (Lim et al., 2007). In fact the differences between RMA and tRMA, involving the transposition of the probeset matrix upon median polish summarization don't influence the correlation patterns between genes, but only that between samples.

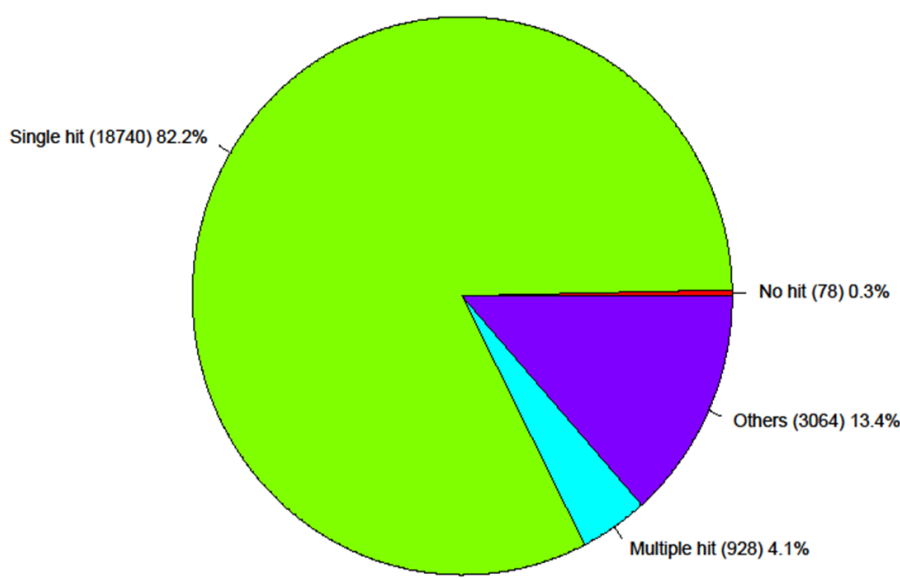


Figure 47 - Probeset population divided by number of hits. "Single hit" indicates a perfect probe agreement (usually, 11 out of 11) for a single perfect gene hit (BLAST bit score threshold 42.1, corresponding, for a 25-mer typical probe, to 21/25 perfect alignment or 25/25 with one mismatch; note: an almost identical pie chart is obtained imposing more strict bit scores). "Other" indicates probesets whose probe population matches several different targets. "Multiple hit" is a particular case of "family" probesets, indicating a perfect probe agreement over more than one target (i.e. for identical, recently duplicated genes).

3.3 Conditional correlation techniques in central gene prediction

The transcript expression data obtained through microarray studies, after establishing a proper understanding of this platform, can therefore be used to make biological inferences about the underlying gene-gene associations. One common way to do this is through the use of networks as comprehensive representations of the underlying transcriptional control mechanisms (McAdams and Shapiro, 1995). Using gene and protein networks instead of tabular or matrix representation of gene-gene interactions has been beneficial not only from a visualization point of view (although it can be argued that a better visualization can lead to a better understanding of the underlying system (Mutwil et al., 2010)). For example, global network properties, such as the scale-free distribution of the network degree, has shed light on the evolution of transcriptional control, which tends to generate novel control steps on pre-existing master transcriptional regulator. This yields a

network structure with many genes connected to a few central nodes which, due to their high degree centrality, are deemed "network hubs" (Barabási and Albert, 1999). Several gene-level network properties have also been studied and associated to particular characteristics of the genes involved. For example, it is known that cancer genes have a higher degree and clustering coefficient centrality than other genes (Rambaldi et al., 2008). In this thesis (Paragraph 2.3), we confirm what previously inferred by separate studies about network centrality and gene essentiality (Jeong et al., 2001), which principally show a high degree for all essential genes. We confirm the significantly higher network degree for the essential genes in *Arabidopsis thaliana* (as shown by (Mutwil et al., 2010) among others) and observe significantly higher clustering coefficient and betweenness too. In addition, we tried to assess how conditional correlation can be used to extract a different concept of "centrality" in expression-based gene networks.

We could initially observe (e.g. Figure 27 and Appendix, page 112, Figure 52) that Conditional correlation can remove high fractions of edges and yield scale-free networks with a low amount of gene "hubs" even when mild thresholds are applied (refer e.g. to Figure 27 for network size reduction and to Appendix, page 112, Figure 52, for scale-free topology generation upon application of Conditional correlation to expression-based networks). This edge reduction is far more prominent than in random correlation networks (data not shown, generated based on the Erdős–Rényi structure (Erdős and Rényi, 1959)), showing that the insurgence of gene hubs upon application of Conditional correlation must be a particular property of gene expression networks. It is interesting to note how Conditional correlation is also one of the techniques used to infer causal relationships in graphs (Pearl, 2000): in these approaches, conditioning is used not only to remove edges, but also to direction the surviving ones, yielding that the most connected surviving nodes are also "central" in cause-effect, information flow relationships.

In our study, we showed how applying Conditional correlation in expression-based networks is beneficial in predicting essential genes in *Arabidopsis thaliana*, since essential genes tend to be fill the role of gene hubs in this scenario, significantly more often than non-essential ones. The novel centrality index based on Conditional correlation, called "Breaking Potential" is able to distinguish efficiently between essential and non essential genes in expression-based networks, although it hasn't been able so far to exclude the possibility for this result to be just a by-product of network degree. We could assess similar properties for essential genes also for networks based on *Escherichia coli* and *Saccharomyces cerevisiae* microarray datasets (data not shown). Therefore, in a broader perspective, an essential gene is characterized by being co-expressed to a high number of genes (high Degree), which in turn tend to be co-expressed to each other (high Clustering coefficient). Furthermore, essential genes seem to act as connectors of different parts of the transcriptional network (high Betweenness) and are likely to be central also in information-flow processes (high Breaking Potential). However, Breaking Potential is not a mere essential gene predictor, since genes can have high Breaking Potential in a non-essential pathway, and therefore be locally central in the information flow of a particular cellular function. The capability of Breaking Potential to depict genes conditionally central in pathways makes it a promising feature to find key regulators not only in molecular biology, but also in other

networks, reverse-engineered from data, for which conditional similarity measures can be defined, e.g., in neurosciences (Fransson and Marrelec, 2008) or economics (Mizuno et al., 2006).

The Breaking Potential seems an interesting network property where causality and precision are important goals to achieve. Unfortunately finding that Breaking Potential results on essential genes predictions are strongly biased by the network degree imposes a reflection on the intrinsic message given this measurement. It will be worth investigating if Breaking Potential contains peculiar properties that separate it from the degree in coexpression networks, trying to use it without a predetermined goal, but assessing the ontology of the nodes it extracts from correlation networks. However, preliminary results, using dynamically adaptable thresholds based on dynamical network quality assessments, hint that the Breaking Potential index would indeed be able to intrinsically outperform degree in identifying "central" genes.

3.4 Gene network reverse engineering

More generally, expression-based gene network reverse engineering, both in a global and in a single-gene perspective, can take advantage of a combination of several different network inference methods. In the present thesis, we tested Correlation, Partial Correlation, Mutual Information and the LASSO, however other unrelated methods are beneficial in this respect, e.g. linear, polynomial or non-linear regression. Standard linear regression wasn't discussed, but it generally provides multi-variable models, with all genes associated to a different weight not necessarily associated to its relevance, and therefore with low *interpretability* and robustness (Tibshirani, 1996). Polynomial and in general non-linear regressions suffer from the same issue as standard regression, however they can detect peculiar co-regulation structures, similarly to what Mutual Information can do (Daub et al., 2004). The performance of these methods has not been investigated in detail, nor included in this dissertation for mainly three reasons. The first is that the amount of "peculiar" non-linear genetic interactions, which would require such sophisticated approaches, are extremely rare in gene-gene co-expression (the example shown in Figure 25 was found after testing several different datasets and bin numbers). The second is overfitting: a polynomial regression model with enough complexity can achieve a perfect prediction of transcriptional behavior, but this wouldn't be necessarily a significant model. The third problem is computational complexity, as trying to achieve polynomial regressions of the n^{th} order for a high number of explanatory variables (such as in microarray datasets) requires a severe increase of the number of variables to consider (not only the genes, but also the powers of the genes), making a broad application of these techniques (such as in Paragraph 2.4.3) still unfeasible given the current implementations.

The performance of the selected techniques, Correlation, Partial Correlation, Mutual Information and the LASSO, was therefore tested applying different combinations of thresholds. Several lessons given by e.g. conditional correlation applied to essential gene detection showed us that a technique has to be evaluated as broadly as possible, principally because the optimal threshold depends on the underlying dataset type, size, and correlation distribution. One way to overcome this issue is the use of mutual ranks, which deem as

significant interactions between genes that respectively hold the other in the top list of co-expressors, evaluated by e.g. correlation (Obayashi and Kinoshita, 2011).

The four classes of network reconstruction methods were compared at several stringency levels, and on different datasets (for practical reasons, we showed only the results for a high quality dataset in this dissertation, however similar conclusions can be taken for any *Arabidopsis* microarray-based condition-independent dataset). Such a combined approach shows that all methods share the capability to infer central transcriptional co-regulation clusters, such as the one composed by ribosomal proteins. It is generally known that genes encoding for subunits of the same complexes tend to be co-regulated (Tischler et al., 2008) (Liu et al., 2009) in order to preserve the stoichiometrical balance between proteins. This co-regulation is particularly strong for ribosomes, since in almost every *Arabidopsis* tissue and physiological condition the ~250 genes coding for ribosomal proteins are expressed (Barakat et al., 2001). We could observe strong co-regulation patterns (found by all co-regulation methods) also within other cellular complexes, like cell wall synthesis rosettes (Cosgrove, 2005) however with lower significance than the ribosomal clusters. However, as previously indicated in the introduction, co-regulation can arise from two mechanisms: in the case of complexes, genes are co-regulated because they are controlled by a common mechanism, and in general induced by a common cause. In the case of ribosomes, this is even more evident due to the high similarity of promoter sequence between groups of ribosome-encoding genes (Barakat et al., 2001). The second mechanism yielding observable co-regulation is a cause-effect relationship, like the one between an inducible transcriptional activator and its downstream activated gene. It has been observed before that conditional techniques (in our case, Partial correlation and the LASSO can be considered as such) excel in finding this second class of co-regulation phenomena, while direct techniques (Pearson correlation and Mutual Information in our study) are better performing in complex-forming co-regulation behaviors. In our investigation, it is not immediately possible to confirm these observations, since the vast majority of cause-effect relationships clearly defined and collected in the *Arabidopsis thaliana* golden set repository for genetic regulation (AtRegNet (Palaniswamy et al., 2006)) are not present in the microarray dataset investigated. However, among the significant functional connections found by the LASSO, it is possible to get an idea on some transcriptional control mechanisms (e.g. between bin 29.5.11, ubiquitination, and bin 27, collecting RNA-processing and regulating genes). In general, it can be said that all methods seem to obtain different areas of the transcriptional co-regulation events in the cell. This fact shows that the dynamics of co-regulation are not identical in all processes and strongly supports the use of a combined methodological approach to take advantage of this network reconstruction algorithm complementarity.

To the main plethora of these algorithms, in my work I refined and applied to large-scale expression dataset the LASSO-based network reconstruction method, in virtue of its several (theoretical) features, highly favorable in dealing with microarray datasets. These are: its robustness to noise and capability to remove indirect effects at once, the high interpretability of its results (not meaningful gene interactions are simply discarded) and the potential to work in scenarios with less samples (microarrays) than variables (genes). LASSO overall accuracy for network reconstruction seems lower when compared to e.g. Correlation, but as

stated before its complementarity to other methods make it applicable in conjunction to other approaches. The first hints for this were based on network quality assessments: LASSO networks significantly improve the quality of other expression-based networks when overlapped, but the complementarity of the LASSO and other Correlation-based methods can also be shown experimentally. Here, we show how genes involved in the seed coat mucilage pathway of *Arabidopsis thaliana* can be extracted using both Pearson Correlation and LASSO. Both these techniques target the network neighborhood of RHM2, a mucilage-deficient sugar mutant. And both yield a high fraction (around 40% of the suggested candidates) of confirmed "true" candidate genes, i.e. genes which, when knocked-out, show significant sugar alteration in their mucilage pectins. The investigation over these genes is currently proceeding, and it will be interesting to use them to test the theory of two separate pathways within the seed coat mucilage: one for pectin synthesis, one for pectin release, since for some genes and their co-regulators mechanical stress or EDTA addition can trigger the release of mucilage (appendix, page 122).

An interesting detail connecting the assembly of the *Thellungiella* transcriptome assembly to the seed coat mucilage pathway is the lack of visible mucilage release (even after mechanical stress) of the salt cress seeds (see appendix, page 123). At the same time, thanks to the availability of the assembled transcriptome, we could find putative orthologous sequences (assessed by BLAST best reciprocal hits) for all the *Arabidopsis* mucilage-related genes listed in Table 8 in *Thellungiella*, which therefore is most likely possessing the ability to synthesize the particular rhamnogalacturonic polysaccharides composing the mucilage. We could deduct that either *Thellungiella* is not synthesizing mucilage, or that perhaps the growing conditions of the seeds used in our preliminary analysis impaired the synthesis of mucilage (variation in mucilage production due to the conditions of seed storage is known to happen, as per personal information by Dr. Björn Usadel). In the first case this can be due to the particular osmotic growth conditions of *Thellungiella* seeds. The mucilage release is in fact impaired even when these seeds are hydrated with differently concentrated NaCl/Water solutions (I tested 0.1M, 0.5M - the molarity of seawater - and 1M solutions). Therefore, if the lack of mucilage will be confirmed, it can be concluded that the processes by which *Thellungiella* is able to germinate in salted water (Wang et al., 2004) do not require the protective effect of seed coat mucilage and must act by adopting different mechanisms.

After assessing the complementarity of gene network reverse engineering methods in the *Arabidopsis* seed coat mucilage pathways, we checked if this characteristic has a broad applicability and therefore co-tested LASSO and Correlation (in this case Spearman Correlation) in an independent scenario, the StHRE-related *Solanum tuberosum* tuber development. In this case, since the gene baits needed for network reconstruction were yet not identified in this organism, we detected them via the use of PCR and primers specific for ERF-VII genes. The baits identified were three, named *StHRE1*, *StHRE2a* and *StHRE2b* and displayed an interesting expression pattern during tuber development. Apart from being the first attempt to identify an ERF pathway in potato, this study once again showed the complementary potential of the two network reconstruction approaches, which yield distinct lists of genes verified to be regulated by oxygen availability, as expected from members of this pathway. The properties of the gene neighbors identified in this study furthermore allowed to

suggest a role for the *S. tuberosum* *StHREs* in the regulation of sucrose and starch metabolism during tuber development. While sugar signaling has been implicated in triggering changes in the routes of sucrose unloading and mobilization as well as in the rates of starch synthesis (Smeekens, 2000), the results of our analysis suggest that the decrease in oxygen tension that develops as tubers grow bigger and the related induction of ERF-type genes might also contribute in regulating these events.

In general, combining the bioinformatical and experimental results, it can be said that the LASSO can yield a unique set of gene connections, and therefore of true gene candidates, at a good rate of false positives (around 60% in the RHM2 model, see Figure 34 in Paragraph 2.5.1), similar to what we achieved by simple Correlation. Furthermore, it should be remarked that, however LASSO has been applied before in biological contexts (Shimamura et al., 2007; Gustafsson et al., 2009; Lu et al., 2011), the majority of the studies were focusing on artificial, ideal and theoretical scenarios, with only hints at the direct application in experimental tasks like gene candidate finding. For example in (Gustafsson et al., 2009) a very interesting pipeline for network reverse engineering is applied, using several nonlinear functions for network inference and the LASSO as a central cog in the algorithm for selecting a subset of significant genes (those possessing non-zero weights). However promising, LASSO results so far have been rarely tested in comparison to other alternatives, making the study described here so far unique in its kind.

3.5 Conclusions and future perspectives

In my work, the main focus has been to optimize expression-based inferences of transcriptional control phenomena in plants, with the final goal of increasing the understanding of biological pathways through LASSO and Correlation gene network reconstruction. However, during the proceeding of these investigations, a plethora of computational methods was generated for microarray data analysis and sample filtering, network reverse engineering and for bridging network and ontology enrichment analysis. This, we thought, can be beneficial *per se*, independently from the applications showed here, for any expression network-related study, and therefore we decided to bundle part of the algorithms described and implemented in this thesis into a tool, dubbed CorTo ("Correlation Tool"). However, CorTo is more than a collection of methods, and can be better defined as a framework for comparative network reconstruction reconstruction, visualization and analysis, and ontology enrichment assessment. Most of these tasks could only be executed until now with custom tailored R scripts (www.r-project.org) or via personal algorithm implementations; furthermore, the networks generated depend on external tools for visualization (e.g. Cytoscape (Shannon et al., 2003)). The completion of CorTo will hopefully bring the Bioinformatics approaches described here towards a more general usage among the scientific community. Further versions of this program will include network quality assessments and ontology-based summaries of gene networks (as in Figure 31D).

In conclusion, this thesis shows how gene expression can be used to infer transcriptional control mechanisms, through a pipeline that can be analyzed at several layers: data retrieval and preprocessing, co-expression and network analysis, biological validation. This approach will tremendously benefit from large-scale availability of publicly available RNA-Seq data, which are able to finally assess the status of the whole transcriptome.

However, I am conscious that the mere study of transcript amount oscillations won't provide a final understanding to the real transcriptional network active in the cell, although the fine tuning of data handling and network inference methods will allow the achievement of a cleaner read of the system. Such an aim will require the massive and harmonic collaboration between Transcriptomics the other -omics techniques, principally Metabolomics, given the high feedbacks between the two populations of molecular species (Hirai et al., 2004). An immediate goal of gene network reconstruction studies such as the one described in this thesis will be to associate expression profiles and promoter/enhancer characteristics (Seipel et al., 1992). There is still an open debate in the scientific community on whether there is co-evolution between expression patterns and sequence features, with positive (McCarroll et al., 2004) and negative examples (Jensen et al., 2006), however it is generally considered that transcriptional control tends to be conserved along with the promoter sequence of a gene (Peng and Weselake, 2011). The recent development in folding prediction of chromatin (Ho and Crabtree, 2010), allowing to know which enhancer regions are actually active over the gene of interest, combined with the improvement of co-expression studies and remote homology detection for promoter sequences (Tirosh et al., 2008), will hopefully allow to add an additional layer to our understanding of transcriptional control.

4. Materials and Methods

4.1 Transcriptome assembly

Raw 454 reads, generated by Yang-Ping Lee (Max Planck Institute for Molecular Plant Physiology), were pre-checked for contamination by non-plant sequences using a BLAST search (Altschul et al., 1990) against the NCBI database of non-redundant sequences (nr; updated at 12-02-2010) with an E-value threshold of 10^{-10} . All reads showing a best match to a sequence originating from an organism outside the *Streptophyta* phylum were considered as "contaminants". In order to be as conservative as possible, we allowed these reads - a minor part of the collection - to be included in the assembly process, but contigs including them were marked as "contaminant contigs" at the end of the process.

The assembly was conducted using the Mira assembler program version 3.1.15 (Chevreux et al., 2004) in accurate mode and default parameters. For comparison to other sequence assembler methods, I used iAssembler v1.0 (bioinfo.bti.cornell.edu/tool/iAssembler/) and CLC Workbench v4.0 (www.clcbio.com), both with default parameters and a minimum contig size filter of 40. Several assemblies using the different 454 libraries were performed with MIRA. Specifically, 454 reads from the normalized library (400,631 reads), the unnormalized library (811,683 reads) and the combined library (1,212,314 reads) were assembled. In every assembly the 44,551 published Sanger EST sequences were included (Wang et al., 2004) (Taji et al., 2008) (Wong et al., 2006) (Zhang et al., 2008) in order to improve the final result. Reads not aligning to any other read, so called singletons, were not included in the final contig population.

Average contig coverage was calculated as the mean of the number of reads per base per contig. The N50 parameter was defined as the contig size above which 50% of the total sequence nucleotides are contained (Table 3). All contigs were checked for presence of open reading frames (ORFs) using the method available at <http://proteomics.yzu.edu/tools/index.html> with default parameters. This method assesses if a nucleotide sequence, in any of its 6 reading frames, contains a putative ORF, or partial ORF, falling in any of the 10 mRNA models described by (Min et al., 2005). In order to assess the completeness of the transcriptome assemblies and the degree of overlap between *Thellungiella* and *Arabidopsis*, I used BLASTX to align the contig sequences to the *Arabidopsis* Information Resource (TAIR9) peptide library (27,739 sequences) (Rhee et al., 2003). The percentage of *Arabidopsis* proteins matching *Thellungiella* contigs was calculated with a loose threshold to account for interspecies variation (E-value < 10^{-10}).

All following steps were conducted on the assembly based on the combined library only. Since the population of *Thellungiella* plants used for this experiment was not homozygous (Dr. Yang Ping Lee, personal communication), many single nucleotide polymorphisms (SNPs) induced the generation of several nearly identical contigs. I aligned each contig to all contigs in the population via BLAST, to identify clusters of contigs matching each other with a sequence coverage and identity higher than 99%. A multiple alignment was produced for each cluster using MUSCLE (Edgar, 2004). Consensus sequences for each cluster were extracted from the multiple alignments using the *consambig* tool from the EMBOSS suite (Rice et al., 2000).

Where disagreeing base pairs were found, the resulting cluster sequence was dubbed using the IUPAC code for nucleotide ambiguity. Of the 46,220 contigs present in the final combined library output, 4,020 contigs were condensed into 610 clusters, leading to a final population of 42,810 (42,200+610) representative distinct sequences for the *Thellungiella* transcriptome.

Functional classification of the 42,810 putative transcripts was performed using the Mercator pipeline (Lohse and Usadel, unpublished). Mercator aligns all sequences against five different databases: TAIR9 proteins (Rhee et al., 2003), SwissProt/Uniprot plant proteins (PPAP) (Schneider et al., 2005), Conserved Domain Database (CDD) (Marchler-Bauer et al., 2005), Clusters of Orthologous Groups (KOG) (Tatusov et al., 2003), and InterProScan (Zdobnov and Apweiler, 2001) and subsequently computes preliminary MapMan BIN codes based on manually curated reference classifications using a majority vote scheme. The programs used to perform the searches were RPSBLAST (Schäffer et al., 2001) for CDD and KOG and BLASTX (Altschul et al., 1990) for TAIR9 and PPAP. Sequence alignments with bit scores lower than 50 were ignored as not significantly similar. A domain matching E-value threshold of 10^{-5} was applied to the InterProScan analysis.

The sequencing procedure does not guarantee that the orientation of the original mRNAs is kept, hence any of the 42,810 transcripts can be either 5'-3' oriented or 3'-5' oriented. In order to unify the orientation, I used the protein models present in the NCBI nr database and, where available, I used the best BLAST hit (E-value < 10^{-10}) to define the correct orientation of the putative *Thellungiella* original transcript.

The sequences, annotated and correctly oriented, were finally used for the generation of a 44k Agilent chip (Wolber et al., 2006). Each probe was designed by the manufacturer using the optimal 60mer subregion of each transcript.

4.2 Comparison of Microarray preprocessing methods

4.2.1 Microarray preprocessing methods

The microarray preprocessing procedures RMA (Irizarry et al., 2003), GCRMA (Wu and Irizarry, 2005) and MAS5 (Hubbell et al., 2002) were compared using the software implementations available from BioConductor (Gentleman et al., 2004). In every case, the default parameters were used. All final outputs, including MAS5 ones, were analyzed on the \log_2 scale.

4.2.2 Microarray datasets

In order to obtain a vast, robust and condition-independent dataset for assessing the performance of different microarray normalization algorithms, all *Arabidopsis thaliana* ATH1 microarrays available from GEO (Edgar et al., 2002) were downloaded, and subsequently filtered for truncated or unreadable files and genomic DNA experiments via human inspection (Venter et al., 2001). This dataset comprised 3707 arrays and is henceforth referred to as the "*Arabidopsis dataset*".

To test the abilities of RMA and tRMA to correctly cluster different tissue samples, I analyzed microarrays from the AtGenExpress stress study (Kilian et al., 2007), contained in the Gene Expression Omnibus series GSE5620-GSE5628. This dataset (*root-shoot dataset*) comprises 248 samples, evenly distributed in shoot and root tissues.

