Comparative study of RNA-seq- and Microarray-derived coexpression networks in Arabidopsis thaliana

BIOINFORMATICS

Vol. 00 no. 00 2012
Pages 1–8

Comparative study of RNA-seq- and Microarray-derived coexpression networks in *Arabidopsis thaliana*

Federico M. Giorgi [1,2,3]*, Cristian Del Fabbro [1], and Francesco Licausi [2]

[1]Institute of Applied Genomics, Udine, 33100 Italy.
[2]Scuola Superiore Sant'Anna, Pisa, 56124 Italy.
[3]University of Udine, Department of Agriculture and Environmental Sciences, Udine, 33100 Italy

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Coexpression networks are data-derived representations of genes behaving in a similar way across tissues and experimental conditions. They have been used for hypothesis generation and guilt-by-association approaches for inferring functions of previously unknown genes. So far, the main platform for expression data has been DNA microarrays, however the recent development of RNA-seq allows for higher accuracy and coverage of transcript populations. It is therefore important to assess the potential for biological investigation of coexpression networks derived from this novel technique in a condition-independent dataset.

Results: We collected 65 publicly available Illumina RNA-seq high quality *Arabidopsis thaliana* samples and generated Pearson correlation coexpression networks. These networks were then compared with those derived from analogous microarray data. We show how Variance-Stabilizing-Transformed (VST) RNA-seq data samples are the most similar to microarray ones, with respect to inter-sample variation, correlation coefficient distribution and network topological architecture. Microarray networks show a slightly higher score in biology-derived quality assessments such as overlap with the known protein-protein interaction network and edge ontological agreement. Different coexpression network centralities are investigated; in particular, we show how betweenness centrality is generally a positive marker for essential genes in *Arabidopsis thaliana*, regardless of the platform originating the data. In the end, we focus on a specific gene network case, showing that, although microarray data seem more suited for gene network reverse engineering, RNA-seq offers the great advantage of extending coexpression analyses to the entire transcriptome.

Contact: fgiorgi@appliedgenomics.org

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

The comprehensive understanding of the functional molecular mechanisms in the cell is a major challenge of modern biology (Kitano, 2002). Network representations have been successfully employed to capture various cellular relationships, ranging from protein-protein interactions (Breitkreutz *et al.*, 2008) to gene regulations (Yilmaz *et al.*, 2011) and metabolic conversions (Yamada and Bork, 2009). In these networks, biological entities (*e.g.*, genes, proteins, and metabolites) are represented as nodes, and their interactions are represented as edges. Biological networks can be assembled either by gathering all existing experimental knowledge over relationships between these entities (Breitkreutz *et al.*, 2008; Yilmaz *et al.*, 2011; Caspi *et al.*, 2012), or alternatively, when this kind of data is missing, they can be reconstructed by educated inferences based on data profiles (Hartemink, 2005). The latter approach, often dubbed "biological network reverse engineering" has great cost and time advantages over, for instance, classical forward genetics approaches, as it allows to reduce the experimental investigation to a subset of candidates potentially involved in a particular biological process (Wang *et al.*, 2012). In the specific case of reverse-engineered gene networks, the last decade has witnessed an avalanche in the availability of transcript expression data, provided mainly by microarray data (Farber and Lusis, 2008), which has in turn fueled the generation of "coexpression" networks (Vandepoele *et al.*, 2009). Coexpression networks are undirected graph representations of transcriptional co-behavior between genes within an organism. In such graphs, genes are connected by edges representing the similarity in their expression pattern across several experiments in which both genes are quantified. These similarities are usually calculated by simple methods like Pearson correlation (D'haeseleer *et al.*, 2000) or by more sophisticated approaches such as mutual information (Daub *et al.*, 2004) or linear modelling (Vasilevski *et al.*, 2012). A significance value is usually associated to each edge to estimate the amount of coexpression between any gene pair; in the case of Pearson correlation, this value is the Pearson Correlation Coefficient (PCC) which ranges from -1 (perfect negative linear coexpression) to +1 (perfect positive linear coexpression), whereas 0 (no correlation) signifies the overall lack of linear relationships between the transcript quantities of the two genes (Usadel *et al.*, 2009). Coexpression networks have been widely generated and exploited in studies aimed at the identification of novel gene functions via the "guilt-by-association" paradigm (Wolfe *et al.*, 2005) which assumes that similar expression patterns correspond to similar functions, times of activation, or cellular compartments (Ryngajllo *et al.*, 2011). Noteworthy successes of this approach have been obtained, for example, in identifying novel genes involved in plant cell wall synthesis (Persson *et al.*, 2005), starch metabolism (Fu and Xue, 2010) and in the human B-cell leukaemia signal transduction (Basso *et al.*, 2005). At the same time, several

*to whom correspondence should be addressed

standalone coexpression tools have been developed (Margolin *et al.*, 2006; Reverter and Chan, 2008; Opgen-Rhein and Strimmer, 2007) together with freely accessible online databases (Usadel *et al.*, 2009; Obayashi *et al.*, 2012).

Coexpression networks have also been studied topologically, generally showing a modular structure (Bassel *et al.*, 2011) and a scale-free distribution of their connectivity (Barabási and Oltvai, 2004; Iancu *et al.*, 2012), meaning that most genes have a small number of coexpressors, while a few genes behave as coexpression "hubs". However, the vast majority of these studies have neglected the growing availability of RNA-seq datasets, which provide several potential advantages over microarrays (Wang *et al.*, 2009). First of all, RNA-seq doesn't require prior knowledge of the studied organism, extending its usage even to poorly-characterized organisms (Balakrishnan *et al.*, 2012), both for "standard" purposes (transcriptome definition and differential gene expression) and for any derived application (such as sample clustering or coexpression analysis). Furthermore, RNA-seq allows for the identification of all transcripts, whereas microarrays usually cover only a subset of the transcriptome. For example, in *Arabidopsis thaliana* the most used microarray for quantitative transcriptomics (Edgar *et al.*, 2002; Giorgi *et al.*, 2010), the Affymetrix ATH1, covers reliably and unambiguously only 21,377 genes (based on the most recent CustomCDF annotation (Dai *et al.*, 2005)), which is only a subset of the entire genome (27,416 genes in the TAIR10 annotation release (Swarbreck *et al.*, 2008)). Finally, RNA-seq has the potential to detect novel transcribed loci on annotated genomes (Roberts *et al.*, 2011), splicing variants (Richard *et al.*, 2010), and allele-specific events (Zhang *et al.*, 2009), massively increasing the investigative capability over these molecular phenomena.