To further assess sample classification performance of RMA and tRMA, I focused on a human breast cancer dataset published by (Signoretti et al., 2002) and reanalyzed by (Eklund and Szallasi, 2008). This dataset contains 98 surgical specimens, 18 of which belong to 9 replicate pairs in which two samples were taken from adjacent sections of the same frozen block.

4.2.3 Permutation of microarrays

In order to compare real samples with completely uninformative samples, I decided to randomly permute the raw signal intensities of the *Arabidopsis* dataset using the same procedure as in (Lim et al., 2007). In brief, every Perfect Match (PM) probe and its Mismatch (MM) counterpart were reassigned to a random probeset within the same microarray. This generates information-less probesets while keeping the properties of the original probe intensity distribution. The code has been courteously provided by Wei-Keat Lim and Andrea Califano (Columbia University) and modified for the ATH1 Affymetrix chip.

4.2.4 Inter-array correlation analysis

The behavior of the three microarray preprocessing procedures was analyzed in the context of randomly selected subsets of the *Arabidopsis* dataset. Different sample sizes were selected (2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 30, 50 and 100) according to the typical dimension of a single-experiment dataset. For each sample size, 1000 subsets were randomly selected and normalized. For each normalized subset, I calculated inter-array Spearman correlations and then plotted the overall mean and standard deviation of these correlations for each sample size.

The same procedure was then repeated for the permuted *Arabidopsis* dataset.

4.2.5 Noise robustness analysis

In order to assess the response of RMA, GCRMA and MAS5 to data perturbation, increasingly noisy samples were generated using the formula:

$$I = w_o \cdot O + w_p \cdot P$$

where I is the final probe intensity, O and P are, respectively, the original intensity and a permuted intensity. w_o and w_p are the weights given to both (where $w_o + w_p = 1$). w_p is referred to as noise level.

4.2.6 Linear model for measuring internal probeset consistency

Internal probeset consistency was analyzed by applying a linear model. Given a matrix for each probeset, where columns are samples and rows are probes, I ranked the values row-wise and determined the model

$$p_{ij} = w \cdot S_j + \text{int} + \epsilon$$

The model tries to predict every i^{th} probe intensity rank in the j^{th} sample (p_{ij}) using as explanatory variable the j^{th} sample effect, S_j , calculated as the probe's rank within the sample j . The model will then try to adjust the sample effect weight w and the intercept int to minimize the unexplained error ϵ . It is apparent that the R^2 for this model will be high when all probes within a probeset behave consistently relative to each other across

different experiments, i.e. when the probe rank in a specific experiment is predicted quite well by the probe's mean rank across experiments. On the other hand, a low R^2 will result from probes acting inconsistently across experiments, e.g. with some probes ranking particularly high in some experiments yet low in others. The modeling procedure is provided in the as an R function (fitLM, publicly available (Giorgi et al., 2010)) to determine internal consistency of probesets.

4.2.7 Transposed RMA (tRMA)

With the goal of reducing inter-array correlation artifacts without losing the positive features of RMA, I modified the RMA median polish source code of the preprocessCore library available on BioConductor (Gentleman et al.). This new method simply changes the order of median substitution, starting from column (sample-wise medians) instead of from rows (probe-wise medians), and was therefore called "transposed RMA" (or tRMA). tRMA code is publicly available (Giorgi et al., 2010) and can be run in the R environment (Dai et al.).

4.2.8 AffyComp benchmark

In order to evaluate and benchmark our newly proposed preprocessing method, tRMA, I adopted the criteria developed for the AffycompII challenge (Irizarry et al., 2003) (Irizarry et al., 2006) using the two Affymetrix spike-in datasets HGU95 and HGU133 and the AffyComp online tool (Irizarry et al.).

4.2.9 Sample classification performance

From the *root-shoot dataset*, I randomly selected 10000 groups of 5 arrays composed of 3 samples from one tissue type, and 2 from the other. Each dataset was normalized using tRMA or RMA and distances between arrays hybridized with the same tissue (intra-tissue distance) and between arrays hybridized with different tissues (inter-tissue distance) were determined. Distances were calculated as (1-Spearman correlation coefficient) using either all probe sets or only the 50% showing the highest variance.

Secondly, a dataset previously used by (Eklund and Szallasi, 2008) to assess microarray performance was used to determine the percentage of correctly clustered subsets of 5 microarrays. From the dataset, two couples of samples coming from the same tumor or non tumor specimen, plus a different specimen were sampled. Probe-sets were selected based on differential expression between the samples using the limma package applying different p-value thresholds corrected using the Benjamini-Hochberg method (Hochberg and Benjamini, 1990). The outcome of the normalization was defined as "correct" if, for every sample in a couple, its highest correlation coefficient against all other samples is the other correct member of the couple, which would lead to them being clustered together. The sampling was repeated 1000 times for each different p-value. The increase in the performance of tRMA when compared to RMA was assessed using a Fisher's exact test with Benjamini Hochberg correction.

The human dataset was used also to perform a test on clustering performance on groups of genes sorted by variance, as described by (Eklund and Szallasi, 2008), but using only subsets of five samples (belonging to three groups). This test was performed for RMA and tRMA at different probe noise levels, added following the procedure described previously in Paragraph 2.2.5.

4.3 Network Centrality and Breaking Potential calculations

In order to analyze a full physiological map of *Arabidopsis thaliana* gene expression, the complete *Arabidopsis* developmental microarray collection generated by the AtGenExpress consortium (Schmid et al., 2005) was obtained and ten normalized using tRMA (Giorgi et al., 2010). Only probesets with at least 95% present values were kept (following the Affymetrix PA calls, see Paragraph 4.5). To avoid overrepresentation of experimental conditions the chosen number of replicates per experimental condition was one; furthermore, to focus on physiological conditions, experiments with mutant plants were excluded. This dataset comprised 63 experiments and 12200 valid measured genes. The dataset is publicly available in CSB.DB (Steinhauser et al., 2004) under the name "atge0100". Network centralities were calculated using the JUNG library (jung.sourceforge.net) and a JAVA implementation of the Breaking Potential calculation. GeneNet networks were obtained using the GeneNet R package (Opgen-Rhein et al., 2007) with default parameters. All Breaking Potential plots (Figure 22 and Figure 23 and Figure 49) were generated using R (www.r-project.org). Joint ROC curves between Breaking Potential and other centralities were calculated by taking the average rank for each gene in both centralities. The list of *Arabidopsis thaliana* essential genes was obtained from SeedGenes 2007 (Tzafrir et al., 2003).

4.4 CorTo tool development

CorTo is a multithreaded Java Swing graphical application and is compatible for Windows, MacOS-X and Linux/Unix platforms supporting JAVA SE 6. The software GUI was designed with the WindowBuilderPro infrastructure (code.google.com/javadevtools/wbpro) within the Eclipse Integrated Development Environment (www.eclipse.org). The co-occurrence algorithms are a Java implementation of the respective algorithms, using the linear algebra calculus functionality provided by the ParallelCOLT (sourceforge.net/projects/parallelcolt) and Apache Commons Math (commons.apache.org/math) projects. The JUNG library (jung.sourceforge.net) was used for network visualization and the JFreeChart library (www.jfree.org/jfreechart) for plot visualization.

All preloaded datasets in CorTo datasets were publicly available on Gene Expression Omnibus (Edgar et al., 2002), microarrays were preprocessed using tRMA (Giorgi et al., 2010) using the Robin tool (Lohse et al., 2010), replicates were averaged and, where available, an updated probeset gene annotation was used (Dai et al., 2005).

4.5 Gene Network Reconstruction and Comparison

The dataset used for comparing gene network reconstruction algorithms is the same available in (Mutwil et al., 2011). In brief, all ATH1 *Arabidopsis thaliana* Affymetrix microarray data was downloaded from Gene Expression Omnibus (Edgar et al., 2002) and ArrayExpress (Parkinson et al., 2007), totaling 8369 samples.

After merging identical entries and removing corrupted entries, this number decreased to 6255 samples. At this point, a Kolmogorov-Smirnov statistics based method was applied using the deleted residuals principle as described in (Persson et al., 2005), with a cutoff KS score of <0.15 , in order to remove samples whose general behavior heavily diverges from the population. This removed 505 samples, bringing the total to 5750 microarrays. Accurate visual inspection of the samples excluded by the KS test showed a high concentration of genomic DNA and laser microdissection (where an additional cDNA amplification with random primers is usually performed) experiments. Finally, in order to condense the dataset to representative samples, similar arrays were grouped and a representative for each cluster was kept (details in (Mutwil et al., 2011), supplementary material), with a final number of 274 samples. These samples were then normalized using tRMA (Giorgi et al., 2010), using the updated probe-to-gene annotation from CustomCDF v13.0, (Dai et al., 2005).

In order to reduce the number of genes to a golden and reliable subset (and to decrease the computational load), I excluded all the probesets that were measured consistently in all, or almost all, samples. To do so, I kept only the 3350 genes that were deemed as "Present" (i.e. having an expression significantly discernable from background signal) in at least 90% of the samples, using the probeset Present/Absent call analysis (Schuster et al., 2007).

Gene network reconstruction was performed using Java implementations for Pearson correlation, Mutual Information Partial Correlation and the LASSO (see Paragraph 4.6), except where specifically noted. The Mutual Information indices (except for Figure 25) are normalized through dividing the unnormalized index by the average of the max self-to-self M.I. index of the two genes analyzed. This was done to allow a fair comparison between indices independently of the original expression distribution of the genes (in Figure 25 this was unnecessary, since we were conducting an exploratory comparative screening between Mutual Information and Pearson Correlation, and not a global network reconstruction).

The LASSO solution is obtained by the LARS algorithm (Efron et al., 2004), using a pre-calculated Gram matrix (Weisstein, 2011) shared by all the models, in order to obtain a fast and parallel global network reconstruction for large gene datasets. In the context of Partial correlation, a first order threshold of "zero" means the removal of an edge that switches sign between its zeroth and first order correlation coefficient, while a first order threshold of "no" correspond to the zeroth order correlation coefficient. For the LASSO the Least Angle Regression algorithm was used (Efron et al., 2004); in order to select the best LASSO model for every gene, I calculated ten-fold cross-validated prediction errors for all models, at every L1 (the sum of absolute weights, see Paragraph 2.4.2), and kept the model with the lowest error for a particular gene. Gene category network enrichment was calculated using a Fisher's Exact test (Upton, 1992) with a p-value cutoff of 0.05, using the 2010 *Arabidopsis thaliana* ontology from MapMan (Usadel et al., 2009) and comparing the edge abundance in the network to the theoretical abundance expected by the ontology groups present in the dataset.

For evaluating convergence to the protein-protein interactome, I calculated Accuracy over the AtPin experimentally validated protein-protein interaction data (Brandão et al., 2009), and excluding proteins not present in the 3350 genes dataset. The fit of the degree distribution to a power-law distribution was calculated as in (Brohée et al., 2008). The Ontology Agreement score percentage of the networks was obtained by counting the number of edges containing two genes with at least one shared MapMan ontology term (Usadel et al., 2009). Due to the highly grained nature of the MapMan bins, I decided to trim the ontology to the third branch (i.e. bin 1.3.1.10 would become 1.3.1). The total percentage agreement is then calculated by dividing the number of agreeing edges by the total number of edges.

4.6 Candidate selection through LASSO and Correlation analysis

4.6.1 *RHM2* - Selection of candidate genes

Here I used the filtered *Arabidopsis thaliana* microarray dataset as in Paragraph 4.5, resulting in 5750 samples. Contrarily to what described in Paragraph 4.5, I didn't condense the dataset, and simply normalized the 5750 remaining samples using RMA and an updated probeset annotation [CustomCDF v12.0, (Dai et al., 2005)]. With this amount of samples, the artifact effects that we described in Paragraph 2.2 would be almost absent for this kind of normalization. As in Paragraph 4.5, in order to reduce the number of genes to a computationally solvable problem, I kept all the probesets that were measured consistently in at least 90% of the samples, ending up with 8237 genes.

As central "bait" for our coexpression analysis I used *RHM2/MUM4* (At1g53500), whose expression is entirely and uniquely measured by the 11 probes of Affymetrix probeset 260985_at. In order to focus our analysis on *RHM2*, I extracted from our dataset the highest (using absolute correlation coefficients) 3000 coexpressors of At1g53500.

I included in the final candidate list genes derived from 4 different coexpression analyses. The first group included the top 30 Pearson correlators using absolute correlation coefficients with At1g53500. The second group included the top 30 correlators using full order Partial Pearson calculated by correlation matrix inversion (Whittaker, 2009). The third group contained the top 30 correlators using Partial correlation calculated via the shrunk correlation matrix approach from (Opgen-Rhein and Strimmer, 2007). I then calculated a LASSO model (Tibshirani, 1996) using At1g53500 as dependent variable and the other 8236 genes as explanatory variables. The lasso2 R package (www.r-project.org) was used for the computations, with default parameters. From this LASSO model, I extracted candidates at L1 thresholds equal to 1%, 2%, 3%, 4% and 5% of the maximum unbound L1 threshold and included these into the fourth and final group of gene candidates.

4.6.2 Multi-gene mucilage networks

I selected the Affymetrix AtGenExpress (Schmid et al., 2005) (GEO accession: GSE5634) seed and silique developmental series, comprising 24 samples, and normalized it via tRMA (Giorgi et al., 2010) with the CustomCDF v12.0 probeset annotation (Dai et al., 2005). Contrarily to the procedure described in paragraph

4.5 and 4.6.1, the reduced number of samples allowed for a full LASSO modeling over all the almost 21504 (CustomCDF) genes, and therefore I didn't apply any Present/Absent call filtering. The models generated by a number of samples so small are also including less variables and are therefore simpler to interpret. This happens because the model will stop exploring the variable space when a number of predictor variables equaling the number of samples minus 2 has been included. At this stage, the model cannot proceed without becoming underdetermined. Therefore, I adopted the heuristic solution to include all LASSO genes introduced by the modeling at any particular LASSO step, and generated a group of putative candidates for every bait gene.

4.6.3 *StHRE1* and *StHRE2a/b* networks

I collected all the 14 samples from the two publicly available potato developmental datasets (Kloosterman et al., 2008; Ferreira et al., 2010). Both datasets used the POCI array as hybridization platform (pgrc.ipk-gatersleben.de/poci/), a 44K 60-mer Agilent oligo array designed from known EST libraries (Kloosterman et al., 2008). The first dataset (Kloosterman et al., 2008) comprises 6 experiments taken from pooled samples at different stages of potato tuber development, specifically at 0 (unswollen stolon), 5, 6-7, 7-8, 9-10 and 15 days after switching from a 16h to a 8h light period. The second dataset (Ferreira et al., 2010), publicly available on ArrayExpress (Parkinson et al., 2007) entry E-MEXP-2482, follows the same sampling conventions as the previous one, but measures two distinct biological pools of 4 stages (0 days, 6-7 days, 7-8 days and 9-10 days) for a total of 8 samples. Both datasets were normalized and quality filtered as described by (Kloosterman et al., 2008) and merged for subsequent analysis.

Of the 42,034 unique probes present on the POCI array I excluded those displaying low signals as in (Kloosterman et al., 2008) and conducted our coexpression analysis on a total of 31,293 probes. *StHRE1* is represented on this chip by three probes (MICRO.3799.C2, MICRO.3799.C3 and ACDA02245D01.T3m.scf), all possessing analogous behavior across the experiments, and therefore a mean of the three was used as the *StHRE1* representative signal. *StHRE2* isoforms both perfectly hybridize to the probe *bf_suspxxxx_0025D01.t3m.scf*.

For both genes, I extracted the top 10 Spearman correlators and calculated LASSO models using all other probes as explanatory variables. In order to increase the final number of candidate partners using the LASSO I didn't keep only the best cross-validated model (see Paragraph 4.5), but I included all genes that were introduced in any LASSO modeling step, obtaining a final count of 23 interactors for *StHRE1* and 15 for *StHRE2a/b*.

4.7 Sugar screening in *Arabidopsis thaliana* seed coat mucilage

Note: Seed coat mucilage extraction, hydrolysis of seed coat mucilage and monosaccharide measurements were carried out by Mr. Aleksandar Vasilevski (Max Planck Institute of Molecular Plant Physiology) as described in (Usadel et al., 2004). A High Performance Anion Exchange Chromatography with Pulsed

Amperometric Detection (HPAEC-PAD) (Ip et al., 1992) system was used via a DIONEX ISC-3000 machine for monosaccharide quantification.

4.7.1 Seed Staining and Microscopy

Arabidopsis thaliana ecotype Columbia-0 seeds were stained with a solution of Ruthenium Red / water 0.01% (weight/volume) for 5-10 minutes under mild shaking. Seeds were visualized using a Leica MZ 12,5 Stereomicroscope (Software: Leica Application Suite).

4.8 Gene expression in *Solanum tuberosum* tubers

Note: I acknowledge Dr. Francesco Licausi (Scuola Superiore Sant'Anna, Pisa) for providing the expertise and materials necessary for the molecular biology sections of this paragraph.

4.8.1 Phylogenetic analysis of *StHRE* genes

The phylogenetic analysis for ERF group VII involved 39 ERF protein sequences, with prefix indicating the species: *Solanum tuberosum* (StHRE1, StHRE2a and StHRE2b), *Arabidopsis thaliana* [AtHRE1 (At1g72360), AtHRE2 (At2g47520), AtRAP2.2 (At3g14230), AtRAP2.3 (At3g16770) and AtRAP2.12 (At1g53910)], *Populus trichocarpa* (PtERF-B2-1, PtERF-B2-2, PtERF-B2-3, PtERF-B2-5, PtERF-B2-6) as named according to Zhao et al. (2007), *Oryza sativa* (OsERF059, OsERF060, OsERF061, OsERF062, OsERF064, OsERF065, OsERF066, OsERF067, OsERF068, OsERF069, OsERF070, OsERF071, OsERF072, named according to (Nakano et al., 2006) and OsSub1A, OsSub1B, OsSub1C, named according to (Fukao and Bailey-Serres, 2008), *Vitis vinifera* [VvERF057, VvERF058, VvERF059, named according to (Licausi et al., 2010)].

The sequences were aligned using MUSCLE (Edgar, 2004) and the phylogenetic tree was inferred using the Maximum Likelihood method based on the JTT matrix-based model (Jones et al., 1992). The latter operation was performed using the MEGA5 software (Tamura et al., 2007).

4.8.2 Tuber growing conditions

Desirée cultivar *Solanum tuberosum* plants were obtained from Saatzucht Lange AG, then were maintained in tissue culture and therefore transferred to greenhouse as described by (Fernie et al., 2002). Developing tubers (Kloosterman et al., 2008) were harvested from healthy 10-week-old plants. High oxygen treatments were carried on in plastic bags applying a stream of premixed air containing 40% O₂, 350 ppm CO₂, and N₂ (Air Liquide, Berlin, Germany) for 12 hours in darkness. Identical conditions, with fluxes of 21% O₂, were applied to control normoxic plants. The actual oxygen concentration inside the bag was checked by means of a phosphorescent oxygen sensor (Presens, Regensburg, Germany).

4.8.3 Sequencing of *StHRE* mRNAs

StHREs were amplified using specific degenerated primers (Table 14) that anneal to the N-terminus of the coding sequence and the conserved region encoding the AP2/ERF DNA-binding domain. The resulting amplicon was sequenced using a T7 promoter primer, as described by the manufacturer. The full length sequence of *StHRE1* and *StHRE2a* and *StHRE2b* was confirmed using a combination of primers annealing to

the N-terminus and C-terminus of the sequences AB085820 (CIP353, (Mine et al., 2003)) and U77655 (STWAAEIRD, (Campbell et al., 1998)).

4.8.4 mRNA extraction

The harvested tubers were immediately frozen in liquid nitrogen, ground into a fine powder using a ball-mill (Retsch, Haan, Germany), and stored at -80°C . Total RNA was isolated from 100 mg ground plant material using the QiagenRNeasy Plant Minikit (Qiagen, Hilden, Germany).

4.8.5 Expression measurement through Realtime RT-qPCR

RNA quality was assessed by agarose gel electrophoresis prior to DNaseI digestion (Promega, Mannheim, Germany) and 1 μg total RNA was used for cDNA synthesis using the Superscript III RT-PCR kit (Invitrogen, Darmstadt, Germany). Quantitative real-time RT-PCR was performed as described by (Czechowski et al., 2004) using the primer pairs described in Table 1. The genes coding for elongation factor 1- α (AB061263) and tubulin (609267) were used as housekeeping genes according to (Nicot et al., 2005). Primers were designed using the Quantprime software (Arvidsson et al., 2008).

| Gene | Forward primer (5'-3') | Reverse primer (5'-3') |
|-----------|--------------------------|--------------------------|
| dgHRE | ATGTGTGGTGGTGCHATHMTY | GCAGCTTCTTCWGCAGTGTTGAA |
| StADH | TGTTGGATGTGTCGCCAAA | GGCCTGTCGAGATTCCACAA |
| StSus4 | GCAAATATATTTTATCTTAATAAG | GAAGTGTGAAGAATTTGAATAGC |
| StHRE1 | TGATTTCTGGCCAACTTCCAC | TGCCTCTTCTTCATCTGCTCA |
| StHRE2a | TGGCAAACCTTCTTCTTTTCCA | TGCCTCTTCTTCATCTGCTCA |
| StHRE2b | TCTGCTGATTTCTGGCCAAC | TCTGCTGATTTCTGGCCAAC |
| StTubulin | CAGACCTGAGGAAATTGGCTG | TTCTTGGCATCCCACATTTGT |
| StEF1a | CATTGCTTGCTTTCACCCTTGGTG | CCTAGCCTTGGAGTACTTGGGGG |
| Stu.10537 | CACACCGGCACGCATATTGATG | TGTCGACATCAAAGCCAGCATCG |
| Stu.18907 | GGACGGTTCCTCCATCTAAAGAGC | AAACAGCGTGACAACGAAGTGC |
| Stu.20202 | TTAGTGGACACAGCGAGCAACG | TAGCTCCTTCACTGTGGGTTCCG |
| Stu.22430 | AGGAGCTACTCTGAAGGTGGATG | TCCACTGATATACTGGGCATCGTC |
| Stu.4930 | TTCAATGCTCCGACCCGGATTCC | TGATCGAACACCGACTCCAC |
| Stu.5337 | AGCGAATGCGTGAAGCTGATCC | TCCGGTGACATGACGAACTTCC |
| Stu.6717 | TGTGATGCCAGAGTTCACACAGC | AAGCTTAGAGCCAGAGCCACTC |
| Stu.7147 | TGTGCTCTTGTTCCTCAGCTTGG | AGCGGTGCTCTGATTGTTTCCG |
| Stu.7176 | TCCAAGAGCAGCACTTACAATGCC | ACAAGGACACAGAGGGTGTTTAGC |
| Stu.15767 | AGGAGCGTTTTCCGCACTATCAC | TCGTCCGTGACATCATACACCAG |
| Stu.16271 | TCTAACGTGTCCACTGAGCATCG | AACAAGGCCGGAGCGATATTCC |
| Stu.18198 | TTGCAGTAGGTTGGGCACTTCC | AGCCACATGCAAGCCAAATGAG |
| Stu.18546 | ACGGTTTGTGCTCCATCATGCC | AGCGGTGCCTTCTCTTTCATC |
| Stu.2176 | TCATGTGCATGTGGAGCCATCAC | ACAGTAGGCAGAGCTGGGATTG |
| Stu.22430 | AGGAGCTACTCTGAAGGTGGATG | TCCACTGATATACTGGGCATCGTC |
| Stu.22641 | ATGTGTCCGTGGTTGACCTCAC | TTCTCCTTGATGGCAGCTTTG |
| Stu.22678 | AGTCCTCCCAAGAGGAATCCTTG | TCATCCAGGAGTGGTGAACGTG |
| Stu.3076 | ACATCCCAATAGCCGTGTCCAAG | TGCTTTCAGCTGGGCAAAGGAG |
| Stu.4348 | CCCGAGAAACCAATCCAGTTGAAG | ATCCGGTGTGTGGTCTTCTGC |
| Stu.4665 | CAGCATTGCAGCAGAGAAGAAGC | ATCCCTTGGCTGACCAATGTCC |
| Stu.4779 | CCTTCCATCGCAATGAAGTTCTC | AGTGAGTAGCTGATCCGCGTTG |
| Stu.9387 | GTCAGCAGCCGGACTTTATGATTC | ACGGGCTACAGCATGTCCAAAC |

Table 14 - List of primers used for amplifying StHREs and their co-regulated genes

5. Appendix

5.1 Sequence length distribution for the *Arabidopsis thaliana* Transcriptome

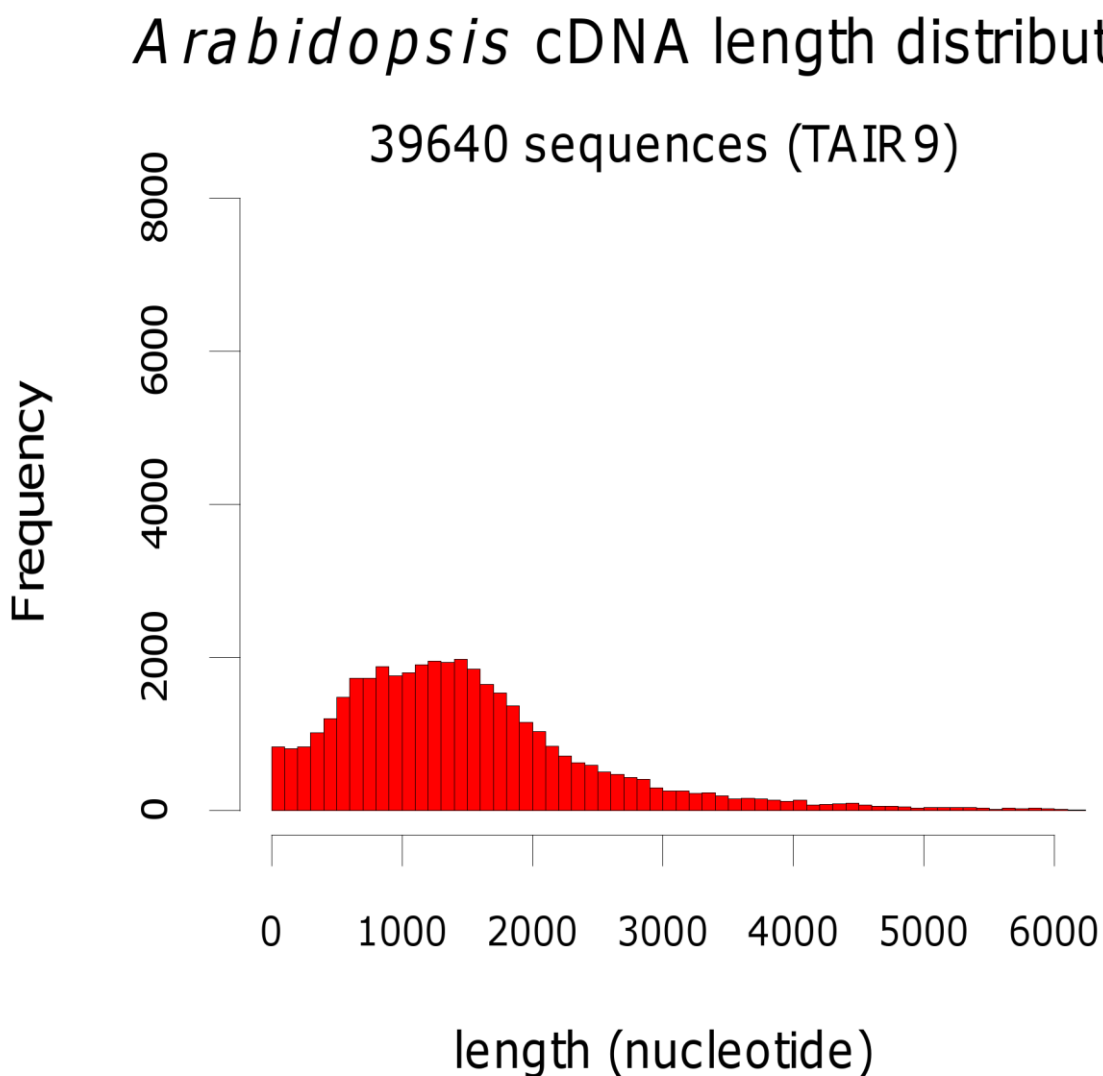


Figure 48 - Sequence length distribution for the 39,640 cDNAs collected in TAIR v9 for *Arabidopsis thaliana* (and therefore collecting multiple transcripts per gene model) (Swarbreck et al., 2008)

5.2 Average Affymetrix inter-array correlation coefficients at different sample sizes, using three different microarray normalization procedures

| Sample size | Original arrays | | | Permutated arrays | | |
|--------------------|------------------------|------------|--------------|--------------------------|------------|--------------|
| | MAS5 | RMA | GCRMA | MAS5 | RMA | GCRMA |
| 2 | 0.7704 | 0.8113 | 0.7825 | 3.31^{-4} | 0.0456 | 0.1795 |
| 3 | 0.7721 | 0.8547 | 0.8524 | -9.88^{-5} | 0.7761 | 0.9144 |
| 4 | 0.7601 | 0.8220 | 0.8039 | -4.63^{-4} | 0.1154 | 0.2647 |
| 5 | 0.7604 | 0.8460 | 0.8432 | 3.01^{-5} | 0.4316 | 0.7140 |
| 6 | 0.7557 | 0.8144 | 0.7949 | -5.90^{-5} | 0.1086 | 0.2544 |
| 7 | 0.7584 | 0.8357 | 0.8212 | -1.09^{-5} | 0.2481 | 0.5436 |
| 8 | 0.7555 | 0.8210 | 0.7981 | 1.00^{-4} | 0.0913 | 0.2274 |
| 9 | 0.7529 | 0.8334 | 0.8125 | 3.46^{-5} | 0.1587 | 0.4140 |
| 10 | 0.7536 | 0.8163 | 0.8048 | 1.20^{-4} | 0.0757 | 0.1929 |
| 11 | 0.7536 | 0.8242 | 0.8097 | -1.01^{-4} | 0.1118 | 0.3170 |
| 12 | 0.7552 | 0.8156 | 0.8000 | 9.03^{-5} | 0.0637 | 0.1665 |
| 13 | 0.7523 | 0.8264 | 0.8061 | 1.85^{-5} | 0.0843 | 0.2546 |
| 14 | 0.7531 | 0.8154 | 0.8014 | 6.90^{-5} | 0.0547 | 0.1482 |
| 15 | 0.7537 | 0.8217 | 0.8087 | 2.82^{-5} | 0.0659 | 0.2048 |
| 16 | 0.7511 | 0.8146 | 0.8052 | 7.68^{-5} | 0.0467 | 0.1307 |
| 17 | 0.7511 | 0.8232 | 0.8062 | -6.80^{-5} | 0.0540 | 0.1654 |
| 18 | 0.7532 | 0.8159 | 0.8022 | -7.89^{-5} | 0.0411 | 0.1150 |
| 19 | 0.7546 | 0.8247 | 0.8059 | 9.65^{-6} | 0.0460 | 0.1418 |
| 20 | 0.7592 | 0.8228 | 0.8024 | 4.69^{-6} | 0.0364 | 0.1008 |
| 30 | 0.7602 | 0.8208 | 0.8022 | 3.09^{-5} | 0.0226 | 0.0620 |
| 50 | 0.7578 | 0.8233 | 0.8038 | 7.72^{-6} | 0.0124 | 0.0326 |
| 100 | 0.7543 | 0.8190 | 0.8023 | -2.41^{-8} | 0.0058 | 0.0157 |

Table 15 - Average inter-array correlation coefficients at different sample sizes, using three different normalization procedures.

5.3 Example of Breaking Potential calculation

Data generation

Expression matrix generated using gaussian distribution of interdependent variables as in Figure 15, panel **a.**, yielding a correlation network as in panel **b.**, and the removal of edges upon conditioning as in panel **c.**

Pearson correlation matrix:

| | V1 | V2 | V3 | V4 | V5 |
|----|--------------|--------------|--------------|--------------|----|
| V1 | 1 | | | | |
| V2 | 0.635 | 1 | | | |
| V3 | 0.827 | 0.863 | 1 | | |
| V4 | 0.703 | 0.816 | 0.899 | 1 | |
| V5 | 0.775 | 0.805 | 0.904 | 0.787 | 1 |

In **bold**, direct correlations, in **red**, indirect (spurious) correlations.

Pairwise first order partial correlations:

Syntax: $\text{pcor}(x,y,z)$ = *partial correlation of x and y, conditioned on z*

$$r_{xy.z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1-r_{xz}^2)(1-r_{yz}^2)}}$$

Only the significant edges ($r > 0.700$, $r^2 > 0.490$) in the previous correlation matrix are tested.

Edges are considered to be "broken" when $r < 0.221$, $r^2 < 0.049$

A sign inversion is also considered by our definition to yield a broken edge

| Edge | Partial correlation | Broken edge? | Breaking vertex |
|----------------|---------------------|--------------|-----------------|
| pcor(V1,V3,V2) | 0.716 | no | |
| pcor(V1,V3,V4) | 0.627 | no | |
| pcor(V1,V3,V5) | 0.467 | no | |
| pcor(V1,V4,V2) | 0.414 | no | |
| pcor(V1,V4,V3) | -0.166 | yes | V3 |
| pcor(V1,V4,V5) | 0.237 | no | |
| pcor(V1,V5,V2) | 0.577 | no | |
| pcor(V1,V5,V3) | 0.115 | yes | V3 |
| pcor(V1,V5,V4) | 0.507 | no | |
| pcor(V2,V3,V1) | 0.779 | no | |
| pcor(V2,V3,V4) | 0.511 | no | |
| pcor(V2,V3,V5) | 0.534 | no | |
| pcor(V2,V4,V1) | 0.673 | no | |
| pcor(V2,V4,V3) | 0.182 | yes | V3 |
| pcor(V2,V4,V5) | 0.497 | no | |
| pcor(V2,V5,V1) | 0.641 | no | |
| pcor(V2,V5,V3) | 0.115 | yes | V3 |
| pcor(V2,V5,V4) | 0.456 | no | |
| pcor(V3,V4,V1) | 0.794 | no | |
| pcor(V3,V4,V2) | 0.666 | no | |
| pcor(V3,V4,V5) | 0.710 | no | |
| pcor(V3,V5,V1) | 0.740 | no | |
| pcor(V3,V5,V2) | 0.698 | no | |
| pcor(V3,V5,V4) | 0.727 | no | |
| pcor(V4,V5,V1) | 0.539 | no | |
| pcor(V4,V5,V2) | 0.379 | no | |
| pcor(V4,V5,V3) | -0.136 | yes | V3 |

Breaking vertices count

V3 $n_{brk} = 5$

Breaking Potential calculation

Number of connections between V3 neighbors in 0th order network $n_{eff} = 5$

$$brk = \frac{n_{brk}}{n_{eff}}$$

$$brk_{V3} = \frac{5}{5} = 1$$

Centrality measures summary

| | Degree | Clustering Coefficient | Betweenness | Breaking Potential |
|-----------|--------|------------------------|-------------|--------------------|
| V1 | 3 | 1.0 | 0.0 | 0.0 |
| V2 | 3 | 1.0 | 0.0 | 0.0 |
| V3 | 4 | 0.833 | 0.333 | 1.0 |
| V4 | 4 | 0.833 | 0.333 | 0.0 |
| V5 | 4 | 0.833 | 0.333 | 0.0 |

5.4 Breaking Potential and other Centrality measures assessed for essential gene prediction power

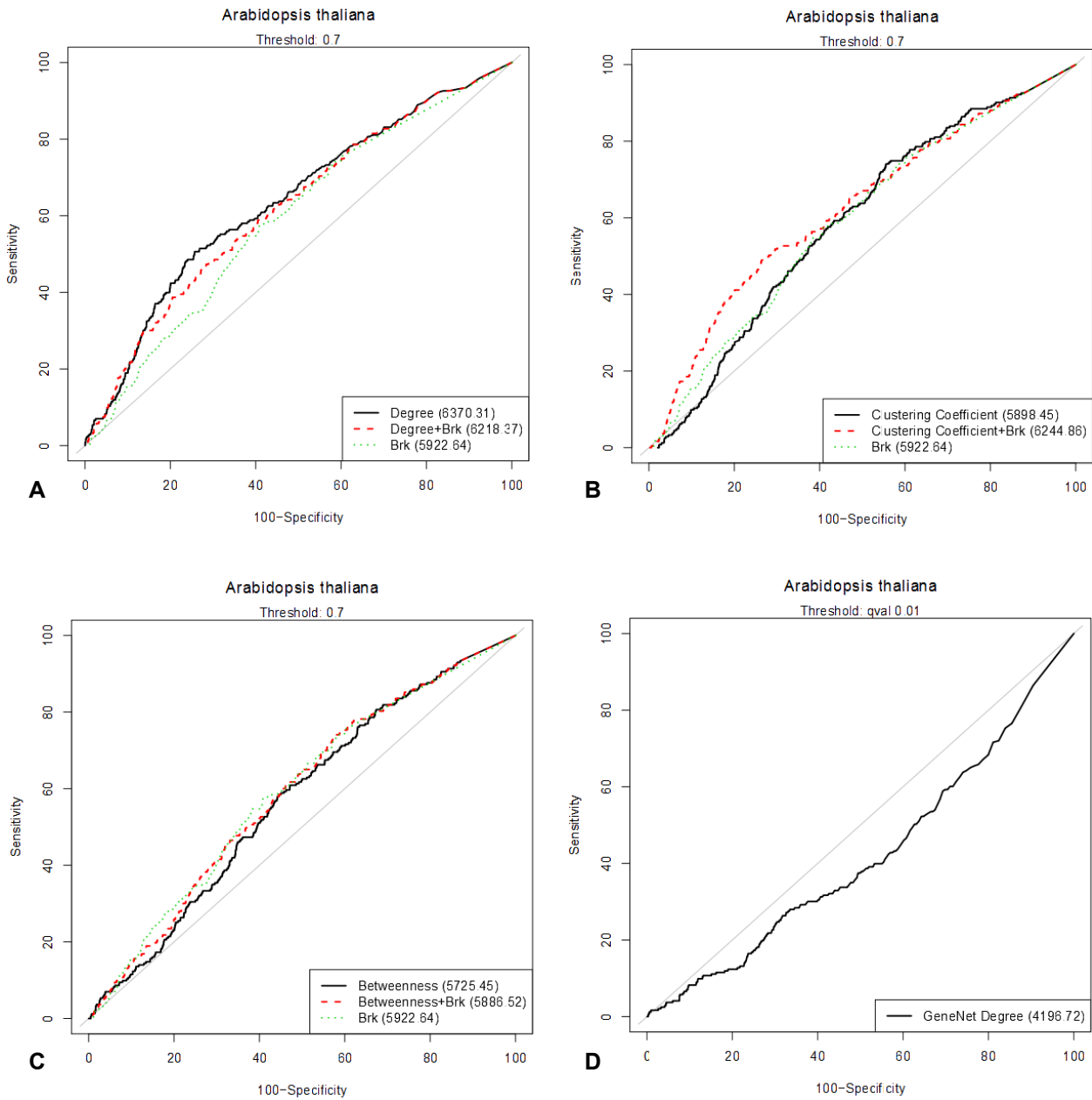


Figure 49 - ROC curves for essential gene prediction using Breaking Potential alone and coupled with degree (panel A), clustering coefficient (panel B) and betweenness (panel C). The GeneNet approach is shown as a counter-indicator for essentiality in panel D.

5.5 Full coding sequences of *StHRE1*, *StHRE2a* and *StHRE2b*.

>StHRE1

ATGTGTGGAGGTGCCATAATCTCCGATTATGAGCCCGCCGAAACTTCTACCGGAAACTCTCTGCTCGTGACCTGTGGGCT
 GAGCTGGACCCTATCTCCGACTACTGGTCCTCTTCCCTCCTCATCCTCAACTGTGCGAAAACCCCTTATTCCGCTCAGTCGCCG
 GTGACTCACTCCGTCGATAAGCCTAAGAAATCAGATTCGGGCAAATCTAATCAACTCAAAAAAGGTAATAAGACTGTGAAG
 GTTGAGAAGGAGAAGAGTACTGGACCAAGGCAGAGAAAAGAACAAGTACAGAGGAATAAGGCAGAGACCATGGGGAAAATGG
 GCTGCTGAGATTTCGCGATCCTCAGAAGGGTGTCCGTGTTGGCTTGGTACATTCAACACAGCAGAGGATGCTGCCAGAGCC
 TATGATGAGGCTGCTAAGCGCATTTCGTGGTAACAAGGCCAAACTCAACTTCCCTGCCCCATCACCACCTGCTAAGCGACAG
 TGCCTAGCACTGTGCTGCTGATCCTCCACCAGCACTACTCCTTGAGAGTTCTAACATAATATCTTATAACAATTCTCCT
 TTAATGAACCTTCGGATATGATGTTTCAGAGCCAAACTCCCTACTACCCAATGGAAATGCCCGTTGCTAGTGATGATTATGAA
 CTTCAAGGAACAGATTTCCAACCTTGGAAATCGTTCCCTGGAATTGGAGCCAGCAGATTTCATCTGATCAGTTTTTCAGGGATCGTC
 GATCCTGATCCTCTTAATGTTTTTCTGATGGAGGATTTTGCTTCAACTCAGCATCAGTTCTATTGA

>StHRE2a

ATGTGTGGTGGTGCAATTCTTTATGATATTATTCCCTCGTGACCGCCGTTTGTTCATCCACCGACTTATGGCCAAGCTCTGCT
 GATTTCTGGCAAACCTTCTTCTTTTTTCCAAGCCAATTTCCACCCAAAATGTTCCCTCCCAAGCCTAAACGAGCTCAACTCTCT
 AGAGGTAGTGAGCAGATGAAGAAGAGGCAAAGGAAGAATCTTTACAGGGGAATCCGACAACGTCCATGGGGTAAATGGGCT
 GCTGAAATTCGTGACCCGAGAAAAAGGGTTAGGGTCTGGTTAGGTACTTTCAACACTGCTGAAGCTGCAAGAGCTTATGAT
 AGAGAAGCTCGTAAAATCAGGGGAAAGAAAGCTAAAGTTAATTTCCCAATGAAGACGACGACCACTACTACAGTCATCCA
 GAGCCGCCTCCTTTGAACATTGTTTATGAATCTTATGATACTACTAGTACTTACAATCAAGAATCAAATAACTGTTACCCC
 TTCCACTCAATCGAAAACACTGAACCTGTTATGGAATTCGCAATTGCTAACAAAAATTCATCTGGGTCTGCTTATAATGGA
 ATTTGAAGATCAGAATGTGGAAGGAGAAGAGCAGACGGTGAAAAATTCAAATAACAGGATCGTAGAGGAAGAGGAAAAACA
 GAGGATGAAGTGCAGATACTTTCTGATGAACTGATGGCTTATGAGTCATTGATGAAGTTCTATGAAATACCGTATGTTGAC
 GGGCAATCAGTGGCGGCACGGTGAATCCAGCGGCGGACACCGAAGTGGGCGGTGGCTCGATGGAGCTTTGGAGTTTTGAT
 GATGTTAGTCGTCTACAACCAAGTTATAATGTTAGTTTGGATTATTGTTTTGTTTAAATTGTTGCATCTTTTTAGTTTGCTG
 AATTAG

>StHRE2b

ATGTGTGGTGGTGCAATTCTTTATGATATTATTCCCTCGTGACCGCCGTTTGTTCATCCACCGACTTATGGCCAAGCTCTGCT
 GATTTCTGGCAAACCTTCTTCTTTTTTCCAAGCCAATTTCCACCCAAAATGTTCCCTCCCAAGCCTAAACGAGCTCAACTCTCT
 AGAGGTAGTGAGCAGATGAAGAAGAGGCAAAGGAAGAATCTTTACAGGGGAATCCGACAACGTCCATGGGGTAAATGGGCT
 GCTGAAATTCGTGACCCGAGAAAAAGGGTTAGGGTCTGGTTAGGTACTTTCAACACTGCTGAAGCTGCAAGAGCTTATGAT
 AGAGAAGCTCGTAAAATCAGGGGAAAGAAAGCTAAAGTTAATTTCCCAATGAAGACGACGACCACTACTACAGTCATCCA
 GAGCCGCCTCCTTTGAACATTGTTTATGAATCTTATGATACTACTAGTACTTACAATCAAGAATCAAATAACTGTTACCCC
 TTCCACTCAATCGAAAACACTGAACCTGTTATGGAATTCGCAATTGCTAACAAAAATTCATCTGGGTCTGCTTATAATGGA
 ATTTGAAGATCAGAATGTGGAAGGAGAAGAGCAGACGGTGAAAAATTCAAATAACAGGATCGTAGAGGAAGAGGAAAAACA
 GAGGATGAAGTGCAGATACTTTCTGATGAACTGATGGCTTATGAGTCATTGATGAAGTTCTATGAAATACCGTATGTTGAC
 GGGCAATCAGTGGCGGCACGGTGAATCCAGCGGCGGACACCGAAGTGGGCGGTGGCTCGATGGAGCTTTGGAGTTTTGAT
 GATGTTAGTCGTCTACAACCAAGTTATAATGTTAGTTTGGATTATTGTTTTGTTTAAATTGTTGCATCTTTTTAGTTTGCTG
 AATTAG

5.6 MapMan ontology bin enrichment analysis for top correlators of *StHRE1* and *StHRE2a/b*

Enrichment for *StHRE1* correlators (Spearman correlation coefficient >0.9, 570 genes)

| bin | bin name | count | p-value |
|----------|--|-------|-----------|
| 29.5.4 | protein.degradation.aspartate protease | 7 | 1.99E-005 |
| 35.2 | not assigned.unknown | 193 | 4.88E-005 |
| 35 | not assigned | 242 | 6.09E-005 |
| 27.3.41 | RNA.regulation of transcription.B3 transcription factor family | 4 | 9.95E-005 |
| 7.2 | OPP.non-reductive PP | 3 | 1.87E-004 |
| 29.5 | protein.degradation | 47 | 2.02E-004 |
| 2.2.2 | major CHO metabolism.degradation.starch | 6 | 3.86E-004 |
| 3.5.1 | minor CHO metabolism.others.Xylose isomerase | 2 | 4.17E-004 |
| 27 | RNA | 68 | 7.87E-004 |
| 26.23 | misc.rhodanese | 3 | 8.26E-004 |
| 7.2.2 | OPP.non-reductive PP.transaldolase | 2 | 8.28E-004 |
| 13.2.5.3 | amino acid metabolism.degradation.serine-glycine-cysteine group.cysteine | 2 | 8.28E-004 |
| 29.8 | protein assembly and cofactor ligation | 4 | 9.45E-004 |
| 11.10.4 | lipid metabolism.glycolipid synthesis.sulfolipid synthase | 2 | 1.37E-003 |
| 2 | major CHO metabolism | 10 | 1.57E-003 |
| 3.3 | minor CHO metabolism.sugar alcohols | 2 | 2.04E-003 |
| 16.8 | secondary metabolism.flavonoids | 6 | 2.91E-003 |
| 7 | OPP | 4 | 2.92E-003 |
| 10.1.1 | cell wall.precursor synthesis.NDP sugar pyrophosphorylase | 2 | 3.74E-003 |
| 29 | protein | 84 | 4.40E-003 |
| 27.4 | RNA.RNA binding | 10 | 8.69E-003 |

Enrichment for *StHRE1* correlators (Spearman correlation coefficient >0.7, 14758 genes)

| bin | bin name | count | p-value |
|---------------|--|-------|----------|
| 29 | protein | 1909 | 3.82E-21 |
| 29.5 | protein.degradation | 866 | 2.01E-15 |
| 35 | not assigned | 7847 | 5.10E-12 |
| 29.5.11 | protein.degradation.ubiquitin | 571 | 1.15E-10 |
| 10 | cell wall | 112 | 1.91E-09 |
| 26 | misc | 529 | 2.27E-09 |
| 35.1 | not assigned.no ontology | 1435 | 1.12E-08 |
| 29.5.11.4.3.2 | protein.degradation.ubiquitin.E3.SCF.FBOX | 152 | 1.90E-08 |
| 27.1 | RNA.processing | 214 | 2.56E-07 |
| 16.2 | secondary metabolism.phenylpropanoids | 28 | 2.68E-07 |
| 29.5.11.4.3 | protein.degradation.ubiquitin.E3.SCF | 164 | 3.36E-07 |
| 29.5.11.4 | protein.degradation.ubiquitin.E3 | 386 | 3.94E-07 |
| 20.1 | stress.biotic | 166 | 8.39E-07 |
| 29.3 | protein.targeting | 184 | 1.33E-06 |
| 27.3.25 | RNA.regulation of transcription.MYB domain transcription factor family | 9 | 5.20E-06 |
| 16 | secondary metabolism | 145 | 5.65E-06 |
| 16.2.1 | secondary metabolism.phenylpropanoids.lignin biosynthesis | 14 | 1.16E-05 |
| 10.6 | cell wall.degradation | 32 | 1.32E-05 |
| 13.1.7 | amino acid metabolism.synthesis.histidine | 15 | 2.01E-05 |
| 17 | hormone metabolism | 192 | 3.90E-05 |
| 29.5.11.4.1 | protein.degradation.ubiquitin.E3.HECT | 14 | 5.21E-05 |
| 29.2.3 | protein.synthesis.initiation | 89 | 5.96E-05 |

Enrichment for StHRE2a/b correlators (Spearman correlation coefficient >0.9, 4 genes)

| bin | bin name | count | p-value |
|-----|----------------------|-------|----------|
| 2 | major CHO metabolism | 2 | 1.87E-04 |

Enrichment for StHRE2a/b correlators (Spearman correlation coefficient >0.7, 927 genes)

| bin | bin name | count | p-value |
|------------|--|-------|-----------|
| 2 | major CHO metabolism | 16 | 8.27E-005 |
| 2.2.1.1 | major CHO metabolism.degradation.sucrose.fructokinase | 4 | 8.53E-005 |
| 16.1.3 | secondary metabolism.isoprenoids.tocopherol biosynthesis | 4 | 2.07E-004 |
| 1.2.7 | PS.photorespiration.glycerate kinase | 2 | 3.71E-004 |
| 13.1.7 | amino acid metabolism.synthesis.histidine | 4 | 4.22E-004 |
| 13 | amino acid metabolism | 25 | 1.18E-003 |
| 13.2.6.2 | amino acid metabolism.degradation.aromatic aa.tyrosine | 4 | 1.46E-003 |
| 2.2 | major CHO metabolism.degradation | 10 | 1.65E-003 |
| 21 | redox.regulation | 14 | 3.06E-003 |
| 26.23 | misc.rhodanese | 3 | 3.31E-003 |
| 13.1.6.1.5 | amino acid metabolism.synthesis.aromatic aa.chorismate.shikimate kinase | 2 | 5.29E-003 |
| 16.1.3.3 | secondary metabolism.isoprenoids.tocopherol biosynthesis.MPBQ/MSBQ methyltransferase | 2 | 5.29E-003 |
| 2.1.2 | major CHO metabolism.synthesis.starch | 6 | 5.74E-003 |
| 27 | RNA | 97 | 5.74E-003 |
| 13.2.6 | amino acid metabolism.degradation.aromatic aa | 5 | 6.00E-003 |
| 2.2.1 | major CHO metabolism.degradation.sucrose | 6 | 7.18E-003 |
| 3.4 | minor CHO metabolism.myo-inositol | 3 | 7.33E-003 |
| 27.4 | RNA.RNA binding | 14 | 8.88E-003 |
| 20.1 | stress.biotic | 5 | 9.45E-003 |
| 17.3.1.2.2 | hormone metabolism.brassinosteroid.synthesis-degradation.sterols.SMT | 2 | 9.62E-003 |

5.7 Expression Intensity and Variance for Transcriptional genes and Essential genes in Arabidopsis thaliana Affymetrix microarrays

See Figure 30A in the main text for an assessment of expression average intensity of Transcriptional genes.