In the present study, we try to determine whether RNA-seq data can be efficiently used for coexpression analysis. In order to do so, we derive coexpression networks from a set of 65 high quality Illumina-based RNA-seq *Arabidopsis thaliana* experiments, and compare them with those extracted from biologically analogous and equally sized microarray data. We determine the nature of these networks, both biologically and topologically, with an overview on different network centralities and their association with gene essentiality (Jeong *et al.*, 2001) in Arabidopsis. Finally, we focus on two specific gene cases, showing how the increased detection range of RNA-seq can indeed cover missing areas of the coexpression networks.

2 METHODS

2.1 Dataset selection and preprocessing

We downloaded 95 samples from the NCBI Sequence Read Archive (Leinonen *et al.*, 2011). After the SRA files were collected, the archives were extracted and saved in FASTQ format. The FASTQ files were trimmed using ERNE-FILTER software[1] with default parameters and minimum read length at least 70% of the original size. All samples with less than 30% of the reads surviving the trimming process were discarded. Surviving reads (986,482,909) were aligned on the TAIR10 *Arabidopsis thaliana* reference genome (Swarbreck *et al.*, 2008) using TopHat v2.0.4 (Trapnell *et al.*, 2009). Samples where less than 30% of the trimmed reads aligned on the Arabidopsis transcriptome were not considered for coexpression analysis. The 831,286,856 aligned reads (corresponding to 65 final samples, see file S1 for details) were then summarized at the gene level based on the TAIR10 annotation by Cuffdiff v2.0.2 (Trapnell *et al.*, 2010), which provided also the raw count and the RPKM (Reads Per Kilobase of gene model per Million mapped reads) values (Mortazavi *et al.*, 2008). Raw counts were modified into normalized values via the Variance Stabilizing Transformation (VST) method implemented in the R package DESeq (Anders and Huber, 2010). Analogous tissue and condition microarray datasets (see file S1) were downloaded from Gene Expression Omnibus (Edgar *et al.*, 2002) and normalized using MAS5 (Hubbell *et al.*, 2002). All microarray samples were quality tested using the Robin software (Lohse *et al.*, 2010).

2.2 Construction and evaluation of coexpression networks

Pearson correlation coefficients between all gene pairs were calculated for each dataset, and networks with varying correlation coefficient thresholds were extracted. Only positive PCCs above the specified thresholds were converted into a network edge, in order to allow for the application of network quality assessments based on the assumption of co-presence (or co-absence) of gene expression; specifically: the existence of protein-protein interaction and/or activating gene regulation, and the belonging to the same functional group (Jordan *et al.*, 2004; Vandepoele *et al.*, 2009).

The Mapman-based iso-ontological percentage in the networks was obtained by counting the number of edges containing two genes with at least one shared Mapman ontology term (Klie and Nikoloski, 2012). Due to the highly grained nature of the Mapman bins, the ontology was trimmed to the third branch (*e.g.*, 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 in the network. Edges containing genes of unknown function (Mapman bin 35) were ignored for this calculation.

The *Arabidopsis thaliana* reference protein-protein interaction network, collecting 96,827 protein interactions, was obtained from AtPin version Jun-2010 (Brandüao *et al.*, 2009). The reference genetic interaction network, composed by 11,355 positive genetic interactions, was obtained from AtRegNet version 15-Sep-2010 (Yilmaz *et al.*, 2011).

The fit of degree distribution of the coexpression networks to a power law was calculated as in (Brohée *et al.*, 2008).

Networks were graphically represented using Cytoscape (Smoot *et al.*, 2011); node coloring was applied to the networks following the Mapman ontology described within the CorTo tool[2] and in file S2.

2.3 Network Centrality and Essential Genes

A manually curated list of 481 essential genes was obtained from SeedGenes v8 (Tzafrir *et al.*, 2003). Degree, shortest path betweenness, and clustering coefficient network centralities were calculated with an implementation of the JUNG library[3]. ROC curves (Beck and Shultz, 1986) were generated for essential genes by using a sliding threshold (τ), namely every different degree (τ_{deg}), betweenness (τ_{btw}), and clustering coefficient (τ_{clc}) values in the population, and then calculating the number of true positive essential genes above each τ. Joint centrality ROC curves were calculated by averaging the ranking in the three centralities (degree, clustering coefficient, and betweenness) for each gene. Further details are available in the Supplementary material (file S3).

2.4 Ontology enrichment analysis

Mapman ontology term over-representation analyses were performed using the most recent *Arabidopsis thaliana* Mapman TAIR9 mapping (Thimm *et al.*, 2004) via a Bonferroni-corrected Fisher's Exact Test (Upton, 1992) as implemented in the CorTo software. Over-represented Mapman bin pairs (to estimate the functional enrichment in edges) were also calculated with an

[1] Available: http://erne.sourceforge.net

[2] Available: http://www.usadellab.org/cms/index.php?page=corto

[3] Available: http://jung.sourceforge.net

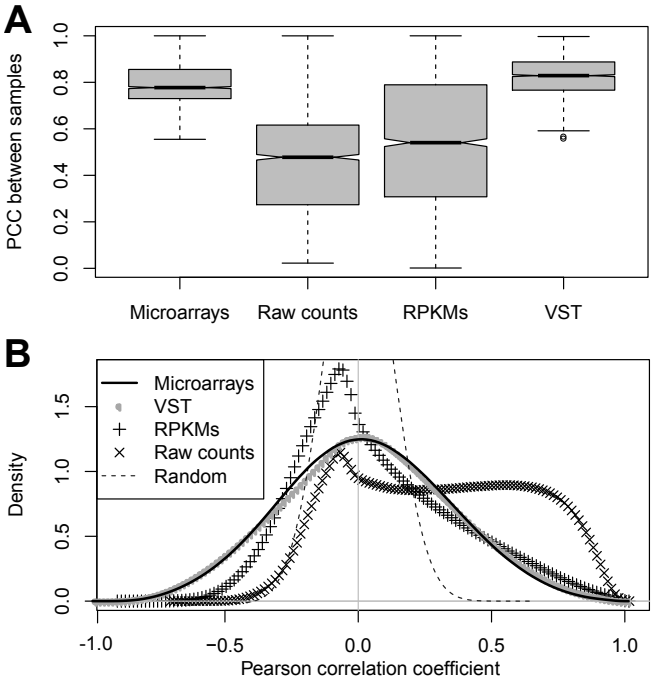

Figure 1. Correlation in expression datasets. (**A**) Box plots showing PCCs between samples. (**B**) Density distributions of PCCs between genes.

implementation of the Fisher's Exact Test based on the theoretical maximum number of combinations between two Mapman bins.