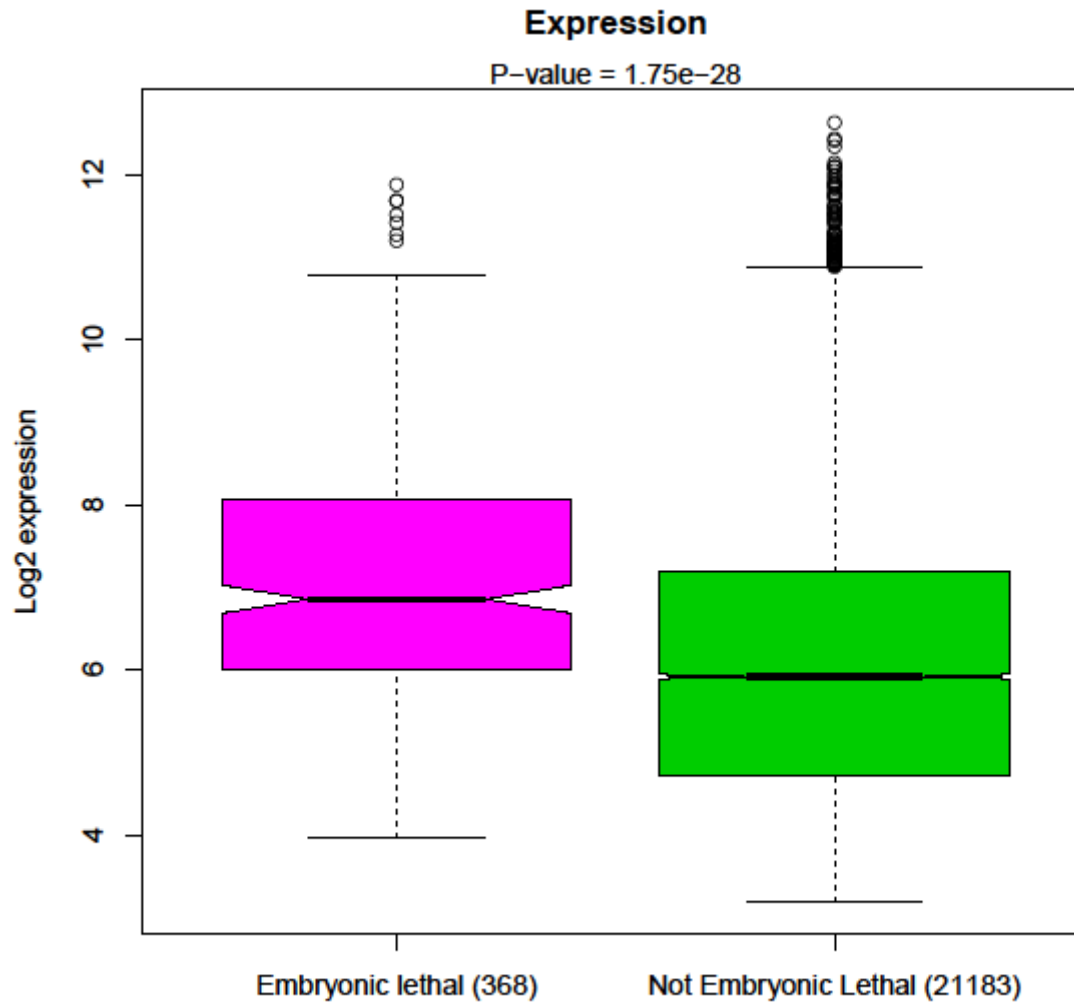


Figure 50 - comparison of tRMA normalized expression levels in a 5750 samples microarray dataset for embryonic lethal genes and other genes, as annotated by (Tzafrir et al., 2003), SeedGenes project v7.

5.8 Comparative Network Reconstruction Methods Analysis - Additional Network Quality Assessments

5.8.1 Example of a Network Degree distribution

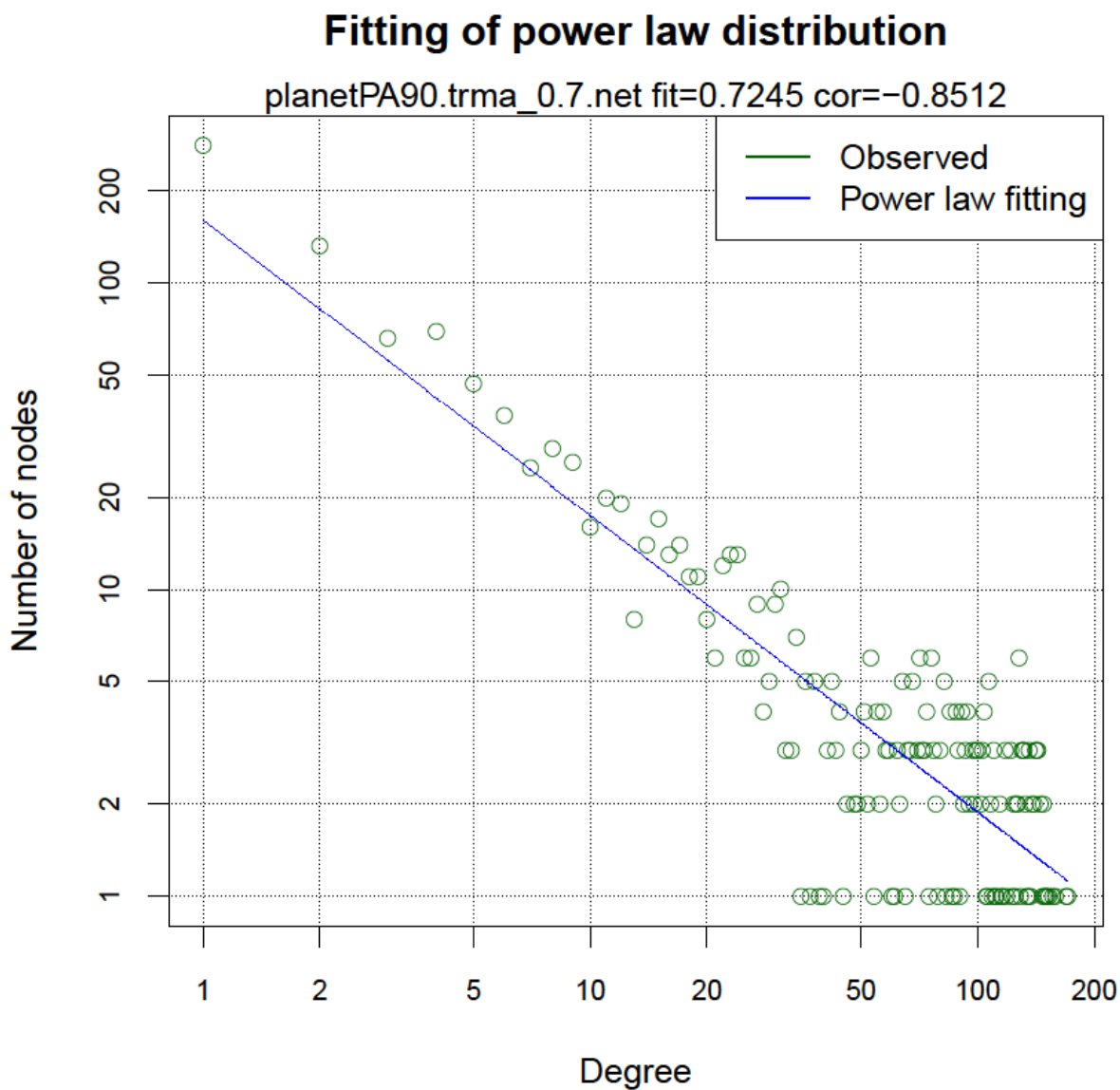


Figure 51 - Example of a Network Degree vs. Number of nodes distribution for an expression based reverse engineered Pearson Correlation network, r_0 threshold = 0.7, inferred from the dataset described in Paragraph 4.5 (274 microarray samples, 3350 genes). The blue line represents a linear model fit to the observed distribution ($R^2 = 0.7245$, Distribution PCC = -0.8512).

5.8.2 Fit to a power law of the Network Degree - adjusted R2

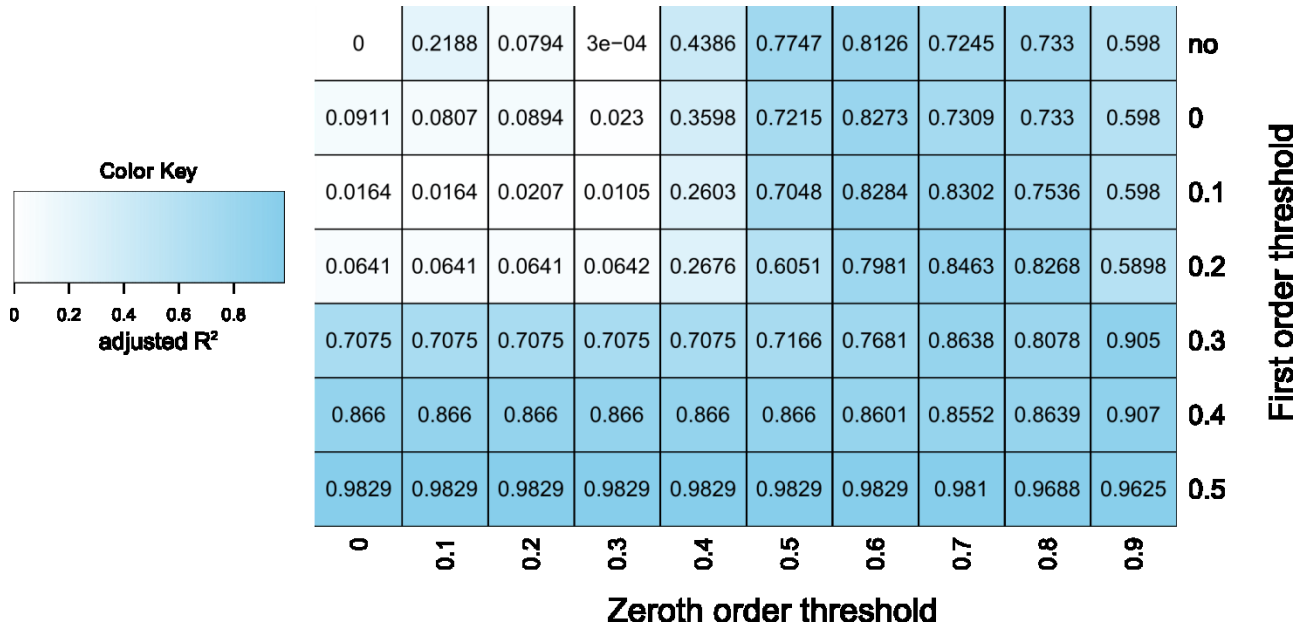


Figure 52 - Fit to a power law distribution of the Network Degree for *Arabidopsis thaliana* expression based Correlation and Partial Correlation networks at different thresholds. Absolute correlation coefficients were considered. A first order threshold of 0 means that edges suffering a sign change were excluded from the resulting networks, while a first order threshold marked with "no" corresponds to the standard zeroth order Pearson Correlation network. The R^2 is calculated based on the procedure described in (Brohée et al., 2008)

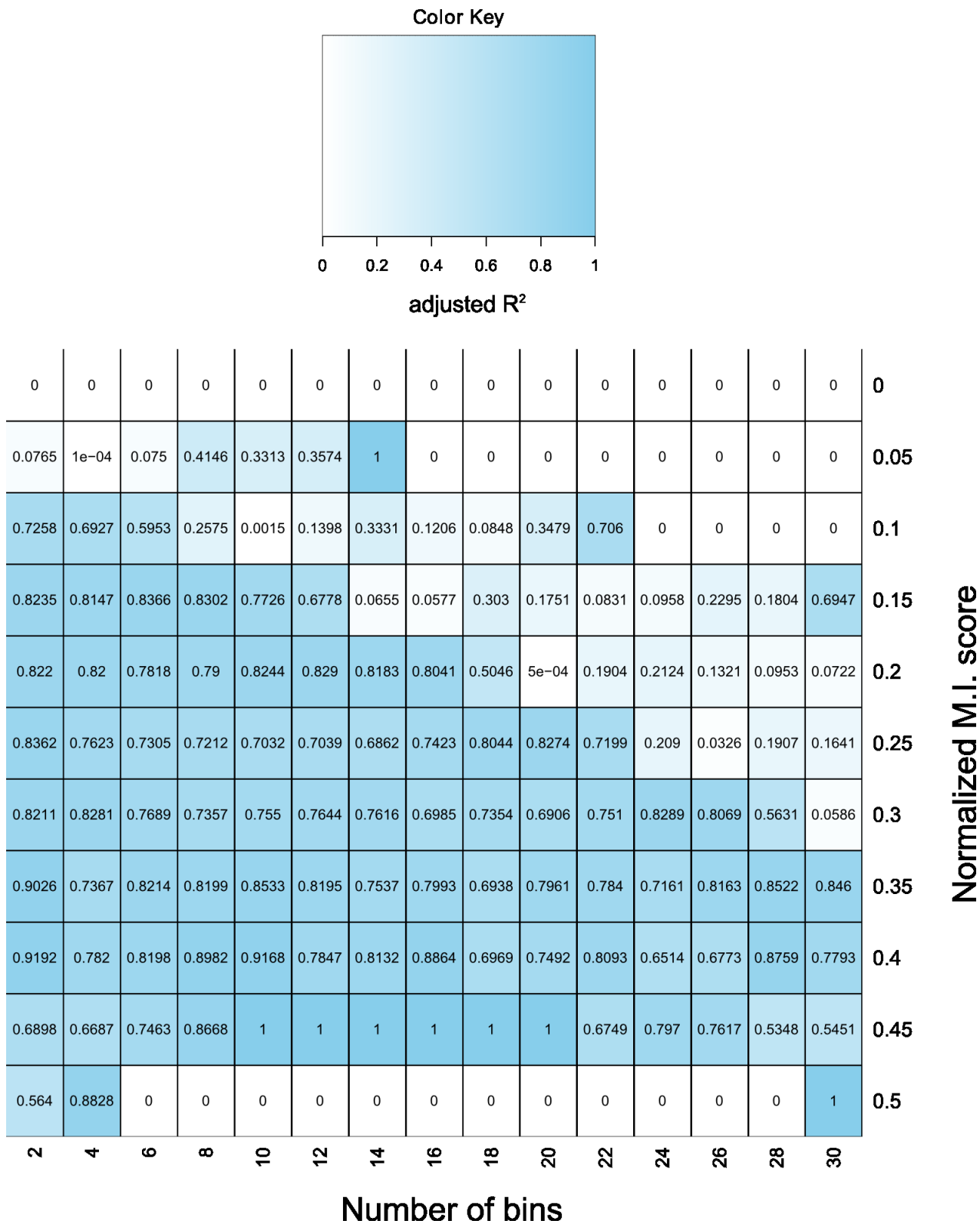


Figure 53 - Fit to a power law distribution of the Network Degree for *Arabidopsis thaliana* expression based Mutual Information networks at different combinations of significance thresholds and bin numbers. The R² is calculated based on the procedure described in (Brohée et al., 2008)

5.8.3 Fit to a power law of the Network Degree - Pearson Correlation coefficient of the Network Degree distribution

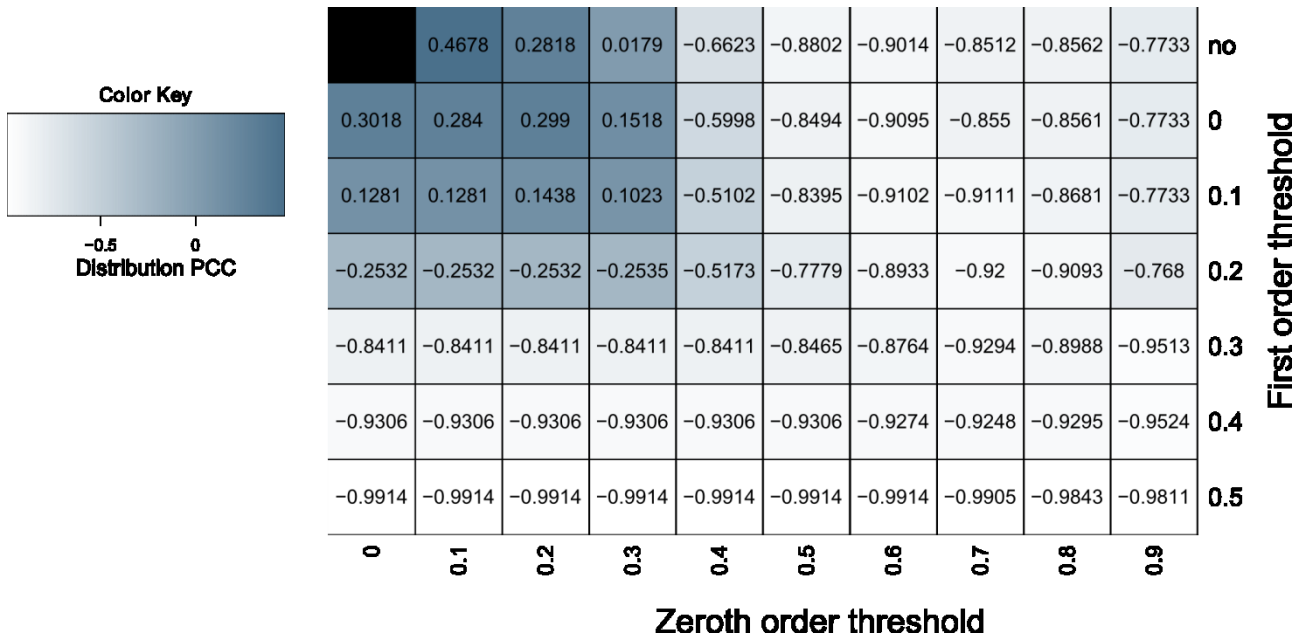


Figure 54 - Pearson's Correlation Coefficient (PCC) of the Network Degree vs. Node Number distribution for *Arabidopsis thaliana* expression based Correlation and Partial Correlation networks at different thresholds, calculated through the power law fitting procedure described in (Brohée et al., 2008)

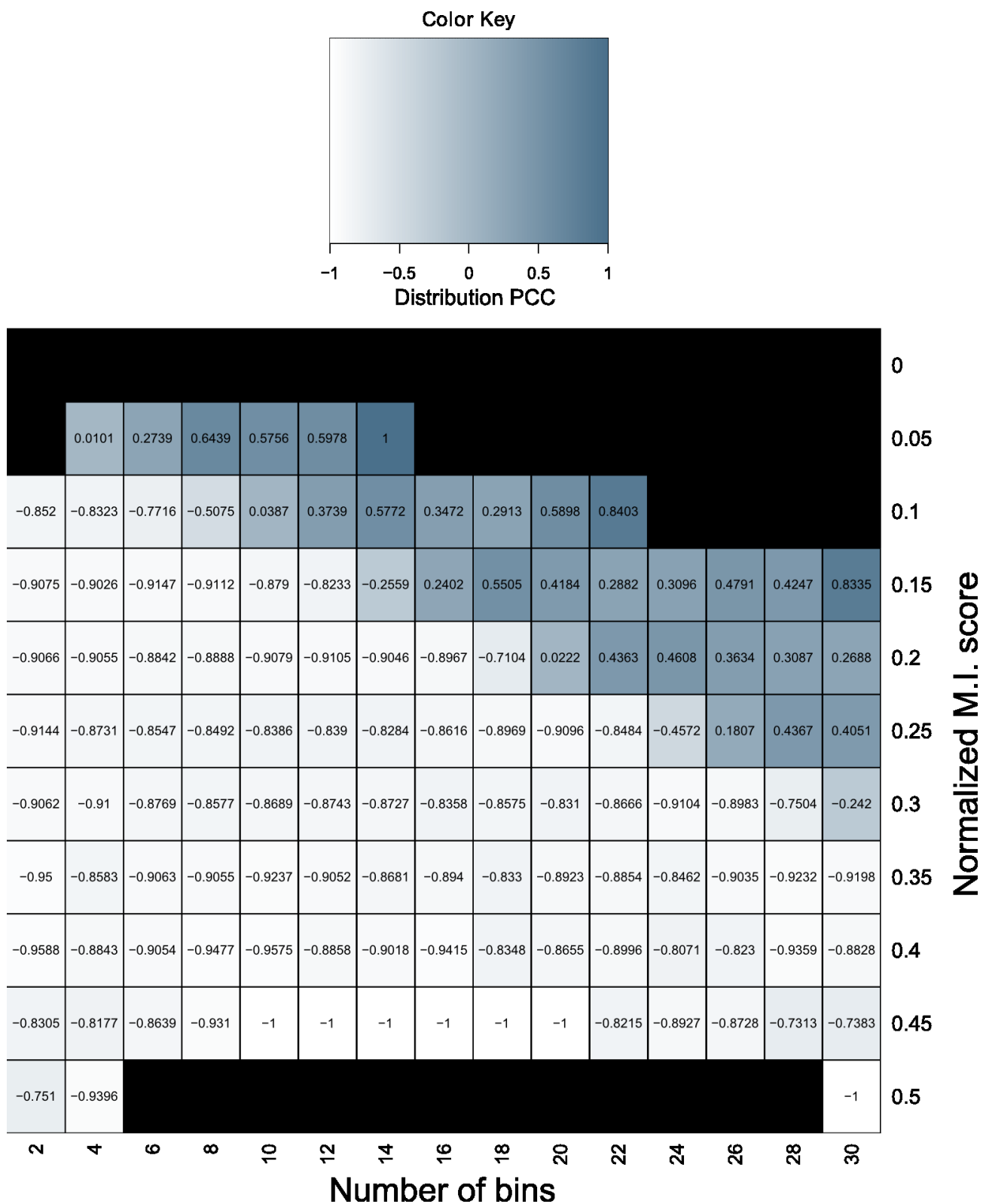


Figure 55 - Pearson's Correlation Coefficient (PCC) of the Network Degree vs. Node Number distribution for *Arabidopsis thaliana* expression based Mutual Information networks at combinations of bin number and M.I. index threshold, calculated through the power law fitting procedure described in (Brohée et al., 2008)