3 RESULTS

3.1 Properties of the coexpression networks

We collected 65 Illumina RNA-seq samples (totalling 831,286,856 aligned reads) representing a wide range of *Arabidopsis thaliana* tissues and conditions. Expression values for each gene were calculated 1) as the count of aligned reads over the transcript sequence ("raw counts"), 2) after RPKM normalization (Mortazavi *et al.*, 2008), which simply adjusts raw counts using the number of mapped reads and gene lengths, and 3) after VST normalization, a method designed to transform count data into values distributed homoscedastically (Anders and Huber, 2010) (see file S3). We decided to pair each of the 65 RNA-seq samples with corresponding microarray experiments, via a manual research of the Gene Expression Omnibus database (Edgar *et al.*, 2002) for the best tissue/condition/ecotype match, in order to keep comparability between these two data sources as high as possible. Despite this, sample clustering shows a clear distinction between the two platforms. However, VST normalization generally brings RNA-seq samples hierarchically closer to microarrays than RPKM normalization or raw counts (file S4).

Correlating samples to each other shows that microarrays are more similar to each other (Fig. 1A). It is known that even with a single-array normalization method such as MAS5, which doesn't overestimate sample correlation (Lim *et al.*, 2007), microarray samples tend to be highly correlated to each other (Giorgi *et al.*, 2010). Correlation coefficients between samples are much lower in publicly available *Arabidopsis thaliana* RNA-seq data when compared to similarly sized combinations of randomly taken publicly available microarrays (Giorgi *et al.*, 2010). RPKM normalization, supposed to increase comparability between samples (Mortazavi *et al.*, 2008), is indeed reducing sample variability when compared to raw counts. VST normalization yields a high inter-sample correlation, comparable to microarray levels (Fig. 1A).

Concerning PCCs between genes, which is the basic parameter on which coexpression networks are built in most studies (Usadel *et al.*, 2009), microarray data yield a symmetrical, bell-shaped distribution (Fig. 1B, solid line), almost perfectly overlapped by VST-normalized RNA-seq data (Fig. 1B, dotted grey line). RNA-seq raw count data show a bimodal correlation distribution (Fig. 1B, cross-pointed line), as noted before in a smaller dataset comparison (Iancu *et al.*, 2012), where this increase was explained by the greater sensitivity and dynamic range of RNA-seq data. RPKM normalization shows a bell-shaped curve slightly skewed towards negative values, and not centered over a zero value (Fig. 1B, plus-pointed line). All data generate correlations between gene expressions which are higher than the random PCC distribution (Fisher, 1915) (Fig. 1B, dashed line). Since the expected random distribution depends on the number of samples in the original dataset (not on the number of variables/genes), these differences are not merely due to the different number of genes detected by microarrays vs. RNA-seq. Distributions of correlation coefficients for raw RNA-seq counts approach a monomodal distribution for log2-transformed data and Spearman correlation coefficients (file S5).

An immediate consequence of different PCC distributions is the difference in the relationship between coexpression network size and PCC threshold used to build it (Fig. 2A). Microarray data, given the same threshold, yield smaller networks than RNA-seq, message that should warn against the application of the same rule-of-thumb significance thresholds applied before in coexpression studies, with PCC=0.7 as a frequently used value (Jordan *et al.*, 2004; Luo *et al.*, 2007; Usadel *et al.*, 2009).

For each PCC threshold plotted in Fig. 2A, we calculated several biological and topological properties. The overlap of a coexpression network with protein-protein interaction networks is a common criterion for biology-based network quality assessment (Lim *et al.*, 2007; Usadel *et al.*, 2009). In fact, direct Pearson correlation has successfully been used before for identifying proteins belonging to the same complex, as these usually require genes to be coexpressed in order to yield stoichiometrically balanced proteic products (Teichmann and Babu, 2002). In this respect, microarray data allow to achieve the highest performance in terms of Matthews coefficient (Baldi *et al.*, 2000) and accuracy in the overlap between coexpression connections and the 96,827 experimentally validated *Arabidopsis thaliana* physical protein-protein interactions (Brandüao *et al.*, 2009) (Fig. 2B and file S6). RNA-seq data allow for positive coexpression-based estimation (*i.e.*, positive Matthews coefficients) only for PCCs higher than 0.8, with raw counts achieving higher prediction power than normalized counts. The accuracy of the coexpression analysis shows a constant increase proportional to the threshold stringency applied to generate the networks (file S6).

Regardless of the expression measurement method or the PCC threshold applied, edges derived from coexpression networks are always negative or null predictors (Fig. 2C) of the manually curated collection of 11,355 *Arabidopsis thaliana* transcription factor-target relationships (Yilmaz *et al.*, 2011). While direct, static coexpression

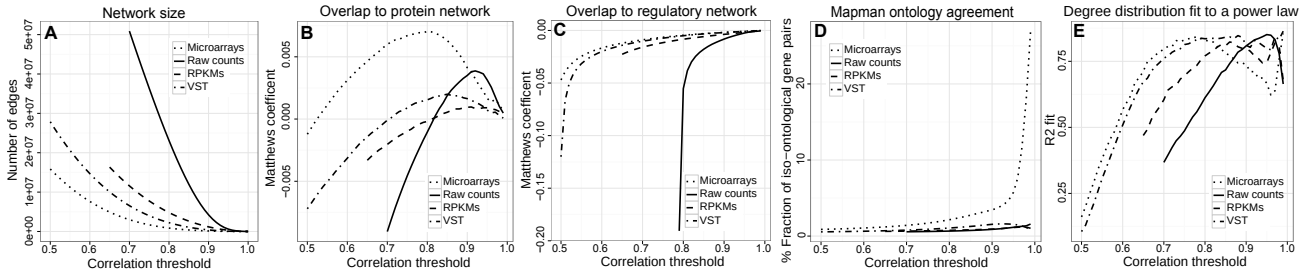

Figure 2. Properties of coexpression networks at different PCC thresholds. (**A**) Network sizes by number of edges. (**B**) Overlap to *Arabidopsis thaliana* AtPin protein-protein interaction network. (**C**) Overlap to *Arabidopsis thaliana* AtRegNet transcription factor-target network. (**D**) Percentage of edges connecting genes with identical Mapman ontology. (**E**) R^2 fit of the degree distribution to a power law.

measures such as Pearson correlation are known to be positive estimators of static protein interactions (Zampieri *et al.*, 2008), they are usually counter-predictive or meaningless for causal relationships like transcription factor-target interactions. In these cases, more complex methods that can remove indirect and spurious edges are suggested, such as Partial Correlation (Schäfer *et al.*, 2001; de la Fuente *et al.*, 2004) or LASSO (Vasilevski *et al.*, 2012). However, even full partial correlation networks (Schäfer *et al.*, 2001) derived from both microarray and RNA-seq data have negative Matthews coefficients with the annotated Arabidopsis genetic network (file S7).

Another common evaluation method of data-derived networks is the assessment of the ontological nature of the edges (Lim *et al.*, 2007), which assumes that a positive-hit edge is the one connecting genes sharing at least one biological function. In order to do so, we assessed the edges of our coexpression networks using the Mapman ontology (Thimm *et al.*, 2004) (Fig. 2D), a plant-oriented finely grained version of the more generic Gene Ontology (Klie and Nikoloski, 2012). The ontological assessment is partially biased, because also genes with different functions can be coregulated in reality (Lim *et al.*, 2007). However, it guarantees a qualitative estimate—independent from experimentally proven interactions—for the 63.1% of Arabidopsis genes which are functionally annotated by Mapman; since the fraction of annotated genes is slightly higher in the population represented by the ATH1 microarray (67.9%, file S8), we used the intersection between genelists in the two data types to perform this analysis. Our data show that microarray-derived networks (Fig. 2D) possess the highest percentage of iso-ontological edges, followed by VST, RPKM and raw counts. A clear connection between threshold stringency and the percentage of edges sharing genes belonging to at least one common ontological term is evident only for microarray-based networks (Fig. 2D), warning against the direct application of functional clustering methods (Mutwil *et al.*, 2010) on RNA-seq-derived coexpression networks.

We also analysed the networks topologically by fitting their global connectivity to a power law distribution (Brohée *et al.*, 2008). All coexpression networks, regardless of the type of transcript data used, show a good fit to a scale-free distribution (Fig. 2E), with R^2 always above 0.7 whenever the PCC threshold reduces the number of connections below 10^7 (Fig. 2A) (Barabási and Oltvai, 2004). There seems to be an optimal scale-free PCC threshold, which is 0.78 for microarray, 0.86 for RPKM, 0.88 for VST and 0.95 for raw count networks. These thresholds correspond also to a positive overlap (Fig. 2B) with the *Arabidopsis thaliana* protein-protein interaction network (also scale-free (Brandüao *et al.*, 2009)).

3.2 Centrality and Essentiality in Coexpression networks

We now focus on specific networks, selected by visual inspection based on the best overall qualities (Fig. 2) at three different sizes (simply dubbed “small”, “medium”, and “large”) and summarized in Table 1. On these networks, we calculated for each gene three different measures for network centrality (Koschützki and Schreiber, 2008), specifically: degree (number of connections), clustering coefficient (normalized amount of connections between the gene’s neighbours) and shortest path betweenness (normalized number of

Table 1. Properties of three ranges of similarly edge-sized *Arabidopsis thaliana* coexpression networks from different input data.

Size range	Data source	PCC threshold	Number of edges	Average node degree	PPI Matthews	% Fraction of iso-ontological edges	Power law R^2
Small	Microarrays	0.90	132,558	26.53	$3.247 \cdot 10^{-3}$	3.442	0.737
	VST	0.94	111,543	20.95	$1.004 \cdot 10^{-3}$	1.582	0.810
	RPKM	0.97	115,485	19.40	$6.790 \cdot 10^{-4}$	1.210	0.819
	Raw counts	0.97	158,314	16.32	$1.623 \cdot 10^{-3}$	1.284	0.832
Medium	Microarrays	0.80	861,676	58.02	$6.725 \cdot 10^{-3}$	2.145	0.831
	VST	0.86	911,096	50.99	$1.904 \cdot 10^{-3}$	1.317	0.843
	RPKM	0.91	954,178	58.42	$9.333 \cdot 10^{-4}$	0.939	0.788
	Raw counts	0.94	997,889	54.29	$3.247 \cdot 10^{-3}$	1.082	0.843
Large	Microarrays	0.70	2,994,674	155.32	$6.222 \cdot 10^{-3}$	1.411	0.796
	VST	0.78	2,857,389	118.19	$1.354 \cdot 10^{-3}$	0.989	0.832
	RPKM	0.84	3,011,806	129.08	$4.043 \cdot 10^{-4}$	0.812	0.814
	Raw counts	0.91	3,161,796	143.95	$3.552 \cdot 10^{-3}$	0.922	0.820

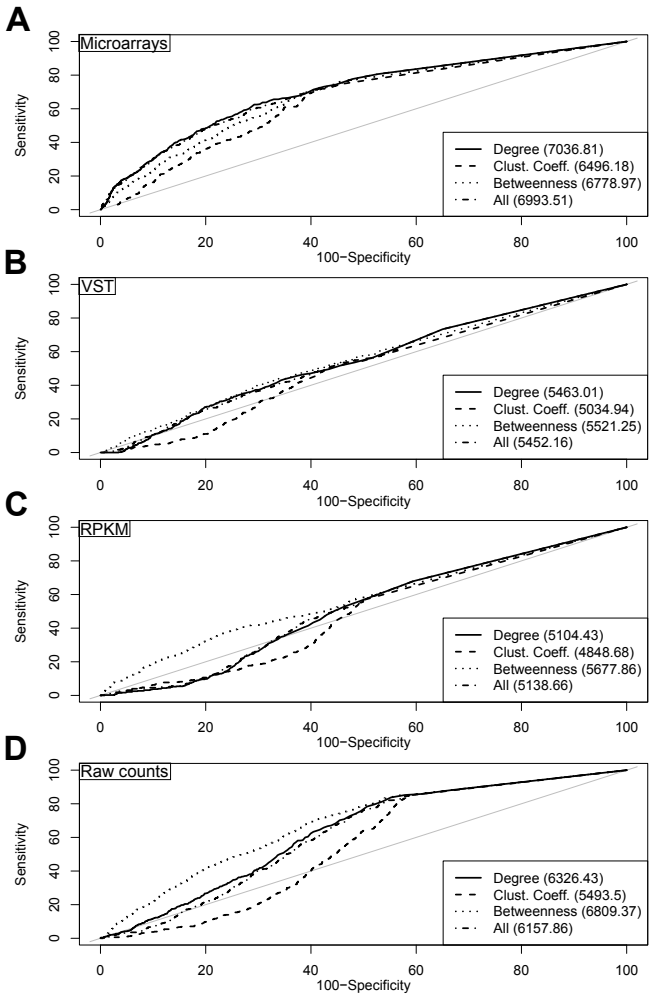

Figure 3. ROC curves showing the discerning capability for essentiality of three centrality measures in medium-sized coexpression networks. (**A**) Microarray. (**B**) RNA-seq VST. (**C**) RNA-seq RPKM. (**D**) RNA-seq raw counts. Areas Under the ROC are indicated in parentheses (total area 10,000).

times the gene is crossed by a shortest path connecting two other genes). Regardless of the data and network size used, we constantly see a positive correlation between degree and betweenness, and a negative correlation between clustering coefficient and betweenness (file S9). In network biology, a strong association between centrality and gene function has been observed for a long time: for instance, essential genes products tend to have more distinct interactors (*i.e.*, a higher degree) than non-essential ones (Jeong *et al.*, 2001), high betweenness genes tend to be key network regulators (Joy *et al.*, 2005) and cancer genes have a significantly higher degree and clustering coefficient than other genes (Rambaldi *et al.*, 2008). In coexpression analysis, this relationship is less investigated; however it has been proven that embryonic-essential Arabidopsis genes (Tzafrir *et al.*, 2003) have a significantly higher degree than the rest of the transcriptome in microarray-based coexpression networks (Mutwil *et al.*, 2010). The same is true in our microarray, VST and raw count (but not RPKM) gene networks, where the essential genes are consistently and significantly more connected than non-essential ones (Table 2). In microarray-derived networks, degree,

Table 2. Wilcoxon tests p-values testing the distributions of centrality values of essential vs. non-essential genes in similarly sized coexpression networks.