5.8.4 Overlap to Protein-Protein interaction networks - Accuracy

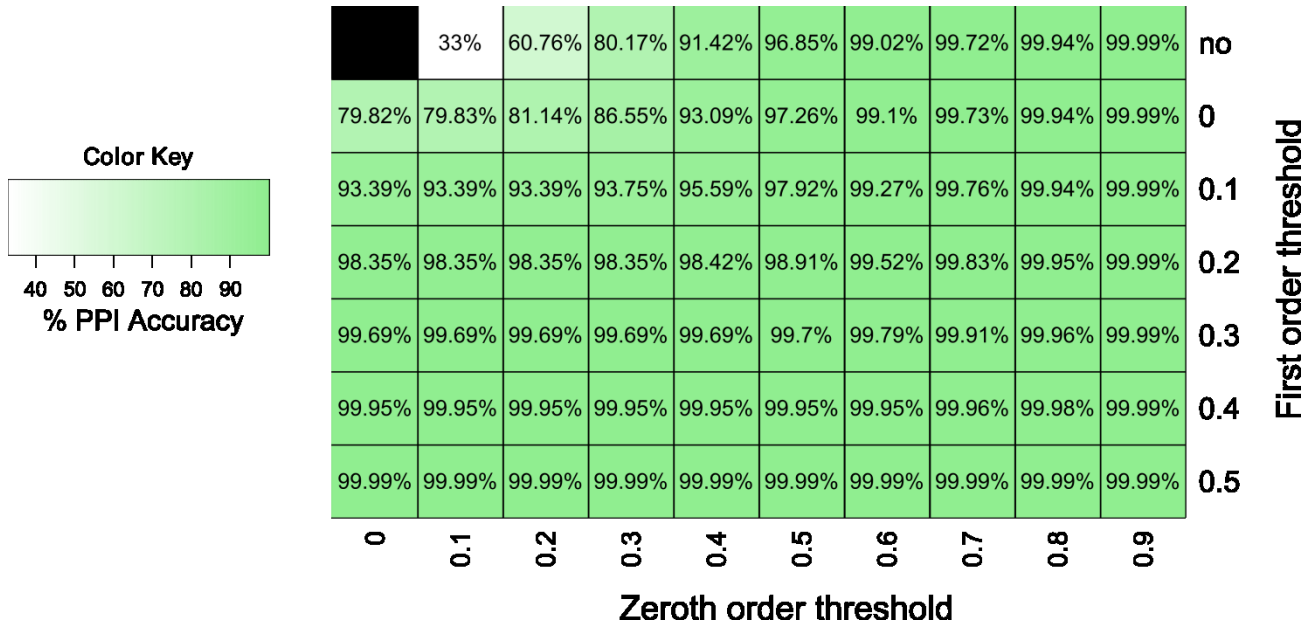


Figure 56 - Accuracy calculated using as golden set a manually curated *Arabidopsis thaliana* Protein-Protein Interaction network (Brandão et al., 2009) for expression-based Pearson Correlation networks at different zeroth and first order threshold combinations. Absolute correlation coefficients were considered. A first order threshold of 0 means that edges suffering a sign change were excluded from the resulting networks, while a first order threshold marked with "no" corresponds to the standard zeroth order Pearson Correlation network

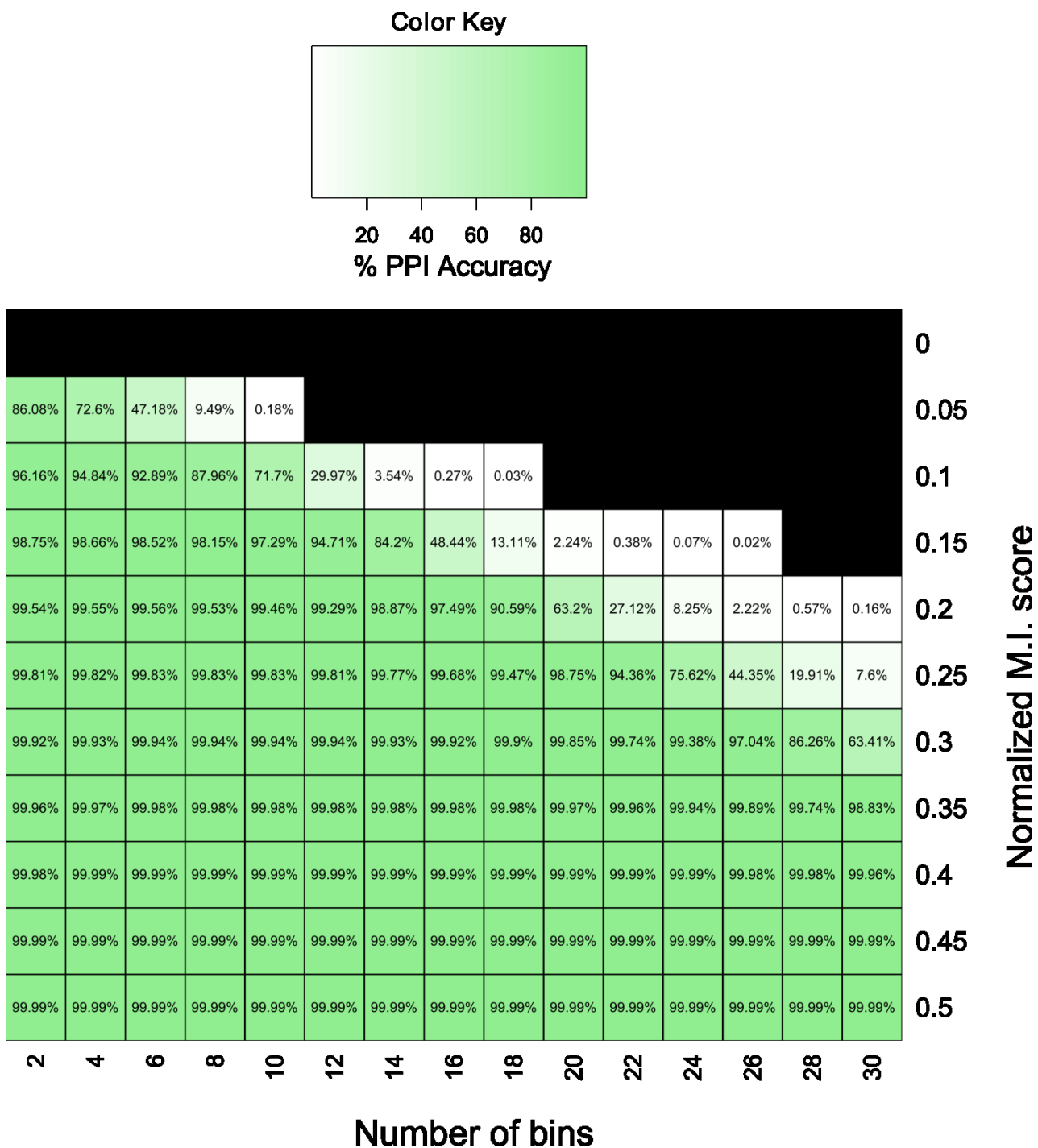


Figure 57 - Accuracy calculated using as golden set a manually curated *Arabidopsis thaliana* Protein-Protein Interaction network (Brandão et al., 2009) for expression-based Mutual Information networks at different bin numbers and normalize M.I. index thresholds

5.8.5 Overlap to Protein-Protein interaction networks - Matthew's coefficient

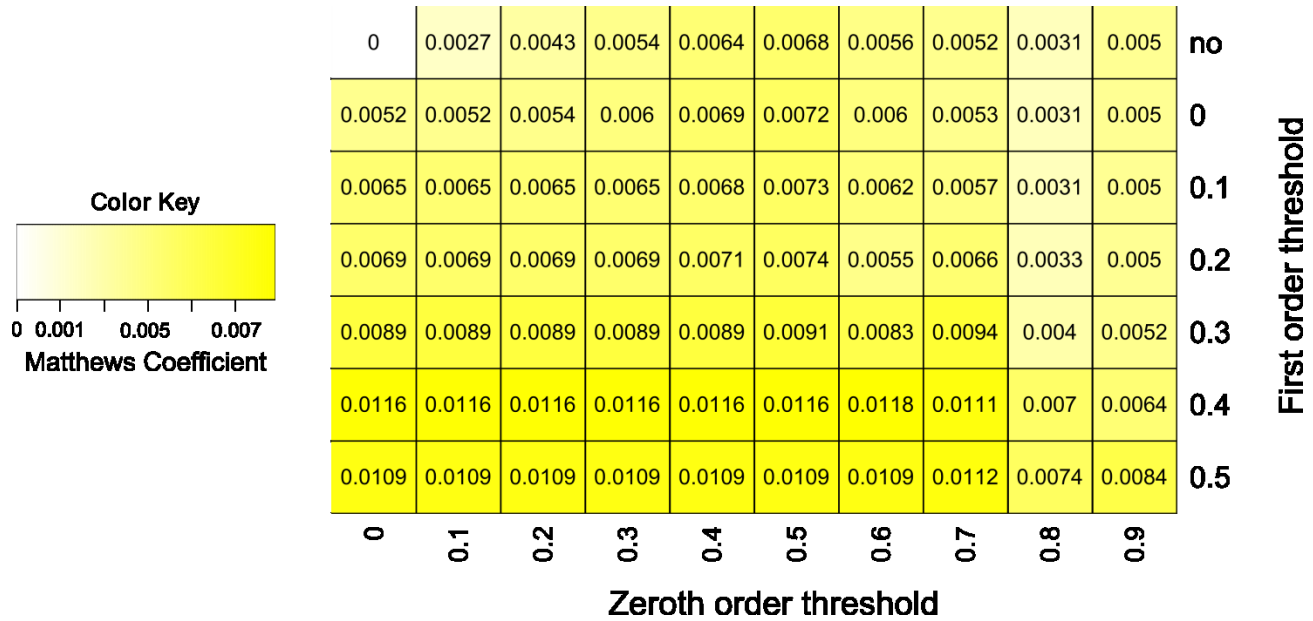


Figure 58 - Matthews Coefficient calculated using as golden set a manually curated *Arabidopsis thaliana* Protein-Protein Interaction network (Brandão et al., 2009) for expression-based Pearson Correlation networks at different zeroth and first order threshold combinations. Absolute correlation coefficients were considered. A first order threshold of 0 means that edges suffering a sign change were excluded from the resulting networks, while a first order threshold marked with "no" corresponds to the standard zeroth order Pearson Correlation network

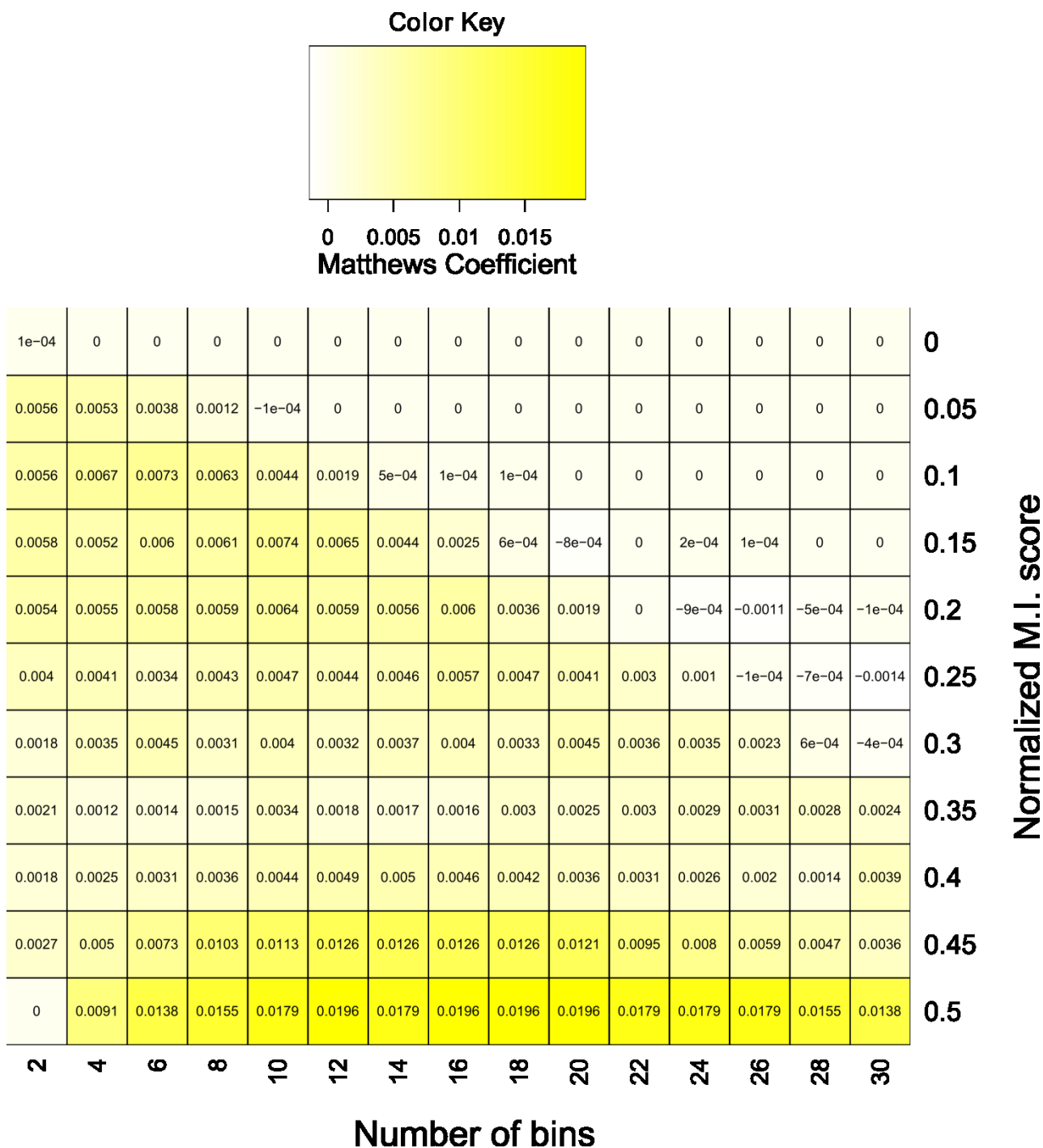


Figure 59 - Matthews Coefficient calculated using as golden set a manually curated *Arabidopsis thaliana* Protein-Protein Interaction network (Brandão et al., 2009) for expression-based Mutual Information networks at different bin numbers and normalize M.I. index thresholds

5.9 Coefficient Distributions for Pearson Correlation and Mutual Information

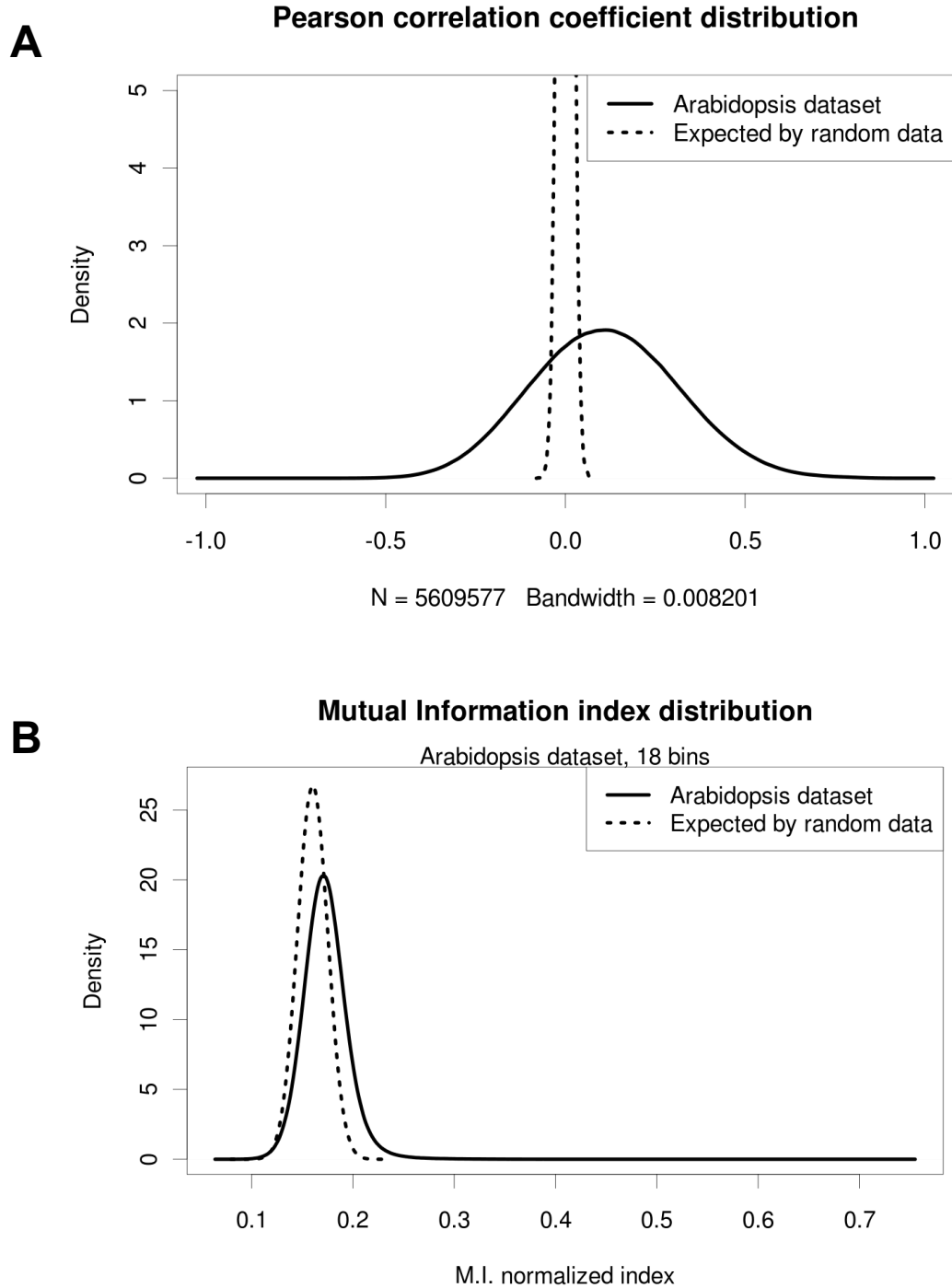


Figure 60 - (A) Pearson Correlation coefficient density distribution and (B) Mutual Information normalized index density distribution for the 274 samples *Arabidopsis thaliana* dataset described in Paragraph 4.5. Both distributions are compared with distributions expected from Gaussian random data.

5.10 LASSO model for the *Arabidopsis* gene RHM2

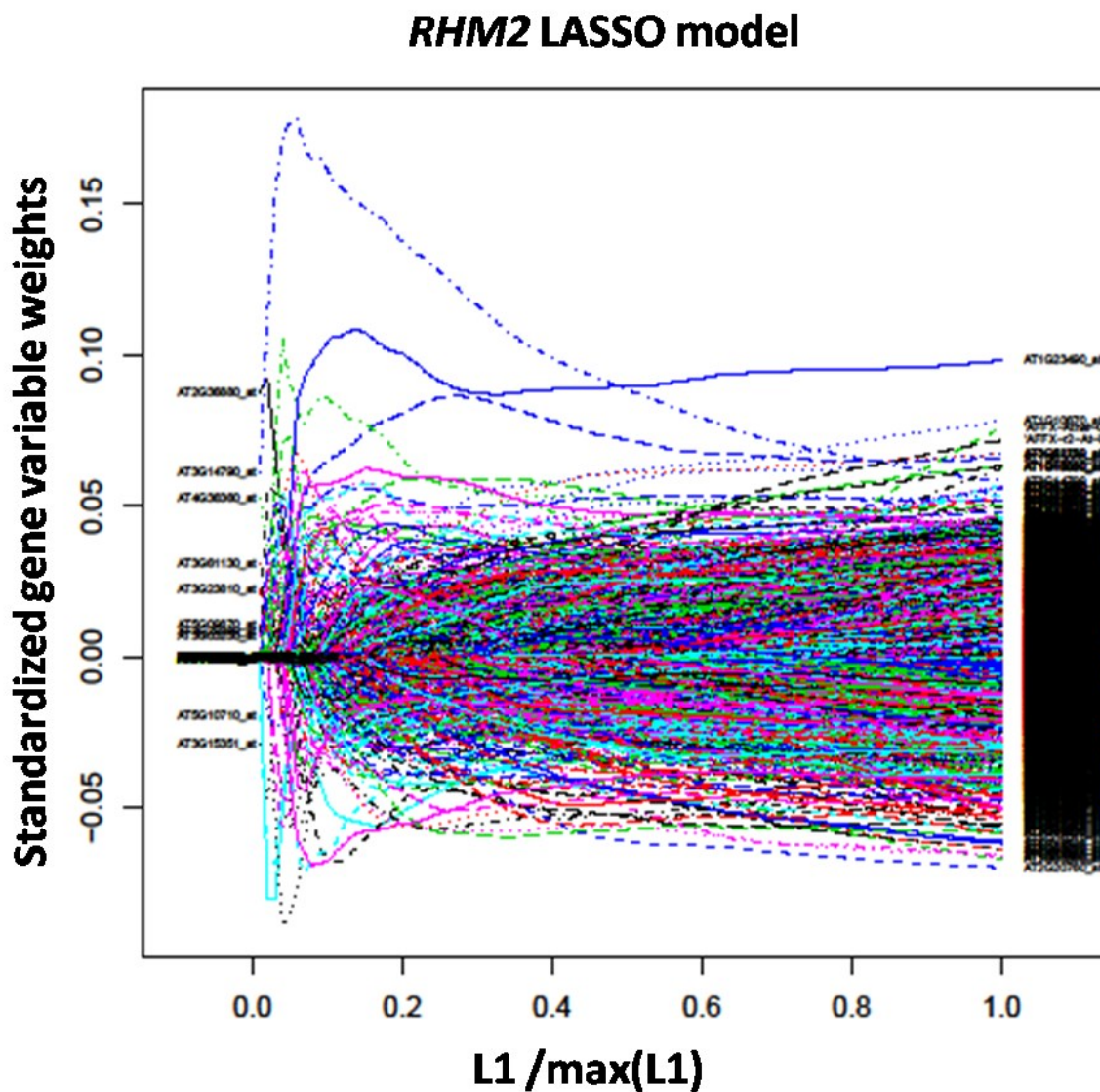


Figure 61 - LASSO model weight plot for the *RHM2* models based on the complete *Arabidopsis thaliana* collection of Affymetrix ATH1 public microarrays, described in Paragraph 2.5.1. On the x axis, varying sum of variable weight constraints, on the y axis, the weights for every gene predictor. Every line corresponds to a gene included as prediction variable for *RHM2*. The picture is shown as a representation of the complex nature of LASSO model generation (which proceeds from left to right on the x axis, increasing continuously the L1 sum of absolute weights