	PCC threshold	*Degree*	*Clustering coefficient*	*Betweenness*
Microarrays	**0.90**	10^{-55}	10^{-53}	10^{-55}
	0.80	10^{-59}	10^{-35}	10^{-48}
	0.70	10^{-60}	10^{-27}	10^{-44}
RNA-seq VST	**0.94**	10^{-4}	0.036	0.002
	0.86	10^{-4}	0.383	10^{-5}
	0.78	10^{-5}	0.135	10^{-10}
RNA-seq RPKM	**0.97**	1	1	1
	0.91	0.208	0.881	10^{-45}
	0.84	0.121	1	10^{-18}
RNA-seq raw counts	**0.97**	10^{-11}	0.008	10^{-13}
	0.94	10^{-24}	10^{-5}	10^{-8}
	0.91	10^{-27}	0.902	10^{-43}

clustering coefficient and betweenness in all three thresholds analyzed are positive predictors for essentiality (Table 2 and Fig. 3A). For networks derived from RPKM-normalized data, betweenness is the only parameter significantly associated with essentiality, albeit not in high threshold networks (Fig. 3C), while RNA-seq raw counts and VST based networks show again the tendency of essential genes to possess a high degree and a high betweenness (Fig. 3B and 3D). The connection with clustering coefficient is lost in larger RNA-seq networks (Table 2).

In general, essential genes possess a significantly higher betweenness in almost all *Arabidopsis thaliana* coexpression networks (file S9), while there seems to be no advantage in combining all three centralities by average gene ranking (Fig. 4), an approach utilized before for essential gene detection (Joy *et al.*, 2005). It is clear however, that coexpression network degree alone, as stated before for degree in protein-protein interaction networks (Wuchty, 2002), is not always a sufficient predictor for gene essentiality in RNA-seq networks, while it is a valid predictor in microarray networks.

3.3 Biological insights from coexpression networks

We then functionally annotated and intersected the networks described in Table 1 and looked at them at a greater detail (Fig. 4, file S2 and S10). There is a very low size overlap between microarray- and RNA-seq-derived coexpression networks (Fig. 4A). This is perhaps not entirely surprising given the technical low correlation between these two techniques, especially for high and low transcript abundances (Wang *et al.*, 2009). However, microarray-derived networks are more similar to VST-derived ones (12.7% shared edges relative to total microarray network size) than those based on RPKM or raw counts (respectively, 6.2% and 5.0%). Also, the overlap between RNA-seq networks is constantly below 50% of their total sizes (file S10, the highest overlap is visible between raw counts and RPKM), posing an interesting caveat about the comparability of coexpression inferences made with differently normalized RNA-seq data (the same issue was reported before for different normalization procedures on microarray data (Lim *et al.*, 2007)). Each of the networks derived by our analysis seems to be focusing on different parts of the cellular transcriptome (Fig. 4B): for example microarrays show a high propensity of coexpression for RNA

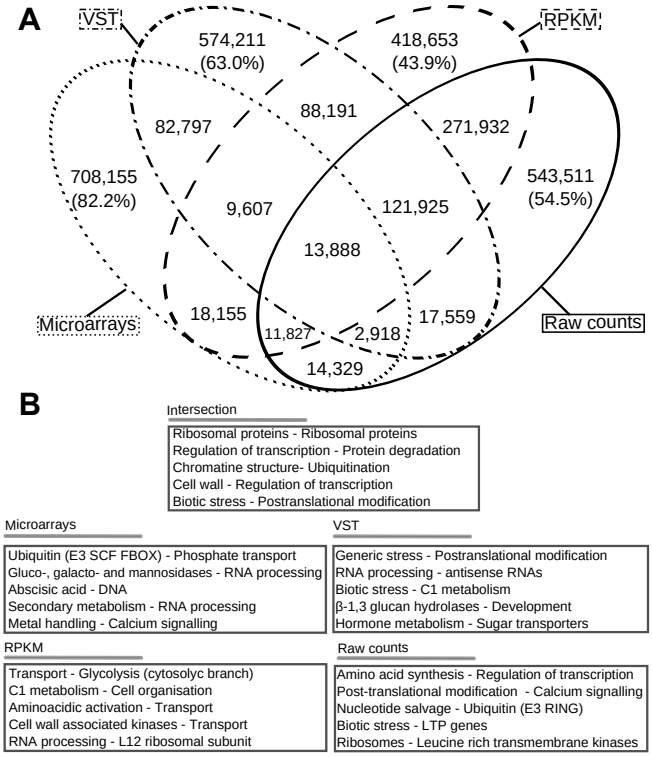

Figure 4. Overlap assessment and functional overview of medium sized coexpression networks described in Table 1. (**A**) Venn diagrams for relative distribution of network edges. In brackets, percentage of edges specific to a particular data type. (**B**) Selection of significantly over-represented (p Bonferroni corrected $< 10^{-100}$) connections between Mapman functional classes.

processing genes with other functional areas, and RPKM-based networks describe several transport-related coexpressions.

An intersection of all networks describes a few backbone coexpression structures well known in literature, like the relationship between cell wall synthesis/degradation and regulation of transcription (Mutwil *et al.*, 2010) or the one between biotic stress and post-translational modification mechanisms (Mishra *et al.*, 2006). Genes coding for ribosomal proteins are also highly coexpressed to each other (Fig. 4B and file S2).

One of the great advantages of coexpression analysis is its possibility to propose novel candidate genes for incompletely characterized biological pathways (Persson *et al.*, 2005; Vasilevski *et al.*, 2012). RNA-seq allows a quantitative assessment of the entire transcriptome, therefore extending this type of inference over genes where microarray-based coexpression investigations are not an option. One of these genes is *Sphavata* (At5g21960), a poorly characterized ethylene-responsive factor gene (ERF) known to be induced by jasmonate (Giuntoli *et al.*, 2009). In fact, the top 100 correlators for *Sphavata* calculated by all three normalizations of RNA-seq data (file S11), show a significant over-representation for genes involved in jasmonate metabolism (Mapman bin 17.7) and belonging to the ERF family (Mapman bin 27.3.3).

On the other hand, a well-studied gene for coexpression analysis is *RHM2*, a NDP-L-rhamnose synthase involved in polysaccharide branching and necessary for Arabidopsis seed coat mucilage pectin biosynthesis (Usadel *et al.*, 2004). This gene has been used as a coexpression bait in order to identify novel genes involved in the mucilage pathway (Haughn and Western, 2012) by using correlation analysis over a microarray seed dataset (Vasilevski *et al.*, 2012). Our analysis shows the potential in identifying novel genes coexpressing with *RHM2* (Fig. 5 and file S12): among the top 10 positive correlators identified using RNA-seq data, four genes not present on the Arabidopsis microarray were identified, three of which putatively involved in polysaccharide synthesis (Swarbreck *et al.*, 2008): At2g26100 (a putative galactosyltransferase), At3g06550 (*RWA2*, involved in polysaccharide O-acetylation) and At5g57270 (a putative N-acetylglucosaminyltransferase). In total, six coexpressors of *RHM2* are already annotated as cell-wall related (Fig. 5, green nodes); a particular coexpressor found with RNA-seq data (*UGP2*) is an essential gene active on nucleotide sugar pyrophosphorylation (Meng *et al.*, 2009). All these genes may be novel candidates in the pectin biosynthesis pathway. At the same time, RNA-seq-based coexpression is able to identify GAUT1, a α-1,4-galacturonosyltransferase already known to be active, as RHM2, within the pectin branching metabolism (Sterling *et al.*, 2006).

4 DISCUSSION

Our results describe the first large scale (65 samples) attempt to use RNA-seq data collected from multiple tissues and experimental conditions for gene network reverse engineering. We show that coexpression networks generated from this novel technology are indeed realistic (Fig. 2) and accurate, with accuracy increasing together with network stringency, validating the assumption that RNA-seq-based coexpression is a better-than-random selector of real biological relationships (file S6). However, our results show that microarray-based coexpression networks based on simple correlation achieve a higher similarity to biological networks, and at the same time show a low overlap with RNA-seq based representations (Fig. 4). All RNA-seq networks show a scale-free topology (Fig. 2E) as previously noted on a smaller dataset (Iancu *et al.*, 2012). In particular, the usage of raw counts with respect to the popular RPKM-normalized counts seems to be advantageous in correlation based analysis for several of the properties investigated here (Fig. 2, Fig. 3 and Table 1).

VST-normalized data possess microarray-like behavior with regards to correlation coefficient distribution and topological network properties (size and degree distribution, Fig. 2A and 2E). Amongst RNA-seq data, VST networks also possess the highest proximity to microarray networks detected by hierarchical clustering (file S4) and edge intersection (file S10); however, this is not directly translated into similar biological network properties.

We also find that coexpression network betweenness centrality can be calculated from RNA-seq data and used as a positive marker for *Arabidopsis thaliana* essential genes (Table 2). The task of identifying essential genes has been called the "most important task of genomics-based target validation" (Chalker and Lunsford, 2002), since these genes are extremely important not only to understand the minimal requirements for life (Li *et al.*, 2011), but also because they are excellent drug targets (Cole, 2002).

Another important application of coexpression analysis is in the identification of novel genes and novel gene functions. To this respect, we show how RNA-seq data can be complementary to microarray data in describing the functional neighborhood of a pectin

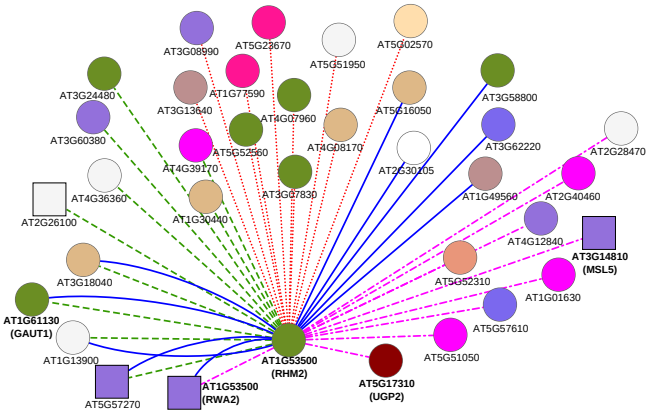

Figure 5. Coexpression network of *RHM2*, obtained by merging the top 10 correlators calculated from four different input data: microarrays (dotted line), RNA-seq VST (dot-dashed line), RNA-seq RPKM (dashed line) and RNA-seq raw counts (solid line). Nodes depicted as rectangles are not represented by the ATH1 Arabidopsis array.

metabolism gene (Fig. 5) or to confirm the connection with jasmonate of a poorly characterized putative transcription factor (file S11). There are at least 6,953 *Arabidopsis thaliana* genes annotated on the TAIR10 genome but not represented by any probeset on the Affymetrix ATH1 microarray platform; 3,578 of these genes have no functional annotation—neither experimentally inferred, nor predicted *in silico* (Thimm *et al.*, 2004)—, which gives RNA-seq the unique possibility to functionally investigate a previously uncovered portion of the transcriptome. This potential can indeed be transposed to other organisms as well, given the fair conservation of coexpression across species, at least in the plant kingdom (Movahedi *et al.*, 2012). All data investigated in this paper are preloaded and can be freely analyzed by the CorTo coexpression tool.

Despite its obvious advantages, the unexpected relative underperformance of RNA-seq vs. microarrays in network reconstruction raises an important caveat on its direct usability for coexpression analysis, at least by the simple Pearson correlation criteria used in this work. The creation of novel approaches to properly normalize and interprete gene count correlations generated by Next Generation Sequencing will pose a future fundamental challenge for coexpression investigators.

ACKNOWLEDGEMENTS

We thank Björn Usadel and Michele Morgante for the fruitful discussions. We also thank Aurora Maurizio, Elisa Del Fabbro, Lupo Giorgi and Maria Julieta D'Onofrio for their precious assistance.

Funding: This work was partially supported by Epigenomics Flagship Project (Progetto Bandiera Epigenomica), MIUR-CNR, the Institute of Applied Genomics and Scuola Superiore Sant'Anna.

Conflict of interest: none declared.

REFERENCES

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

Balakrishnan, C. N., Lin, Y.-C., London, S. E., and Clayton, D. F. (2012). Rna-seq transcriptome analysis of male and female zebra finch cell lines. *Genomics*.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A., and Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, **16**(5), 412–424.

Barabási, A.-L. and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nat Rev Genet*, **5**(2), 101–113.