5.11 Mucilage release upon mechanical stress in a *Myb5* knockout line

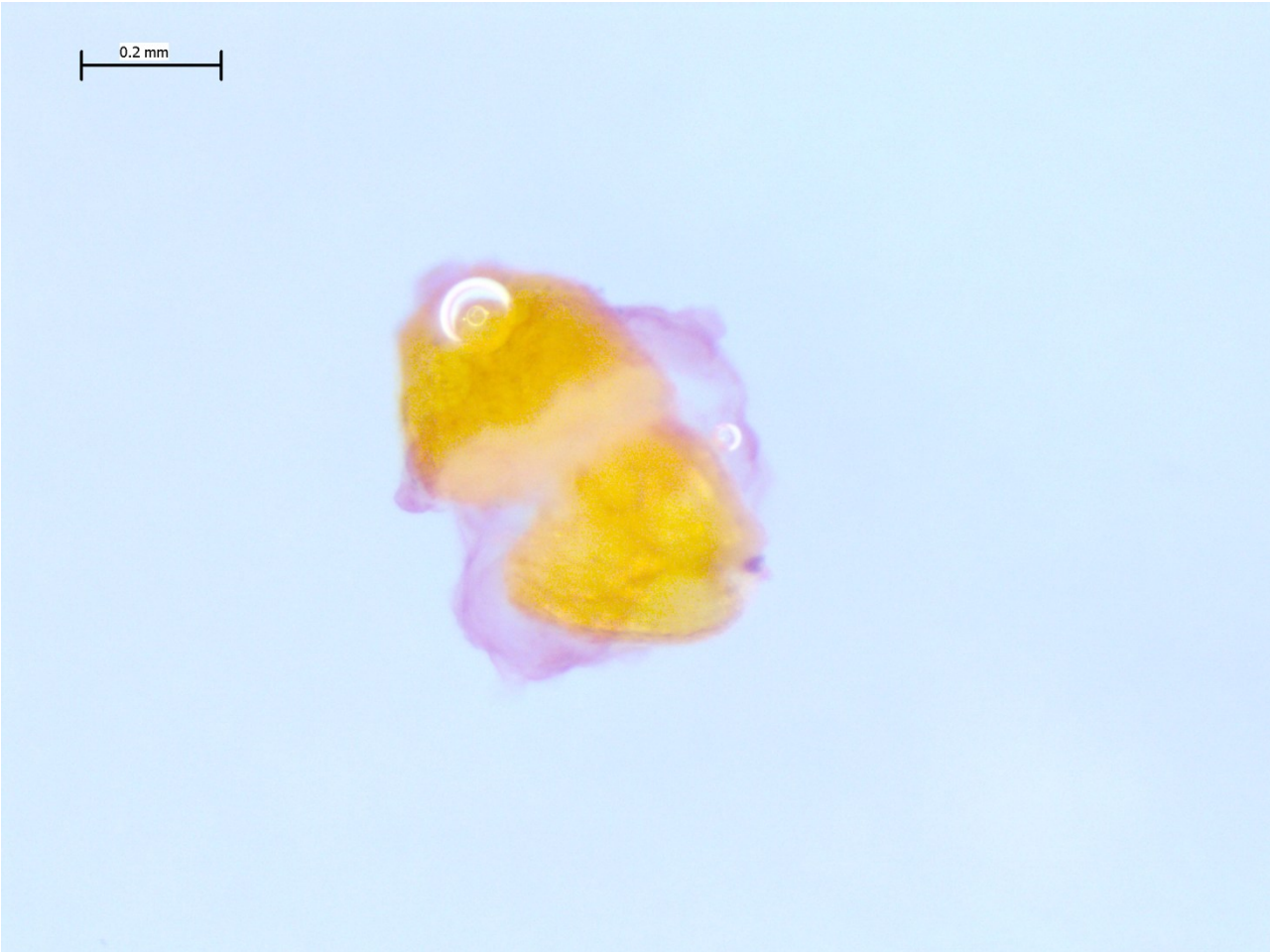


Figure 62 - Ruthenium Red staining of a *Myb5* knockout line seed after mechanical stress

5.12 *Thellungiella salsuginea* seeds upon hydration

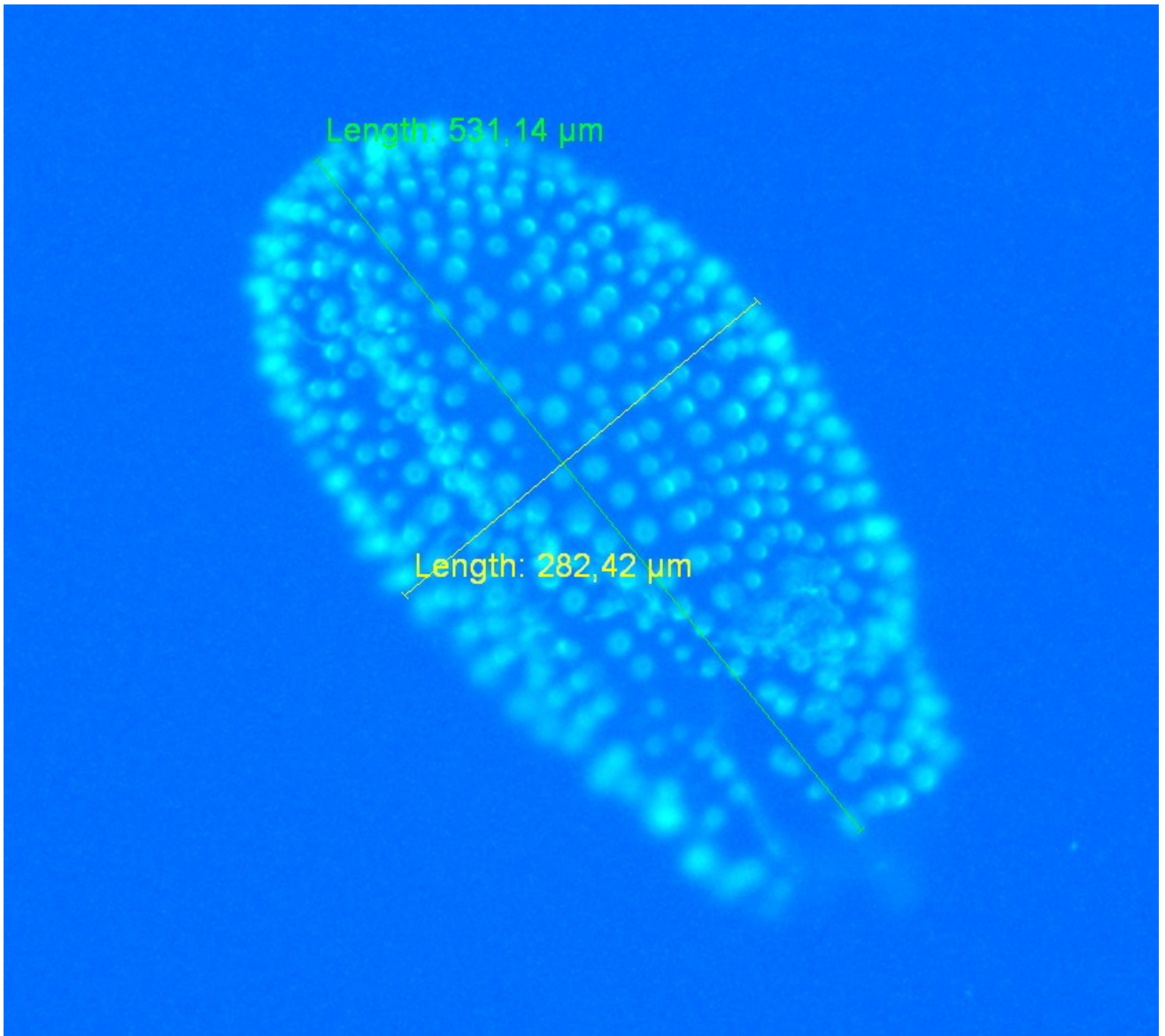


Figure 63 - *Thellungiella salsuginea* seed stained with calcofluor to highlight cell wall and pectin components. Although the seed is hydrated, no mucilage release is visible (Herth and Schnepf, 1980)

Publications

In brackets, personal contributions. Updated 28th July 2011.

- Rambaldi D, **Giorgi FM**, Capuani F, Ciliberto A, Ciccarelli FD. “*Low duplicability and network fragility of cancer genes*”, Trends Genet. 2008. (Analysis of the local network properties of specific cancer genes and development of a web interface for data mining).
- Civril F, Wehenkel A, **Giorgi FM**, Santaguida S, Di Fonzo A, Grigorean G, Ciccarelli FD, Musacchio A. “*Structural analysis of the RZZ complex reveals common ancestry with multisubunit vesicle tethering machinery*”, Structure 2010. (Remote homology analysis, protein domain characterization and detailed structure folding prediction).
- Usadel B, Obayashi T, Mutwil M, **Giorgi FM**, Bassel GW, Tanimoto M, Chow A, Steinhauser D, Persson S, Provart NJ. “*Co-expression tools for plant biology: opportunities for hypothesis generation and caveats*”, Plant Cell Environ. 2009. (Study on the influence of data normalization procedures on gene coexpression networks).
- Mutwil M, Ruprecht C, **Giorgi FM**, Bringmann M, Usadel B, Persson S. “*Transcriptional wiring of cell wall-related genes in Arabidopsis*”, Mol Plant. 2009. (Protein family association and statistical analysis).
- Lohse M, Nunes-Nesi A, Krüger P, Nagel A, Hannemann J, **Giorgi FM**, Childs L, Osorio S, Walther D, Selbig J, Sreenivasulu N, Stitt M, Fernie AR, Usadel B. “*Robin: an intuitive wizard application for R-based expression microarray quality assessment and analysis*”, Plant Physiol. 2010. (Microarray datasets quality controls and program debugging).
- **Giorgi FM**, Bolger AM, Lohse M and Usadel B. “*Algorithm-driven Artifacts in median polish summarization of Microarray data*”, BMC Bioinformatics 2010. (Conceived the study and carried out the statistical experiments).
- Licausi F, **Giorgi FM**, Zenoni S, Osti F, Pezzotti M and Perata P. “*Genomic and transcriptomic analysis of the AP2/ERF superfamily in Vitis vinifera*”, BMC Genomics 2010. (Cross-species gene prediction through domain scan of the grapevine genome, statistical analysis of expression).
- Mutwil M, Klie S, Tohge T, **Giorgi FM**, Wilkins O, Campbell MM, Fernie AR, Usadel B, Nikoloski Z and Persson S. “*PlaNet: Combined Sequence and Expression Comparisons across Plant Networks Derived from Seven Species*”, The Plant Cell 2011. (Analysis of relationships between protein domains and expression across species).
- **Giorgi FM**, Licausi F, Schmaelzlin E, Usadel B, Perata P, van Dongen J and Geigenberger P. “*HRE-Type genes are Regulated by growth-related changes in Internal oxygen concentrations during the normal development of potato (Solanum tuberosum) Tubers*”, Plant & Cell Physiology, provisionally accepted. (Conceived the coexpression analysis and carried out transcript expression measurements).

Posters

- Giorgi FM et al., “*A novel centrality framework for causal gene regulatory network reverse engineering*”, International Conference of Arabidopsis Research, Edinburgh, 2009.
- Giorgi FM et al., “*Using a model of the transcript response of Arabidopsis thaliana as a predictor for important factors during day-night cycles*”, International Conference of Arabidopsis Research, Edinburgh, 2009.
- Giorgi FM et al., “*Algorithm-driven Artifacts in median polish summarization of Microarray data*”, European Conference of Computational Biology, Ghent, 2010.

Curriculum vitae

PERSONAL INFORMATION

Born in Bologna, November 30th, 1982

Address: Zillestrasse 42, Berlin, Germany

Telephone: +39 339 4832 494

E mail: federico.giorgi@gmail.com

EDUCATION

| | |
|-------------------------------------|--|
| August 2008 - July 2011 | PhD student at Max Planck Institute of Molecular Plant Physiology, Golm, Integrative Carbon Biology group, supervisor Prof. Dr. Björn Usadel. |
| 27th March 2008 | Master's Degree in Bioinformatics, course of Pharmaceutical Biotechnology, Università di Bologna. Thesis title: " <i>Construction of a Database to analyze network features of cancer-related genes</i> ", supervisor Prof. Dr. Francesca Ciccarelli. Final grade 110/110 cum laude. |
| March 2006 - June 2006 | Trainee at Dipartimento di Fisiologia Umana e Generale, University of Bologna, supervisor Prof. Dr. Giorgio Aicardi |
| 26th October 2004 | Bachelor's Degree in Eukaryotic Molecular Biology, course of Biotechnology, Università di Bologna. Thesis title: " <i>Analysis of Erb-B1 gene mutations in relation to sensitivity to Iressa in human tumoral cells</i> ", supervisor Prof. Dr. Giovanni Capranico, cosupervisor Prof. Dr. Giuseppe Giaccone. Final grade 110/110 cum laude. |
| April 2004 - August 2004 | Trainee at Department of Oncology, Vrije Universiteit of Amsterdam, supervisor Prof. Dr. Giuseppe Giaccone |
| 1st July 2001 | Scientific High School degree, Liceo "Enrico Fermi", Bologna. Final grade 100/100 cum laude. |

LIST OF TECHNICAL SKILLS

Molecular biology. Extraction and purification of DNA, DNA amplification (PCR), DNA Sanger sequencing, western blotting, plant genotyping, ChIP, RT-PCR etc. **Cell biology.** Primary cell cultures (astrocytes), DNA transfection of cell lines, protein extraction from cell lines. **Electrophysiology.** Field potential in hippocampal and cortical brain slices. **Bioinformatics.** Fluency in programming languages: Perl, R, Java. Web and Database development (CGI-Perl, GWT, MySQL). Statistics. Sequence analysis (phylogenesis, structural bioinformatics, remote homology detection). Differential gene expression and coexpression analysis. Next-gen Transcriptome and Genome assembly (mira, CLC-gw, ABySS)

Selbständigkeitserklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbständig und unter Verwendung keiner anderen als den von mir angegebenen Quellen und Hilfsmitteln verfasst habe. Teile des Inhalts in den Absätzen 1.2, 2.2 und 4.2 sind in der wissenschaftlichen Publikation "*Algorithm driven Artifacts in median polish summarization of Microarray data*" von Giorgi FM et al. in BMC Bioinformatics erschienen. Teile der Absätze 1.4.2, 2.6 und 4.8 sind in der wissenschaftlichen Publikation "*HRE-type Genes are regulated by Growth-Related Changes in Internal Oxygen Concentrations during the normal development of Potato (*Solanum tuberosum*) tubers*" von Licausi F, Giorgi FM et al. mit geteilter Erstautorenschaft in Plant Cell Physiology erschienen. Hiermit erkläre ich, dass ich in beiden Fällen der alleinige Autor der entsprechenden Teile bin.

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

Potsdam, den 31.07.2011

Federico Manuel Giorgi

Acknowledgments

This thesis is dedicated to my son, **Lupo**, which was born 1 year and a half ago, and therefore has accompanied me through half of the work presented here. Whatever kind of man he will become, scientist or not, I hope he will be proud of his old man and his German adventure. Next, I would like to thank my current group leader **Björn Usadel**, who gave me the opportunity to work on the extremely fascinating and complex field of Bioinformatics, and especially for his strong multi-disciplinary attitude, which is tremendously rare in the scientific world, filled with ivory towers and overspecialization. And my former group leader **Francesca Ciccarelli**, who taught me how a researcher should reason and approach any scientific problem.

I would also like to thank the following people for their various degrees of contribution to this PhD thesis.

Francesco Licausi, for his endless motivation and for showing me how from simple ideas great discoveries can be obtained. Despite the fact that he's 20 days younger than me, I've always regarded him as a wise example to follow for his almost incredible intrinsic scientific instinct. **Marc Lohse** and **Anthony Bolger**, who taught me a lot from a technical and a human point of view. They spent way too much time in trying to inculcate programming principles in my biologist mind, and if I've finally succeeded (or better, I've not totally failed) I owe this to them. **Aleksandar Vasilevski** for being the best example of hard-working and dedicated PhD student I've met so far. He was the first one who believed in an application of the LASSO to real-life problems and has been a constantly motivating presence in the group. **Daan Weits**, an extremely promising young scientist in the institute, for showing me how it is possible to have exceptionally lucid scientific thought without renouncing to the healthy aspects of life. **Zoran Nikoloski**, for believing in the goodness of my ideas even in times when all seemed lost, and for keeping an ever-positive attitude. **Heike Riegler**, postdoc in my group: we have largely different working fields, working hours and almost opposite characters; however, many times I've felt her as the closest person in the institute, I felt to understand her wholly and to be wholly understood. **Malgorzata Ryngajlo**, for the chats, for the science, and because she's the only one to which I show my true self. My Italian friends, who helped as beta-testers for the CorTo tool in its early and buggy days: my wife **Lara Giuffrida**, Agostino Carandente, Salvatore Insalaco, Raul Picciotti, Francesco Sicurella, Alessio Cantore, Giuseppe Pennoni, Chiara Bacci, Alessia Milani, Giacomo Lanzoni, Benjamin Jaegle, Tania Molteni, Eoin, Serenella, Valentina Panebianco, Davide Rambaldi, Anna de Grassi, Fabio Pes, Micaela Radivo and Silvia Gulisano for kindly offering themselves as beta-testers of the CorTo tool in its early and buggy days. **Luca Molteni** for always being my potential best friend, potential because we unfortunately always live at thousands of kms from each other. **Mimi Lyda Brown**, who survived my leadership and followed my directions when I still had doubts about the directions to follow myself. **Oliver Drechsel**, an extremely bright scientist who unfortunately I had the luck to meet only in the final months of my PhD. **Colin Ruprecht**, for helping me in stressful moments, both in the lab and in the outside world. **Vittoria Offeddu**, for the too few but great and inspiring evenings spent chatting. **Marek Mutwil**, for teaching me how to play the bass. **Kenny Billiau**, **Alex Ivakov** and **Anna Flis**, great minds both for everyday and scientific subjects. **Eleonora Paparelli**, for the help in the lab and for answering to all my out-of-the-blue questions about *Arabidopsis* enzymatic activity. **Antonietta Santaniello**, for being able to merge comic books and biological experiments. **Beatrice Giuntoli**, for her great help in graphical matters. My mother **Katia**, my father **Cito**, my grandparents **Anna**, **Domenico**, **Frida** and **Bepino**. The authors who turned on my brain every morning **Randall Munroe**, **Jeph Jacques** and **Ryan North**. **Andrea Califano** for writing great papers on gene networks who inspired my early and not-so-early steps in Bioinformatics. **Marie Hopkins**, for being always very kind to me despite my mood shifts, and the other members of the group: Ewelina, Ewelina, Mareike, Kati, Anja, Florian, Diana and Thomas. **Joost van Dongen**, for showing me how a researcher should formulate scientific questions. **Lilian Matallana** and **Maria Cristina**, for their empathy and their Spanish lessons. Prof. **Dittmann** and Prof. **Gaedke** for their tremendous help during the final moments of the PhD. All these people, and many more which I have now forgotten to add, made these three years in Germany an extremely instructive period of my life. I will always look back at my days in Golm with a nostalgic smile, and this will all be thanks to you.

Bibliography

Affymetrix <http://www.affymetrix.com/>

- Agarwal A, Koppstein D, Rozowsky J, Sboner A, Habegger L, Hillier LDW, Sasidharan R, Reinke V, Waterston RH, Gerstein M** (2010) Comparison and calibration of transcriptome data from RNA-Seq and tiling arrays. *BMC genomics* **11**: 1-16
- Akaike H** (2002) A new look at the statistical model identification. *Automatic Control, IEEE Transactions on* **19**: 716-723
- Alonso JM, Stepanova AN, Lisse TJ, Kim CJ, Chen H, Shinn P, Stevenson DK, Zimmerman J, Barajas P, Cheuk R** (2003) Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. *Science* **301**: 653
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ** (1990) Basic local alignment search tool. *J Mol Biol* **215**: 403-410
- Alwine JC, Kemp DJ, Stark GR** (1977) Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proceedings of the National Academy of Sciences* **74**: 5350
- Aoki K, Ogata Y, Shibata D** (2007) Approaches for extracting practical information from gene co-expression networks in plant biology. *Plant Cell Physiol* **48**: 381-390
- Appeldoorn NJG, de Bruijn SM, Koot-Gronsveld EAM, Visser RGF, Vreugdenhil D, van der Plas LHW** (1997) Developmental changes of enzymes involved in conversion of sucrose to hexose-phosphate during early tuberisation of potato. *Planta* **202**: 220-226
- Arsovski AA, Haughn GW, Western TL** (2010) Seed coat mucilage cells of *Arabidopsis thaliana* as a model for plant cell wall research. *Plant Signal Behav* **5**: 796-801
- Arsovski AA, Villota MM, Rowland O, Subramaniam R, Western TL** (2009) MUM ENHANCERS are important for seed coat mucilage production and mucilage secretory cell differentiation in *Arabidopsis thaliana*. *Journal of experimental botany*
- Arvidsson S, Kwasniewski M, Riaño-Pachón DM, Mueller-Roeber B** (2008) QuantPrime – a flexible tool for reliable high-throughput primer design for quantitative PCR. *BMC bioinformatics* **9**: 465
- Bai X, Mamidala P, Rajarapu SP, Jones SC, Mittapalli O** (2011) Transcriptomics of the bed bug (*Cimex lectularius*). *PLoS One* **6**: e16336
- Baldi P, Brunak S, Chauvin Y, Andersen CAF, Nielsen H** (2000) Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* **16**: 412
- Banks NH** (1983) Evaluation of methods for determining internal gases in banana fruit. *Journal of experimental botany* **34**: 871
- Barabási AL, Albert R** (1999) Emergence of scaling in random networks. *Science* **286**: 509
- Barakat A, Szick-Miranda K, Chang IF, Guyot R, Blanc G, Cooke R, Delseny M, Bailey-Serres J** (2001) The organization of cytoplasmic ribosomal protein genes in the *Arabidopsis* genome. *Plant physiology* **127**: 398
- Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A** (2005) Reverse engineering of regulatory networks in human B cells. *Nat Genet* **37**: 382-390
- Bauer S, Vasu P, Persson S, Mort AJ, Somerville CR** (2006) Development and application of a suite of polysaccharide-degrading enzymes for analyzing plant cell walls. *Proceedings of the National Academy of Sciences* **103**: 11417
- Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL** (2008) GenBank. *Nucleic acids research* **36**: D25
- Biais B, Beauvoit B, William Allwood J, Deborde C, Maucourt M, Goodacre R, Rolin D, Moing A** (2009) Metabolic acclimation to hypoxia revealed by metabolite gradients in melon fruit. *J Plant Physiol* **167**: 242-245
- Boeswinkel FD, Bouman F** (1995) The seed: structure and function. *Seed development and germination*: 1-24
- Bolstad B** (2008) 3 Preprocessing and Normalization for Affymetrix GeneChip Expression Microarrays. *Methods in microarray normalization*: 41
- Bolstad BM, Irizarry RA, Astrand M, Speed TP** (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**: 185-193
- Bor YCH, M.** (2006) Northern Blot analysis of mRNA from mammalian polyribosomes. *Nature Protocols Protocol Exchange*
- Borisjuk L, Rolletschek H** (2009) The oxygen status of the developing seed. *New Phytol* **182**: 17-30