Bassel, G. W., Lan, H., Glaab, E., Gibbs, D. J., Gerjets, T., Krasnogor, N., Bonner, A. J., Holdsworth, M. J., and Provart, N. J. (2011). Genome-wide network model capturing seed germination reveals coordinated regulation of plant cellular phase transitions. *Proc Natl Acad Sci U S A*, **108**(23), 9709–9714.

Basso, K., Margolin, A. A., Stolovitzky, G., Klein, U., Dalla-Favera, R., and Califano, A. (2005). Reverse engineering of regulatory networks in human b cells. *Nat Genet*, **37**(4), 382–390.

Beck, J. R. and Shultz, E. K. (1986). The use of relative operating characteristic (roc) curves in test performance evaluation. *Arch Pathol Lab Med*, **110**(1), 13–20.

Brandüao, M. M., Dantas, L. L., and Silva-Filho, M. C. (2009). Atpin: Arabidopsis thaliana protein interaction network. *BMC Bioinformatics*, **10**, 454.

Breitkreutz, B.-J., Stark, C., Reguly, T., Boucher, L., Breitkreutz, A., Livstone, M., Oughtred, R., Lackner, D. H., Bähler, J., Wood, V., Dolinski, K., and Tyers, M. (2008). The biogrid interaction database: 2008 update. *Nucleic Acids Res*, **36**(Database issue), D637–D640.

Brohée, S., Faust, K., Lima-Mendez, G., Vanderstocken, G., and van Helden, J. (2008). Network analysis tools: from biological networks to clusters and pathways. *Nat Protoc*, **3**(10), 1616–1629.

Caspi, R., Altman, T., Dreher, K., Fulcher, C. A., Subhraveti, P., Keseler, I. M., Kothari, A., Krummenacker, M., Latendresse, M., Mueller, L. A., Ong, Q., Paley, S., Pujar, A., Shearer, A. G., Travers, M., Weerasinghe, D., Zhang, P., and Karp, P. D. (2012). The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic Acids Res*, **40**(Database issue), D742–D753.

Chalker, A. F. and Lunsford, R. D. (2002). Rational identification of new antibacterial drug targets that are essential for viability using a genomics-based approach. *Pharmacol Ther*, **95**(1), 1–20.

Cole, S. T. (2002). Comparative mycobacterial genomics as a tool for drug target and antigen discovery. *Eur Respir J Suppl*, **36**, 78s–86s.

Dai, M., Wang, P., Boyd, A. D., Kostov, G., Athey, B., Jones, E. G., Bunney, W. E., Myers, R. M., Speed, T. P., Akil, H., Watson, S. J., and Meng, F. (2005). Evolving gene/transcript definitions significantly alter the interpretation of genechip data. *Nucleic Acids Res*, **33**(20), e175.

Daub, C. O., Steuer, R., Selbig, J., and 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., and Mendes, P. (2004). Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics*, **20**(18), 3565–3574.

D'haeseleer, P., Liang, S., and Somogyi, R. (2000). Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics*, **16**(8), 707–726.

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic Acids Res*, **30**(1), 207–210.

Farber, C. R. and Lusis, A. J. (2008). Integrating global gene expression analysis and genetics. *Adv Genet*, **60**, 571–601.

Fisher, R. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, **10**(4), 507–521.

Fu, F.-F. and Xue, H.-W. (2010). Coexpression analysis identifies rice starch regulator1, a rice ap2/erebp family transcription factor, as a novel rice starch biosynthesis regulator. *Plant Physiol*, **154**(2), 927–938.

Giorgi, F. M., Bolger, A. M., Lohse, M., and Usadel, B. (2010). Algorithm-driven artifacts in median polish summarization of microarray data. *BMC Bioinformatics*, **11**, 553.

Giuntoli, B., Licausi, F., Parlanti, S., W., D.-L., Weiste, C., and Perata, P. (2009). Sphavata, a ja-induced ap2/erf transcription factor of arabidopsis thaliana. In *20th International Conference on Arabidopsis Research*.

Hartemink, A. J. (2005). Reverse engineering gene regulatory networks. *Nat Biotechnol*, **23**(5), 554–555.

Haughn, G. W. and Western, T. L. (2012). Arabidopsis seed coat mucilage is a specialized cell wall that can be used as a model for genetic analysis of plant cell wall structure and function. *Front Plant Sci*, **3**, 64.

Hubbell, E., Liu, W.-M., and Mei, R. (2002). Robust estimators for expression analysis. *Bioinformatics*, **18**(12), 1585–1592.

Iancu, O. D., Kawane, S., Bottomly, D., Searles, R., Hitzemann, R., and McWeeney, S. (2012). Utilizing rna-seq data for de novo coexpression network inference. *Bioinformatics*, **28**(12), 1592–1597.

Jeong, H., Mason, S. P., Barabási, A. L., and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, **411**(6833), 41–42.

Jordan, I., Mariño-Ramírez, L., Wolf, Y., and Koonin, E. (2004). Conservation and coevolution in the scale-free human gene coexpression network. *Molecular biology and evolution*, **21**(11), 2058–2070.

Joy, M. P., Brock, A., Ingber, D. E., and Huang, S. (2005). High-betweenness proteins in the yeast protein interaction network. *J Biomed Biotechnol*, **2005**(2), 96–103.

Kitano, H. (2002). Systems biology: a brief overview. *Science*, **295**(5560), 1662–1664.

Klie, S. and Nikoloski, Z. (2012). The choice between mapman and gene ontology for automated gene function prediction in plant science. *Front Genet*, **3**, 115.

Koschützki, D. and Schreiber, F. (2008). Centrality analysis methods for biological networks and their application to gene regulatory networks. *Gene Regul Syst Bio*, **2**, 193–201.

Leinonen, R., Sugawara, H., Shumway, M., and Collaboration, I. N. S. D. (2011). The sequence read archive. *Nucleic Acids Res*, **39**(Database issue), D19–D21.

Li, M., Wang, J., Chen, X., Wang, H., and Pan, Y. (2011). A local average connectivity-based method for identifying essential proteins from the network level. *Comput Biol Chem*, **35**(3), 143–150.

Lim, W. K., Wang, K., Lefebvre, C., and Califano, A. (2007). Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks. *Bioinformatics*, **23**(13), i282–i288.

Lohse, M., Nunes-Nesi, A., Krüger, P., Nagel, A., Hannemann, J., Giorgi, F. M., Childs, L., Osorio, S., Walther, D., Selbig, J., Sreenivasulu, N., Stitt, M., Fernie, A. R., and Usadel, B. (2010). Robin: an intuitive wizard application for r-based expression microarray quality assessment and analysis. *Plant Physiol*, **153**(2), 642–651.