- Boutillier K, Offringa R, Sharma VK, Kieft H, Ouellet T, Zhang L, Hattori J, Liu CM, van Lammeren AAM, Miki BLA** (2002) Ectopic expression of BABY BOOM triggers a conversion from vegetative to embryonic growth. *The Plant Cell Online* **14**: 1737
- Boutros PC, Okey AB** (2005) Unsupervised pattern recognition: an introduction to the whys and wherefores of clustering microarray data. *Brief Bioinform* **6**: 331-343
- Brandão MM, Dantas LL, Silva-Filho MC** (2009) AtPIN: Arabidopsis thaliana protein interaction network. *BMC bioinformatics* **10**: 454
- Breitkreutz BJ, Stark C, Reguly T, Boucher L, Breitkreutz A, Livstone M, Oughtred R, Lackner DH, Bahler J, Wood V** (2007) The BioGRID interaction database: 2008 update. *Nucleic acids research*
- Brohée S, Faust K, Lima-Mendez G, Vanderstocken G, van Helden J** (2008) Network Analysis Tools: from biological networks to clusters and pathways. *Nat Protoc* **3**: 1616-1629
- Brown DM, Zeef LAH, Ellis J, Goodacre R, Turner SR** (2005) Identification of novel genes in Arabidopsis involved in secondary cell wall formation using expression profiling and reverse genetics. *The Plant Cell Online* **17**: 2281
- Brown MPS, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M, Haussler D** (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences of the United States of America* **97**: 262
- Butte AJ, Kohane IS** (1999) Unsupervised knowledge discovery in medical databases using relevance networks. *Proc AMIA Symp*: 711-715
- Butte AJ, Kohane IS** (2000) Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac Symp Biocomput*: 418-429
- Caesele LV, Mills JT, Sumner M, Gillespie R** (1981) Cytology of mucilage production in the seed coat of Candle canola (*Brassica campestris*). *Canadian Journal of Botany* **59**: 292-300
- Campbell MA, Woodring A, Stidd J** (1998) Isolation of a cDNA from potato with structural similarity to the AP2 gene superfamily. *Plant Physiol* **117**: 1127
- Carlson MR, Zhang B, Fang Z, Mischel PS, Horvath S, Nelson SF** (2006) Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. *BMC Genomics* **7**: 40
- Carmeliet P, Dor Y, Herbert JM, Fukumura D, Brusselmans K, Dewerchin M, Neeman M, Bono F, Abramovitch R, Maxwells P** (1998) Role of HIF-1 in hypoxia-mediated apoptosis, cell proliferation and tumour angiogenesis. *NATURE-LONDON*: 485-490
- Caspi R, Foerster H, Fulcher CA, Kaipa P, Krummenacker M, Latendresse M, Paley S, Rhee SY, Shearer AG, Tissier C, Walk TC, Zhang P, Karp PD** (2008) The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res* **36**: D623-631
- Chalker AF, Lunsford RD** (2002) Rational identification of new antibacterial drug targets that are essential for viability using a genomics-based approach. *Pharmacol Ther* **95**: 1-20
- Cheung F, Haas BJ, Goldberg S, May GD, Xiao Y, Town CD** (2006) Sequencing *Medicago truncatula* expressed sequenced tags using 454 Life Sciences technology. *BMC genomics* **7**: 272
- Chevreux B** (2005) MIRA: An automated genome and EST assembler. Duisburg: Heidelberg
- Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Müller WEG, Wetter T, Suhai S** (2004) Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Research* **14**: 1147
- Christianson JA, Llewellyn DJ, Dennis ES, Wilson IW** (2010) Comparisons of early transcriptome responses to low-oxygen environments in three dicotyledonous plant species. *Plant signaling & behavior* **5**
- Chun KT, Goebel MG** (2004) Necessity's sharp pinch. *Mol Cell* **15**: 166-168
- Churchill GA** (2002) Fundamentals of experimental design for cDNA microarrays. *Nature genetics* **32**: 490-495
- Cole ST** (2002) Comparative mycobacterial genomics as a tool for drug target and antigen discovery. *Eur Respir J Suppl* **36**: 78s-86s
- Cole ST, Saint Girons I** (1994) Bacterial genomics. *FEMS Microbiol Rev* **14**: 139-160
- Copas JB** (1983) Regression, prediction and shrinkage. *Journal of the Royal Statistical Society. Series B (Methodological)* **45**: 311-354
- Cope LM, Irizarry RA, Jaffee HA, Wu Z, Speed TP** (2004) A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics* **20**: 323-331
- Cosgrove DJ** (2005) Growth of the plant cell wall. *Nature Reviews Molecular Cell Biology* **6**: 850-861

- Crick F** (1970) Central dogma of molecular biology. *Nature* **227**: 561-563
- Czechowski T, Bari RP, Stitt M, Scheible WR, Udvardi MK** (2004) Real time RT PCR profiling of over 1400 Arabidopsis transcription factors: unprecedented sensitivity reveals novel root and shoot specific genes. *The Plant Journal* **38**: 366-379
- D'haeseleer P, Liang S, Somogyi R** (2000) Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics* **16**: 707
- Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, Bunney WE, Myers RM, Speed TP, Akil H, Watson SJ, Meng F** (2005) Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res* **33**: e175
- Datta S** (2003) Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics* **19**: 459-466
- Daub CO, Steuer R, Selbig J, Kloska S** (2004) Estimating mutual information using B-spline functions – an improved similarity measure for analysing gene expression data. *BMC bioinformatics* **5**: 118
- de la Fuente A, Bing N, Hoeschele I, Mendes P** (2004) Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics* **20**: 3565-3574
- Dean GH, Zheng H, Tewari J, Huang J, Young DS, Hwang YT, Western TL, Carpita NC, McCann MC, Mansfield SD** (2007) The Arabidopsis MUM2 gene encodes a -galactosidase required for the production of seed coat mucilage with correct hydration properties. *The Plant Cell Online* **19**: 4007
- Dunn MJ** (2011) PROTEOMICS: Into the next decade. *Proteomics* **11**: 1-3
- Durand GM, Zukin RS** (1993) Developmental regulation of mRNAs encoding rat brain kainate/AMPA receptors: a northern analysis study. *J Neurochem* **61**: 2239-2246
- Edgar R, Domrachev M, Lash AE** (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* **30**: 207-210
- Edgar RC** (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* **32**: 1792
- Efron B, Hastie T, Johnstone I, Tibshirani R** (2004) Least angle regression. *Annals of statistics* **32**: 407-451
- Efron B, Tibshirani R** (1995) Cross-validation and the bootstrap: Estimating the error rate of a prediction rule. Division of Biostatistics, Stanford University
- Ehling J, Provart NJ, Werck-Reichhart D** (2006) Functional annotation of the Arabidopsis P450 superfamily based on large-scale co-expression analysis. *Biochemical Society Transactions* **34**: 1192-1198
- Eisen MB, Spellman PT, Brown PO, Botstein D** (1998) Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences* **95**: 14863
- Eisen MB, Spellman PT, Brown PO, Botstein D** (1998) Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* **95**: 14863-14868
- Eklund AC, Szallasi Z** (2008) Correction of technical bias in clinical microarray data improves concordance with known biological information. *Genome Biol* **9**: R26
- Erdős P, Rényi A** (1959) On random graphs, I. *Publicationes Mathematicae (Debrecen)* **6**: 290-297
- Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS** (2007) Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol* **5**: e8
- Fernie AR, Tiessen A, Stitt M, Willmitzer L, Geigenberger P** (2002) Altered metabolic fluxes result from shifts in metabolite levels in sucrose phosphorylase expressing potato tubers. *Plant, Cell & Environment* **25**: 1219-1232
- Ferreira SJ, Senning M, Sonnewald S, Keßling PM, Goldstein R, Sonnewald U** (2010) Comparative transcriptome analysis coupled to X-ray CT reveals sucrose supply and growth velocity as major determinants of potato tuber starch biosynthesis. *BMC genomics* **11**: 93
- Fitzsimmons D, Hagman J** (1996) Regulation of gene expression at early stages of B-cell and T-cell differentiation. *Curr Opin Immunol* **8**: 166-174
- Forristal CE, Wright KL, Hanley NA, Oreffo ROC, Houghton FD** (2010) Hypoxia inducible factors regulate pluripotency and proliferation in human embryonic stem cells cultured at reduced oxygen tensions. *Reproduction* **139**: 85
- Fransson P, Marrelec G** (2008) The precuneus/posterior cingulate cortex plays a pivotal role in the default mode network: Evidence from a partial correlation network analysis. *Neuroimage* **42**: 1178-1184
- Frenzel S, Pompe B** (2007) Partial mutual information for coupling analysis of multivariate time series. *Physical review letters* **99**: 204101
- Friedman N** (2004) Inferring cellular networks using probabilistic graphical models. *Science* **303**: 799-805

- Friedman N, Linial M, Nachman I, Pe'er D** (2000) Using Bayesian networks to analyze expression data. *Journal of computational biology* **7**: 601-620
- Fukao T, Bailey-Serres J** (2004) Plant responses to hypoxia-is survival a balancing act? *Trends in Plant Science* **9**: 449-456
- Fukao T, Bailey-Serres J** (2008) Submergence tolerance conferred by Sub1A is mediated by SLR1 and SLRL1 restriction of gibberellin responses in rice. *Proceedings of the National Academy of Sciences* **105**: 16814
- Fukao T, Xu K, Ronald PC, Bailey-Serres J** (2006) A variable cluster of ethylene response factor-like genes regulates metabolic and developmental acclimation responses to submergence in rice. *The Plant Cell Online* **18**: 2021
- Gama-Castro S, Jiménez-Jacinto V, Peralta-Gil M, Santos-Zavaleta A, Peñaloza-Spinola MI, Contreras-Moreira B, Segura-Salazar J, Muñoz-Rascado L, Martínez-Flores I, Salgado H** (2008) RegulonDB (version 6.0): gene regulation model of *Escherichia coli* K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic acids research* **36**: D120
- Gama-Castro S, Salgado H, Peralta-Gil M, Santos-Zavaleta A, Muñoz-Rascado L, Solano-Lira H, Jimenez-Jacinto V, Weiss V, García-Sotelo JS, López-Fuentes A** (2011) RegulonDB version 7.0: transcriptional regulation of *Escherichia coli* K-12 integrated within genetic sensory response units (Gensor Units). *Nucleic acids research* **39**: D98
- Gasieniec L, Li CY, Sant P, Wong PWH** (2006) Efficient Probe Selection in Microarray Design. *In*. IEEE, pp 1-8
- Geigenberger P** (2003) Response of plant metabolism to too little oxygen. *Current Opinion in Plant Biology* **6**: 247-256
- Geigenberger P, Fernie AR, Gibon Y, Christ M, Stitt M** (2000) Metabolic activity decreases as an adaptive response to low internal oxygen in growing potato tubers. *Biological Chemistry* **381**: 723-740
- Gentleman R, Carey VJ, Huber W, Irizarry RA, Dudoit S** (2005) Quality Assessment of Affymetrix GeneChip Data. *In* *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer
- Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J** (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**: R80
- Giorgi FM, Bolger AM, Lohse M, Usadel B** (2010) Algorithm-driven Artifacts in median polish summarization of Microarray data. *BMC bioinformatics* **11**: 553
- Glickman MH, Ciechanover A** (2002) The ubiquitin-proteasome proteolytic pathway: destruction for the sake of construction. *Physiological reviews* **82**: 373
- Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES** (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**: 531-537
- Gómez LD, Baud S, Gilday A, Li Y, Graham IA** (2006) Delayed embryo development in the ARABIDOPSIS TREHALOSE 6 PHOSPHATE SYNTHASE 1 mutant is associated with altered cell wall structure, decreased cell division and starch accumulation. *The Plant Journal* **46**: 69-84
- Gouy M, Gautier C** (1982) Codon usage in bacteria: correlation with gene expressivity. *Nucleic acids research* **10**: 7055
- Gurkan C, Lapp H, Hogenesch JB, Balch WE** (2005) Exploring trafficking GTPase function by mRNA expression profiling: use of the SymAtlas web-application and the Membrane datasets. *Methods in enzymology* **403**: 1-10
- Gustafsson M, Hörnquist M, Lundström J, Björkegren J, Tegnér J** (2009) Reverse engineering of gene networks with LASSO and nonlinear basis functions. *Annals of the New York Academy of Sciences* **1158**: 265-275
- Gyorffy B, Molnar B, Lage H, Szallasi Z, Eklund AC** (2009) Evaluation of microarray preprocessing algorithms based on concordance with RT-PCR in clinical samples. *PLoS One* **4**: e5645
- Hahn MW, Kern AD** (2005) Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Mol Biol Evol* **22**: 803-806
- Harr B, Schlotterer C** (2006) Comparison of algorithms for the analysis of Affymetrix microarray data as evaluated by co-expression of genes in known operons. *Nucleic Acids Res* **34**: e8
- Hartemink AJ** (2005) Reverse engineering gene regulatory networks. *Nat Biotechnol* **23**: 554-555
- Hastie T, Tibshirani R, Friedman J** (2001) *The elements of statistical learning*.

- Hastings MH, Maywood ES, O'Neill JS** (2008) Cellular circadian pacemaking and the role of cytosolic rhythms. *Curr Biol* **18**: R805-R815
- He F, Balling R, Zeng AP** (2009) Reverse engineering and verification of gene networks: principles, assumptions, and limitations of present methods and future perspectives. *J Biotechnol* **144**: 190-203
- Herth W, Schnepf E** (1980) The fluorochrome, calcofluor white, binds oriented to structural polysaccharide fibrils. *Protoplasma* **105**: 129-133
- Hinz M, Wilson IW, Yang J, Buerstenbinder K, Llewellyn D, Dennis ES, Sauter M, Dolferus R** (2010) Arabidopsis RAP2. 2: an ethylene response transcription factor that is important for hypoxia survival. *Plant physiology* **153**: 757
- Hirai MY, Yano M, Goodenowe DB, Kanaya S, Kimura T, Awazuhara M, Arita M, Fujiwara T, Saito K** (2004) Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in Arabidopsis thaliana. *Proceedings of the National Academy of Sciences of the United States of America* **101**: 10205
- Ho L, Crabtree GR** (2010) Chromatin remodelling during development. *Nature* **463**: 474-484
- Hobert O** (2008) Gene regulation by transcription factors and microRNAs. *Science* **319**: 1785
- Hochberg Y, Benjamini Y** (1990) More powerful procedures for multiple significance testing. *Stat Med* **9**: 811-818
- Holder CB, Cary JW** (1984) Soil oxygen and moisture in relation to Russet Burbank potato yield and quality. *American Journal of Potato Research* **61**: 67-75
- Horton P, Park KJ, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, Nakai K** (2007) WoLF PSORT: protein localization predictor. *Nucleic acids research* **35**: W585
- Hsiao YY, Chen YW, Huang SC, Pan ZJ, Chen WH, Tsai WC, Chen HH** (2011) Gene discovery using next-generation pyrosequencing to develop ESTs for Phalaenopsis orchids. *BMC Genomics* **12**: 360
- Huang J, Bowles D, Esfandiari E, Dean G, Carpita NC, Haughn GW** (2011) The Arabidopsis Transcription Factor LUH/MUM1 Is Required for Extrusion of Seed Coat Mucilage. *Plant physiology*
- Huang X, Madan A** (1999) CAP3: A DNA sequence assembly program. *Genome Research* **9**: 868
- Hubbell E, Liu WM, Mei R** (2002) Robust estimators for expression analysis. *Bioinformatics* **18**: 1585-1592
- Hundertmark M, Hinch DK** (2008) LEA(Late Embryogenesis Abundant) proteins and their encoding genes in Arabidopsis thaliana. *BMC genomics* **9**: 118
- Hurst LD, Smith NG** (1999) Do essential genes evolve slowly? *Curr Biol* **9**: 747-750
- Inan G, Zhang Q, Li P, Wang Z, Cao Z, Zhang H, Zhang C, Quist TM, Goodwin SM, Zhu J** (2004) Salt cress. A halophyte and cryophyte Arabidopsis relative model system and its applicability to molecular genetic analyses of growth and development of extremophiles. *Plant physiology* **135**: 1718
- Ip CC, Manam V, Hepler R, Hennessey JP, Jr.** (1992) Carbohydrate composition analysis of bacterial polysaccharides: optimized acid hydrolysis conditions for HPAEC-PAD analysis. *Anal Biochem* **201**: 343-349
- Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP** (2003) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res* **31**: e15
- Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP** (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**: 249-264
- Irizarry RA, Wu Z, Jaffee HA** (2006) Comparison of Affymetrix GeneChip expression measures. *Bioinformatics* **22**: 789-794
- Jackson MB, Colmer TD** (2005) Response and adaptation by plants to flooding stress. *Annals of Botany* **96**: 501
- Jensen LJ, Jensen TS, de Lichtenberg U, Brunak S, Bork P** (2006) Co-evolution of transcriptional and post-translational cell-cycle regulation. *NATURE-LONDON* **443**: 594
- Jeong H, Mason SP, Barabasi AL, Oltvai ZN** (2001) Lethality and centrality in protein networks. *Nature* **411**: 41-42
- Jones DT, Taylor WR, Thornton JM** (1992) The rapid generation of mutation data matrices from protein sequences. *Computer applications in the biosciences: CABIOS* **8**: 275
- Kapranov P, Cheng J, Dike S, Nix DA, Dutttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermüller J, Hofacker IL** (2007) RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316**: 1484
- Karin M, Ben-Neriah Y** (2000) Phosphorylation meets ubiquitination: the control of NF- B activity. *Annual Review of Immunology* **18**: 621-663

- Kevil CG, Walsh L, Laroux FS, Kalogeris T, Grisham MB, Alexander JS** (1997) An improved, rapid Northern protocol. *Biochemical and Biophysical Research Communications* **238**: 277-279
- Khanin R, Wit E** (2006) How scale-free are biological networks. *Journal of computational biology* **13**: 810-818
- Kilian J, Whitehead D, Horak J, Wanke D, Weini S, Batistic O, D'Angelo C, Bornberg-Bauer E, Kudla J, Harter K** (2007) The AtGenExpress global stress expression data set: protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses. *Plant J* **50**: 347-363
- Kirschner MW** (2005) The meaning of systems biology. *Cell* **121**: 503
- Kitano H** (2002) Systems biology: a brief overview. *Science* **295**: 1662
- Kloosterman B, De Koeyer D, Griffiths R, Flinn B, Steuernagel B, Scholz U, Sonnewald S, Sonnewald U, Bryan GJ, Prat S** (2008) Genes driving potato tuber initiation and growth: identification based on transcriptional changes using the POCI array. *Functional & integrative genomics* **8**: 329-340
- Koschützki D, Schreiber F** (2008) Centrality analysis methods for biological networks and their application to gene regulatory networks. *Gene Regul Syst Bio* **2**: 193-201
- Koßmann J, Visser RGF, Müller-Röber B, Willmitzer L, Sonnewald U** (1991) Cloning and expression analysis of a potato cDNA that encodes branching enzyme evidence for co-expression of starch biosynthetic genes. *Molecular and General Genetics MGG* **230**: 39-44
- Lasanthi-Kudahettige R, Magneschi L, Loreti E, Gonzali S, Licausi F, Novi G, Beretta O, Vitulli F, Alpi A, Perata P** (2007) Transcript profiling of the anoxic rice coleoptile. *Plant physiology* **144**: 218
- Lee PR, Cohen JE, Tendi EA, Farrer R, De Vries GH, Becker KG, Fields RD** (2004) Transcriptional profiling in an MPNST-derived cell line and normal human Schwann cells. *Neuron glia biology* **1**: 135-147
- Lee YS, Mulugu S, York JD, O'Shea EK** (2007) Regulation of a cyclin-CDK-CDK inhibitor complex by inositol pyrophosphates. *Science* **316**: 109
- Leinonen R, Sugawara H, Shumway M** (2011) The sequence read archive. *Nucleic acids research* **39**: D19
- Letunic I, Doerks T, Bork P** (2009) SMART 6: recent updates and new developments. *Nucleic acids research* **37**: D229
- Li C, Wong WH** (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc Natl Acad Sci U S A* **98**: 31-36
- Licausi F, Giorgi FM, Zenoni S, Osti F, Pezzotti M, Perata P** (2010) Genomic and transcriptomic analysis of the AP 2/ERF superfamily in *Vitis vinifera*. *BMC genomics* **11**: 719
- Licausi F, Van Dongen JT, Giuntoli B, Novi G, Santaniello A, Geigenberger P, Perata P** (2010) HRE1 and HRE2, two hypoxia inducible ethylene response factors, affect anaerobic responses in *Arabidopsis thaliana*. *The Plant Journal* **62**: 302-315
- Lim WK, Wang K, Lefebvre C, Califano A** (2007) Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks. *Bioinformatics* **23**: i282-288
- Liolios K, Chen I, Min A, Mavromatis K, Tavernarakis N, Hugenholtz P, Markowitz VM, Kyripides NC** (2009) The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. *Nucleic acids research* **38**: D346
- Liu CT, Yuan S, Li KC** (2009) Patterns of co-expression for protein complexes by size in *Saccharomyces cerevisiae*. *Nucleic acids research* **37**: 526
- Lodish H, Berk A, Matsudaira P, Kaiser CA, Krieger M, Scott MP, Zipursky SL, Darnell J** (2003) *Molecular cell biology*. WH Freeman and Company, New York
- Lohse M, Nunes-Nesi A, Krueger P, Nagel A, Hannemann J, Giorgi FM, Childs L, Osorio S, Walther D, Selbig J** (2010) Robin: an intuitive wizard application for R-based expression microarray quality assessment and analysis. *Plant physiology* **153**: 642
- Lohse M, Usadel B** (unpublished) Mercator: A fast and simple functional annotation web server for genome scale sequence data <http://mapman.gabipd.org/web/guest/app/mercator>. *In*,
- Lu Y, Huggins P, Bar-Joseph Z** (2009) Cross species analysis of microarray expression data. *Bioinformatics* **25**: 1476
- Lu Y, Zhou Y, Qu W, Deng M, Zhang C** (2011) A Lasso regression model for the construction of microRNA-target regulatory networks. *Bioinformatics*
- Ma S, Gong Q, Bohnert HJ** (2007) An *Arabidopsis* gene network based on the graphical Gaussian model. *Genome Research* **17**: 1614
- Macquet A, Ralet MC, Kronenberger J, Marion-Poll A, North HM** (2007) In situ, chemical and macromolecular study of the composition of *Arabidopsis thaliana* seed coat mucilage. *Plant and Cell Physiology* **48**: 984
- Madan Babu M, Teichmann SA** (2003) Evolution of transcription factors and the gene regulatory network in *Escherichia coli*. *Nucleic acids research* **31**: 1234

- Maeo K, Tokuda T, Ayame A, Mitsui N, Kawai T, Tsukagoshi H, Ishiguro S, Nakamura K** (2009) An AP2 type transcription factor, WRINKLED1, of *Arabidopsis thaliana* binds to the AW box sequence conserved among proximal upstream regions of genes involved in fatty acid synthesis. *The Plant Journal* **60**: 476-487
- Magness JR** (1920) Composition of gases in intercellular spaces of apples and potatoes. *Botanical Gazette* **70**: 308-316
- Malone J, Oliver B** (2011) Microarrays, deep sequencing and the true measure of the transcriptome. *BMC biology* **9**: 34
- Marchler-Bauer A, Anderson JB, Cherukuri PF, DeWeese-Scott C, Geer LY, Gwadz M, He S, Hurwitz DI, Jackson JD, Ke Z** (2005) CDD: a Conserved Domain Database for protein classification. *Nucleic acids research* **33**: D192
- Mardis ER** (2008) The impact of next-generation sequencing technology on genetics. *Trends in genetics* **24**: 133-141
- Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, Califano A** (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* **7 Suppl 1**: S7
- Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z** (2005) Genome sequencing in open microfabricated high density picoliter reactors. *Nature* **437**: 376
- McAdams HH, Shapiro L** (1995) Circuit simulation of genetic networks. *Science* **269**: 650
- McCarroll SA, Murphy CT, Zou S, Pletcher SD, Chin CS, Jan YN, Kenyon C, Bargmann CI, Li H** (2004) Comparing genomic expression patterns across species identifies shared transcriptional profile in aging. *Nature genetics* **36**: 197-204
- Meinke D, Muralla R, Sweeney C, Dickerman A** (2008) Identifying essential genes in *Arabidopsis thaliana*. *Trends in Plant Science* **13**: 483-491
- Millenaar FF, Okyere J, May ST, van Zanten M, Voeselek LA, Peeters AJ** (2006) How to decide? Different methods of calculating gene expression from short oligonucleotide array data will give different results. *BMC Bioinformatics* **7**: 137
- Miller JR, Koren S, Sutton G** (2010) Assembly algorithms for next-generation sequencing data. *Genomics* **95**: 315-327
- Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U** (2002) Network motifs: simple building blocks of complex networks. *Science* **298**: 824
- Min XJ, Butler G, Storms R, Tsang A** (2005) OrfPredictor: predicting protein-coding regions in EST-derived sequences. *Nucleic acids research* **33**: W677
- Mine T, Hiyoshi T, Kasaoka K, Ohyama A** (2003) CIP353 encodes an AP2/ERF-domain protein in potato (*Solanum tuberosum* L.) and responds slowly to cold stress. *Plant and Cell Physiology* **44**: 10
- Mizuno T, Takayasu H, Takayasu M** (2006) Correlation networks among currencies. *Physica A: Statistical Mechanics and its Applications* **364**: 336-442
- Mockler TC, Ecker JR** (2005) Applications of DNA tiling arrays for whole-genome analysis. *Genomics* **85**: 1-15
- Morrison TB, Weis JJ, Wittwer CT** (1998) Quantification of low-copy transcripts by continuous SYBR® Green I monitoring during amplification. *Biotechniques* **24**: 954-962
- Mudryj M, Devoto SH, Hiebert SW, Hunter T, Pines J, Nevins JR** (1991) Cell cycle regulation of the E2F transcription factor involves an interaction with cyclin A. *Cell* **65**: 1243-1253
- Mustroph A, Lee SC, Oosumi T, Zanetti ME, Yang H, Ma K, Yaghoubi-Masihi A, Fukao T, Bailey-Serres J** (2010) Cross-kingdom comparison of transcriptomic adjustments to low-oxygen stress highlights conserved and plant-specific responses. *Plant physiology* **152**: 1484
- Mutwil M, Klie S, Tohge T, Giorgi FM, Wilkins O, Campbell MM, Fernie AR, Usadel B, Nikoloski Z, Persson S** (2011) PlaNet: Combined Sequence and Expression Comparisons across Plant Networks Derived from Seven Species. *The Plant Cell Online* **23**: 895
- Mutwil M, Usadel B, Schutte M, Loraine A, Ebenhoh O, Persson S** (2010) Assembly of an interactive correlation network for the *Arabidopsis* genome using a novel heuristic clustering algorithm. *Plant Physiol* **152**: 29-43
- Nakano T, Suzuki K, Fujimura T, Shinshi H** (2006) Genome-wide analysis of the ERF gene family in *Arabidopsis* and rice. *Plant physiology* **140**: 411
- Narsai R, Ivanova A, Ng S, Whelan J** (2010) Defining reference genes in *Oryza sativa* using organ, development, biotic and abiotic transcriptome datasets. *BMC Plant Biology* **10**: 56

- Nasmyth K, Shore D** (1987) Transcriptional regulation in the yeast life cycle. *Science* **237**: 1162-1170
- Nicot N, Hausman JF, Hoffmann L, Evers D** (2005) Housekeeping gene selection for real-time RT-PCR normalization in potato during biotic and abiotic stress. *Journal of experimental botany* **56**: 2907
- Nielsen HB, Mundy J, Willenbrock H** (2007) Functional Associations by Response Overlap (FARO), a functional genomics approach matching gene expression phenotypes. *PLoS One* **2**: e676
- Nolan T, Hands RE, Bustin SA** (2006) Quantification of mRNA using real-time RT-PCR. *Nature Protocols* **1**: 1559-1582
- Obayashi T, Kinoshita K** (2011) COXPRESdb: a database to compare gene coexpression in seven model animals. *Nucleic acids research* **39**: D1016
- Ohme-Takagi M, Shinshi H** (1995) Ethylene-inducible DNA binding proteins that interact with an ethylene-responsive element. *The Plant Cell Online* **7**: 173
- Oka T, Nemoto T, Jigami Y** (2007) Functional analysis of Arabidopsis thaliana RHM2/MUM4, a multidomain protein involved in UDP-D-glucose to UDP-L-rhamnose conversion. *Journal of Biological Chemistry* **282**: 5389
- Okamuro JK, Caster B, Villarroel R, Van Montagu M, Jofuku KD** (1997) The AP2 domain of APETALA2 defines a large new family of DNA binding proteins in Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America* **94**: 7076
- Opgen-Rhein R, Schaefer J, Strimmer K, Strimmer MK** (2007) The GeneNet Package.
- Opgen-Rhein R, Strimmer K** (2007) Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach. *Stat Appl Genet Mol Biol* **6**: Article9
- Palaniswamy SK, James S, Sun H, Lamb RS, Davuluri RV, Grotewold E** (2006) AGRIS and AtRegNet. A platform to link cis-regulatory elements and transcription factors into regulatory networks. *Plant physiology* **140**: 818
- Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, Farne A, Holloway E, Kolesnykov N, Lilja P, Lukk M, Mani R, Rayner T, Sharma A, William E, Sarkans U, Brazma A** (2007) ArrayExpress--a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res* **35**: D747-750
- Pearl J** (2000) Causality: models, reasoning and inference. Cambridge Univ Press
- Peng F, Weselake R** (2011) Gene coexpression clusters and putative regulatory elements underlying seed storage reserve accumulation in Arabidopsis. *BMC genomics* **12**: 286
- Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, Niranjan V, Muthusamy B, Gandhi TK, Gronborg M, Ibarrola N, Deshpande N, Shanker K, Shivashankar HN, Rashmi BP, Ramya MA, Zhao Z, Chandrika KN, Padma N, Harsha HC, Yatish AJ, Kavitha MP, Menezes M, Choudhury DR, Suresh S, Ghosh N, Saravana R, Chandran S, Krishna S, Joy M, Anand SK, Madavan V, Joseph A, Wong GW, Schiemann WP, Constantinescu SN, Huang L, Khosravi-Far R, Steen H, Tewari M, Ghaffari S, Blobe GC, Dang CV, Garcia JG, Pevsner J, Jensen ON, Roepstorff P, Deshpande KS, Chinnaiyan AM, Hamosh A, Chakravarti A, Pandey A** (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res* **13**: 2363-2371
- Persson S, Wei H, Milne J, Page GP, Somerville CR** (2005) Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets. *Proc Natl Acad Sci U S A* **102**: 8633-8638
- Pruitt KD, Tatusova T, Maglott DR** (2006) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research* **35**: D61
- Qiu P, Wang ZJ, Liu KJ, Hu ZZ, Wu CH** (2007) Dependence network modeling for biomarker identification. *Bioinformatics* **23**: 198
- Rambaldi D, Giorgi FM, Capuani F, Ciliberto A, Ciccarelli FD** (2008) Low duplicability and network fragility of cancer genes. *Trends in Genetics* **24**: 427-430
- Ramsay G** (1998) DNA chips: state-of-the art. *Nat Biotechnol* **16**: 40-44
- Reverter A, Chan EK** (2008) Combining partial correlation and an information theory approach to the reversed engineering of gene co-expression networks. *Bioinformatics* **24**: 2491-2497
- Rhee SY, Beavis W, Berardini TZ, Chen G, Dixon D, Doyle A, Garcia-Hernandez M, Huala E, Lander G, Montoya M** (2003) The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic acids research* **31**: 224
- Rice P, Longden I, Bleasby A** (2000) EMBOSS: the European molecular biology open software suite. *Trends in genetics* **16**: 276-277

- Ryngajllo M, Childs L, Lohse M, Giorgi FM, Lude A, Selbig J, Usadel B** (unpublished) SLocX: Predicting Arabidopsis subcellular localization leveraging gene expression data.
- Sakuma Y, Liu Q, Dubouzet JG, Abe H, Shinozaki K, Yamaguchi-Shinozaki K** (2002) DNA-binding specificity of the ERF/AP2 domain of Arabidopsis DREBs, transcription factors involved in dehydration-and cold-inducible gene expression. *Biochemical and Biophysical Research Communications* **290**: 998-1009
- Sasaki K, Mitsuhashi I, Seo S, Ito H, Matsui H, Ohashi Y** (2007) Two novel AP2/ERF domain proteins interact with cis element VWRE for wound induced expression of the Tobacco tpoxN1 gene. *The Plant Journal* **50**: 1079-1092
- Schäfer J, Strimmer K** (2005) An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics* **21**: 754
- Schäffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, Koonin EV, Altschul SF** (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic acids research* **29**: 2994
- Scheffner M, Nuber U, Huibregtse JM** (1995) Protein ubiquitination involving an E1–E2–E3 enzyme ubiquitin thioester cascade. *Nature* **373**: 81-83
- Schena M, Shalon D, Davis RW, Brown PO** (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**: 467-470
- Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, Vingron M, Scholkopf B, Weigel D, Lohmann JU** (2005) A gene expression map of Arabidopsis thaliana development. *Nature genetics* **37**: 501-506
- Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, Vingron M, Scholkopf B, Weigel D, Lohmann JU** (2005) A gene expression map of Arabidopsis thaliana development. *Nat Genet* **37**: 501-506
- Schmittgen TD, Lee EJ, Jiang J** (2008) High-throughput real-time PCR. *Methods Mol Biol* **429**: 89-98
- Schneider M, Bairoch A, Wu CH, Apweiler R** (2005) Plant protein annotation in the UniProt Knowledgebase. *Plant physiology* **138**: 59
- Schneider M, Lane L, Boutet E, Lieberherr D, Tognolli M, Bougueleret L, Bairoch A** (2009) The UniProtKB/Swiss-Prot knowledgebase and its plant proteome annotation program. *Journal of proteomics* **72**: 567-573
- Schuster EF, Blanc E, Partridge L, Thornton JM** (2007) Correcting for sequence biases in present/absent calls. *Genome biology* **8**: R125
- Seipel K, Georgiev O, Schaffner W** (1992) Different activation domains stimulate transcription from remote ('enhancer') and proximal ('promoter') positions. *The EMBO journal* **11**: 4961
- Seringhaus M, Paccanaro A, Borneman A, Snyder M, Gerstein M** (2006) Predicting essential genes in fungal genomes. *Genome Res* **16**: 1126-1135
- Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T** (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research* **13**: 2498
- Shimamura T, Imoto S, Yamaguchi R, Miyano S** (2007) Weighted lasso in graphical Gaussian modeling for large gene network estimation based on microarray data. *Genome Informatics* **19**: 142-153
- Shipley B** (2002) Cause and correlation in biology: a user's guide to path analysis, structural equations and causal inference. Cambridge Univ Pr
- Signoretti S, Di Marcotullio L, Richardson A, Ramaswamy S, Isaac B, Rue M, Monti F, Loda M, Pagano M** (2002) Oncogenic role of the ubiquitin ligase subunit Skp2 in human breast cancer. *J Clin Invest* **110**: 633-641
- Smeekens S** (2000) Sugar-induced signal transduction in plants. *Annual Review of Plant Biology* **51**: 49-81
- Soares MB, Bonaldo MF, Jelene P, Su L, Lawton L, Efstratiadis A** (1994) Construction and characterization of a normalized cDNA library. *Proceedings of the National Academy of Sciences of the United States of America* **91**: 9228
- Southern EM** (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. *Journal of molecular biology* **98**: 503-517
- Srivastava A, Rogers WL, Breton CM, Cai L, Malmberg RL** (2011) Transcriptome Analysis of Sarracenia, an Insectivorous Plant. *DNA Research*
- Steinhauser D, Usadel B, Luedemann A, Thimm O, Kopka J** (2004) CSB. DB: a comprehensive systems-biology database. *Bioinformatics* **20**: 3647
- Stuart JM, Segal E, Koller D, Kim SK** (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302**: 249

- Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D, Schmidt D, O'Keeffe S, Haas S, Vingron M, Lehrach H, Yaspo ML** (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* **321**: 956-960
- Swarbreck D, Wilks C, Lamesch P, Berardini TZ, Garcia-Hernandez M, Foerster H, Li D, Meyer T, Muller R, Ploetz L** (2008) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic acids research* **36**: D1009
- Taiz L, Zeiger E** (2006) *Plant physiology*, Ed 4th. Sinauer Associates, Sunderland, Mass.
- Taji T, Sakurai T, Mochida K, Ishiwata A, Kurotani A, Totoki Y, Toyoda A, Sakaki Y, Seki M, Ono H** (2008) Large-scale collection and annotation of full-length enriched cDNAs from a model halophyte, *Thellungiella halophila*. *BMC Plant Biology* **8**: 115
- Tamura K, Dudley J, Nei M, Kumar S** (2007) MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. *Molecular biology and evolution* **24**: 1596
- Taniguchi M, Miura K, Iwao H, Yamanaka S** (2001) Quantitative assessment of DNA microarrays--comparison with Northern blot analyses. *Genomics* **71**: 34-39
- Tanya V, Ben L** (2008) A simple principle concerning the robustness of protein complex activity to changes in gene expression. *BMC Systems Biology* **2**
- Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN** (2003) The COG database: an updated version includes eukaryotes. *BMC bioinformatics* **4**: 41
- Therneau TM, Ballman KV** (2008) What Does PLIER Really Do? *Cancer Inform* **6**: 423-431
- Thimm O, Bläsing O, Gibon Y, Nagel A, Meyer S, Krüger P, Selbig J, Müller LA, Rhee SY, Stitt M** (2004) mapman: a user driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *The Plant Journal* **37**: 914-939
- Thomas PS** (1983) Hybridization of denatured RNA transferred or dotted nitrocellulose paper. *Methods in enzymology* **100**: 255
- Tibshirani R** (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*: 267-288
- Tirosh I, Weinberger A, Bezael D, Kaganovich M, Barkai N** (2008) On the relation between promoter divergence and gene expression evolution. *Molecular systems biology* **4**
- Tischler J, Lehner B, Fraser AG** (2008) Evolutionary plasticity of genetic interaction networks. *Nature genetics* **40**: 390-391
- Tsaparas P, Mariño-Ramírez L, Bodenreider O, Koonin EV, Jordan IK** (2006) Global similarity and local divergence in human and mouse gene co-expression networks. *BMC evolutionary biology* **6**: 70
- Tukey JW** (1977) *Exploratory data analysis*. Addison-Wesley, Reading, Mass.
- Tzafrir I, Dickerman A, Brazhnik O, Nguyen Q, McElver J, Frye C, Patton D, Meinke D** (2003) The Arabidopsis SeedGenes Project. *Nucleic acids research* **31**: 90
- Tzafrir I, Pena-Muralla R, Dickerman A, Berg M, Rogers R, Hutchens S, Sweeney TC, McElver J, Aux G, Patton D** (2004) Identification of genes required for embryo development in Arabidopsis. *Plant physiology* **135**: 1206
- Upton GJG** (1992) Fisher's exact test. *Journal of the Royal Statistical Society. Series A (Statistics in society)*: 395-402
- Usadel B, Blasing OE, Gibon Y, Retzlaff K, Hohne M, Gunther M, Stitt M** (2008) Global transcript levels respond to small changes of the carbon status during progressive exhaustion of carbohydrates in Arabidopsis rosettes. *Plant Physiol* **146**: 1834-1861
- Usadel B, Kuschinsky AM, Rosso MG, Eckermann N, Pauly M** (2004) RHM2 is involved in mucilage pectin synthesis and is required for the development of the seed coat in Arabidopsis. *Plant physiology* **134**: 286
- Usadel B, Obayashi T, Mutwil M, Giorgi FM, Bassel GW, Tanimoto M, Chow A, Steinhauser D, Persson S, Provart NJ** (2009) Coexpression Tools for Plant Biology: Opportunities for Hypothesis Generation and Caveats. *Plant Cell Environ*
- Usadel B, Poree F, Nagel A, Lohse M, Czedik-Eysenberg A, Stitt M** (2009) A guide to using MapMan to visualize and compare Omics data in plants: a case study in the crop species, Maize. *Plant Cell Environ* **32**: 1211-1229
- Van Dongen JT, Schurr U, Pfister M, Geigenberger P** (2003) Phloem metabolism and function have to cope with low internal oxygen. *Plant physiology* **131**: 1529

- Veiga DF, Vicente FF, Grivet M, de la Fuente A, Vasconcelos AT** (2007) Genome-wide partial correlation analysis of *Escherichia coli* microarray data. *Genet Mol Res* **6**: 730-742
- Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA** (2001) The sequence of the human genome. *Science* **291**: 1304
- Vera JC, Wheat CW, Fescemyer HW, Frilander MJ, Crawford DL, Hanski I, Marden JH** (2008) Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Molecular Ecology* **17**: 1636-1647
- Wang H, Hubbell E, Hu J, Mei G, Cline M, Lu G, Clark T, Siani-Rose MA, Ares M, Kulp DC** (2003) Gene structure-based splice variant deconvolution using a microarray platform. *Bioinformatics* **19**: i315
- Wang Z, Gerstein M, Snyder M** (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10**: 57-63
- Wang Z, Gerstein M, Snyder M** (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**: 57-63
- Wang Z, Li P, Fredricksen M, Gong Z, Kim CS, Zhang C, Bohnert HJ, Zhu JK, Bressan RA, Hasegawa PM** (2004) Expressed sequence tags from *Thellungiella halophila*, a new model to study plant salt-tolerance. *Plant science* **166**: 609-616
- Waterhouse AM, Procter JB, Martin D, Clamp M, Barton GJ** (2009) Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**: 1189
- Watson JD, Crick FHC** (1953) Molecular structure of nucleic acids. *Nature* **171**: 737-738
- Watt G, Leoff C, Harper AD, Bar-Peled M** (2004) A bifunctional 3, 5-epimerase/4-keto reductase for nucleotide-rhamnose synthesis in *Arabidopsis*. *Plant physiology* **134**: 1337
- Weckwerth W** (2010) Metabolomics: an integral technique in systems biology. *Metabolomics* **2**: 829-836
- Wei H, Persson S, Mehta T, Srinivasasainagendra V, Chen L, Page GP, Somerville C, Loraine A** (2006) Transcriptional coordination of the metabolic network in *Arabidopsis*. *Plant physiology* **142**: 762
- Weisstein EW** (2011) Gram Matrix. *MathWorld - A Wolfram Web Resource* <http://mathworld.wolfram.com/GramMatrix.html>
- Wen X, Fuhrman S, Michaels GS, Carr DB, Smith S, Barker JL, Somogyi R** (1998) Large-scale temporal gene expression mapping of central nervous system development. *Proceedings of the National Academy of Sciences of the United States of America* **95**: 334
- Werhli AV, Grzegorzczak M, Husmeier D** (2006) Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical Gaussian models and Bayesian networks. *Bioinformatics* **22**: 2523
- Werhli AV, Husmeier D** (2007) Reconstructing gene regulatory networks with Bayesian networks by combining expression data with multiple sources of prior knowledge. *Statistical applications in genetics and molecular biology* **6**: 15
- Western TL** (2006) Changing spaces: the *Arabidopsis* mucilage secretory cells as a novel system to dissect cell wall production in differentiating cells. *Botany* **84**: 622-630
- Western TL, Burn J, Tan WL, Skinner DJ, Martin-McCaffrey L, Moffatt BA, Haughn GW** (2001) Isolation and characterization of mutants defective in seed coat mucilage secretory cell development in *Arabidopsis*. *Plant physiology* **127**: 998
- Whittaker J** (2009) Graphical models in applied multivariate statistics.
- Wilke N, Sganga M, Barhite S, Miles MF** (1994) Effects of alcohol on gene expression in neural cells. *EXS* **71**: 49-59
- Wingender E, Chen X, Hehl R, Karas H, Liebich I, Matys V, Meinhardt T, Pruss M, Reuter I, Schacherer F** (2000) TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res* **28**: 316-319
- Wolber PK, Collins PJ, Lucas AB, De Witte A, Shannon KW** (2006) The Agilent In Situ-Synthesized Microarray Platform. *Methods in enzymology* **410**: 28-57
- Wolfe CJ, Kohane IS, Butte AJ** (2005) Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. *BMC bioinformatics* **6**: 227
- Wolfe AP, Matzke MA** (1999) Epigenetics: regulation through repression. *Science* **286**: 481
- Wong CE, Li Y, Labbe A, Guevara D, Nuin P, Whitty B, Diaz C, Golding GB, Gray GR, Weretilnyk EA** (2006) Transcriptional profiling implicates novel interactions between abiotic stress and hormonal responses in *Thellungiella*, a close relative of *Arabidopsis*. *Plant physiology*
- Wong CE, Li Y, Whitty BR, Diaz-Camino C, Akhter SR, Brandle JE, Golding GB, Weretilnyk EA, Moffatt BA, Griffith M** (2005) Expressed sequence tags from the Yukon ecotype of *Thellungiella* reveal that

- gene expression in response to cold, drought and salinity shows little overlap. *Plant molecular biology* **58**: 561-574
- Wu Z, Irizarry RA** (2004) Preprocessing of oligonucleotide array data. *Nat Biotechnol* **22**: 656-658; author reply 658
- Wu Z, Irizarry RA** (2005) Stochastic models inspired by hybridization theory for short oligonucleotide arrays. *J Comput Biol* **12**: 882-893
- Wunderlich Z, Mirny LA** (2006) Using the topology of metabolic networks to predict viability of mutant strains. *Biophys J* **91**: 2304-2311
- Xu K, Xu X, Fukao T, Canlas P, Maghirang-Rodriguez R, Heuer S, Ismail AM, Bailey-Serres J, Ronald PC, Mackill DJ** (2006) Sub1A is an ethylene-response-factor-like gene that confers submergence tolerance to rice. *Nature* **442**: 705-708
- Xu X, Pan S, Cheng S, Zhang B, Mu D, Ni P, Zhang G, Yang SPI, Li RPI, Wang JPI, Orjeda GPI, Guzman F, Torres M, Lozano R, Ponce O, Martinez D, De la Cruz G, Chakrabarti SKPI, Patil VU, Skryabin KGPI, Kuznetsov BB, Ravin NV, Kolganova TV, Beletsky AV, Mardanov AV, Di Genova A, Bolser DM, Martin DMPI, Li G, Yang Y, Kuang H, Hu Q, Xiong X, Bishop GJ, Sagredo BPI, Mejia N, Zagorski WPI, Gromadka R, Gawor J, Szczesny P, Huang SPI, Zhang Z, Liang C, He J, Li Y, He Y, Xu J, Zhang Y, Xie B, Du Y, Qu DPI, Bonierbale M, Ghislain M, Del Rosario Herrera M, Giuliano GPI, Pietrella M, Perrotta G, Facella P, O'Brien K, Feingold SEPI, Barreiro LE, Massa GA, Diambra L, Whitty BR, Vaillancourt B, Lin H, Massa AN, Geoffroy M, Lundback S, Dellapenna D, Robin Buell CPI, Sharma SK, Marshall DF, Waugh R, Bryan GJPI, Destefanis M, Nagy I, Milbourne DPI, Thomson SJ, Fiers M, Jacobs JMPI, Nielsen KLPI, Sonderkaer M, Iovene M, Torres GA, Jiang JPI, Veilleux RE, Bachem CWPI, de Boer J, Borm T, Kloosterman B, van Eck H, Datema E, Te Lintel Hekkert B, Goverse A, van Ham RC, Visser RG** (2011) Genome sequence and analysis of the tuber crop potato. *Nature*
- Yamada T, Bork P** (2009) Evolution of biomolecular networks—lessons from metabolic and protein interactions. *Nature Reviews Molecular Cell Biology* **10**: 791-803
- Yamamoto S, Suzuki K, Shinshi H** (1999) Elicitor responsive, ethylene independent activation of GCC box mediated transcription that is regulated by both protein phosphorylation and dephosphorylation in cultured tobacco cells. *The Plant Journal* **20**: 571-579
- Yonekura-Sakakibara K, Tohge T, Matsuda F, Nakabayashi R, Takayama H, Niida R, Watanabe-Takahashi A, Inoue E, Saito K** (2008) Comprehensive flavonol profiling and transcriptome coexpression analysis leading to decoding gene–metabolite correlations in Arabidopsis. *The Plant Cell Online* **20**: 2160
- Youm JW, Jeon JH, Choi D, Yi SY, Joung H, Kim HS** (2008) Ectopic expression of pepper CaPF1 in potato enhances multiple stresses tolerance and delays initiation of in vitro tuberization. *Planta* **228**: 701-708
- Yu H, Greenbaum D, Xin Lu H, Zhu X, Gerstein M** (2004) Genomic analysis of essentiality within protein networks. *Trends in genetics* **20**: 227-231
- Yu H, Luscombe NM, Qian J, Gerstein M** (2003) Genomic analysis of gene expression relationships in transcriptional regulatory networks. *Trends in genetics* **19**: 422-427
- Zampieri M, Soranzo N, Altafini C** (2008) Discerning static and causal interactions in genome-wide reverse engineering problems. *Bioinformatics* **24**: 1510-1515
- Zdobnov EM, Apweiler R** (2001) InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**: 847
- Zhang Y, Lai J, Sun S, Li Y, Liu Y, Liang L, Chen M, Xie Q** (2008) Comparison analysis of transcripts from the halophyte *Thellungiella halophila*. *Journal of Integrative Plant Biology* **50**: 1327-1335
- Zhou Y, Abagyan R** (2003) Algorithms for high-density oligonucleotide array. *Curr Opin Drug Discov Devel* **6**: 339-345
- Zhu JK** (2002) Salt and drought stress signal transduction in plants. *Annual Review of Plant Biology* **53**: 247-273
- Zimmermann P, Hirsch-Hoffmann M, Hennig L, Gruissem W** (2004) GENEVESTIGATOR. Arabidopsis microarray database and analysis toolbox. *Plant physiology* **136**: 2621
- Zotenko E, Mestre J, O'Leary DP, Przytycka TM** (2008) Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality. *PLoS Comput Biol* **4**: e1000140