Luo, F., Yang, Y., Zhong, J., Gao, H., Khan, L., Thompson, D., and Zhou, J. (2007). Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory. *BMC bioinformatics*, **8**(1), 299.

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R. D., and Califano, A. (2006). Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, **7 Suppl 1**, S7.

Meng, M., Geisler, M., Johansson, H., Harholt, J., Scheller, H., Mellerowicz, E., and Kleczkowski, L. (2009). Udp-glucose pyrophosphorylase is not rate limiting, but is essential in arabidopsis. *Plant and cell physiology*, **50**(5), 998–1011.

Mishra, N. S., Tuteja, R., and Tuteja, N. (2006). Signaling through map kinase networks in plants. *Arch Biochem Biophys*, **452**(1), 55–68.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by rna-seq. *Nat Methods*, **5**(7), 621–628.

Movahedi, S., Bel, M. V., Heyndrickx, K. S., and Vandepoele, K. (2012). Comparative co-expression analysis in plant biology. *Plant Cell Environ*, **35**(10), 1787–1798.

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

Obayashi, T., Okamura, Y., Ito, S., Tadaka, S., Motoike, I., and Kinoshita, K. (2012). Coxpresdb: a database of comparative gene coexpression networks of eleven species for mammals. *Nucleic Acids Research*.

Opgen-Rhein, R. and Strimmer, K. (2007). From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. *BMC Syst Biol*, **1**, 37.

Persson, S., Wei, H., Milne, J., Page, G. P., and Somerville, C. R. (2005). Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets. *Proc Natl Acad Sci U S A*, **102**(24), 8633–8638.

Rambaldi, D., Giorgi, F. M., Capuani, F., Ciliberto, A., and Ciccarelli, F. D. (2008). Low duplicability and network fragility of cancer genes. *Trends Genet*, **24**(9), 427–430.

Reverter, A. and Chan, E. K. F. (2008). Combining partial correlation and an information theory approach to the reversed engineering of gene co-expression networks. *Bioinformatics*, **24**(21), 2491–2497.

Richard, H., Schulz, M. H., Sultan, M., Nürnberger, A., Schrinner, S., Balzereit, D., Dagand, E., Rasche, A., Lehrach, H., Vingron, M., Haas, S. A., and Yaspo, M.-L. (2010). Prediction of alternative isoforms from exon expression levels in rna-seq experiments. *Nucleic Acids Res*, **38**(10), e112.

Roberts, A., Pimentel, H., Trapnell, C., and Pachter, L. (2011). Identification of novel transcripts in annotated genomes using rna-seq. *Bioinformatics*, **27**(17), 2325–2329.

Ryngajllo, M., Childs, L., Lohse, M., Giorgi, F., Lude, A., Selbig, J., and Usadel, B. (2011). Slocx: predicting subcellular localization of arabidopsis proteins leveraging gene expression data. *Frontiers in plant science*, **2**.

Schäfer, J., Opgen-Rhein, R., and Strimmer, K. (2001). Reverse engineering genetic networks using the genenet package. *Journal of the American Statistical Association*, **96**, 1151–1160.

Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L., and Ideker, T. (2011). Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, **27**(3), 431–432.

Sterling, J. D., Atmodjo, M. A., Inwood, S. E., Kolli, V. S. K., Quigley, H. F., Hahn, M. G., and Mohnen, D. (2006). Functional identification of an arabidopsis pectin biosynthetic homogalacturonan galacturonosyltransferase. *Proc Natl Acad Sci U S A*, **103**(13), 5236–5241.

Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T. Z., Garcia-Hernandez, M., Foerster, H., Li, D., Meyer, T., Muller, R., Ploetz, L., Radenbaugh, A., Singh, S., Swing, V., Tissier, C., Zhang, P., and Huala, E. (2008). The arabidopsis information resource (tair): gene structure and function annotation. *Nucleic Acids Res*, **36**(Database issue), D1009–D1014.

Teichmann, S. A. and Babu, M. M. (2002). Conservation of gene co-regulation in prokaryotes and eukaryotes. *Trends Biotechnol*, **20**(10), 407–10; discussion 410.

Thimm, O., Bläsing, O., Gibon, Y., Nagel, A., Meyer, S., Krüger, P., Selbig, J., Müller, L. A., Rhee, S. Y., and Stitt, M. (2004). Mapman: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J*, **37**(6), 914–939.

Trapnell, C., Pachter, L., and Salzberg, S. L. (2009). Tophat: discovering splice junctions with rna-seq. *Bioinformatics*, **25**(9), 1105–1111.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2010). Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*, **28**(5), 511–515.

Tzafrir, I., Dickerman, A., Brazhnik, O., Nguyen, Q., McElver, J., Frye, C., Patton, D., and Meinke, D. (2003). The arabidopsis seedgenes project. *Nucleic Acids Res*, **31**(1), 90–93.

Upton, G. J. G. (1992). Fisher's exact test. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, **155**, 395–402.

Usadel, B., Kuschinsky, A. M., Rosso, M. G., Eckermann, N., and Pauly, M. (2004). Rhm2 is involved in mucilage pectin synthesis and is required for the development of the seed coat in arabidopsis. *Plant Physiol*, **134**(1), 286–295.

Usadel, B., Obayashi, T., Mutwil, M., Giorgi, F. M., Bassel, G. W., Tanimoto, M., Chow, A., Steinhauser, D., Persson, S., and Provart, N. J. (2009). Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. *Plant Cell Environ*, **32**(12), 1633–1651.

Vandepoele, K., Quimbaya, M., Casneuf, T., De Veylder, L., and Van de Peer, Y. (2009). Unraveling transcriptional control in arabidopsis using cis-regulatory elements and coexpression networks. *Plant physiology*, **150**(2), 535–546.

Vasilevski, A., Giorgi, F. M., Bertinetti, L., and Usadel, B. (2012). Lasso modeling of the arabidopsis thaliana seed/seedling transcriptome: a model case for detection of novel mucilage and pectin metabolism genes. *Mol Biosyst*, **8**(10), 2566–2574.

Wang, S., Yin, Y., Ma, Q., Tang, X., Hao, D., and Xu, Y. (2012). Genome-scale identification of cell-wall related genes in arabidopsis based on co-expression network analysis. *BMC Plant Biol*, **12**(1), 138.

Wang, Z., Gerstein, M., and Snyder, M. (2009). Rna-seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, **10**(1), 57–63.

Wolfe, C. J., Kohane, I. S., and Butte, A. J. (2005). Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. *BMC Bioinformatics*, **6**, 227.

Wuchty, S. (2002). Interaction and domain networks of yeast. *Proteomics*, **2**(12), 1715–1723.

Yamada, T. and Bork, P. (2009). Evolution of biomolecular networks: lessons from metabolic and protein interactions. *Nat Rev Mol Cell Biol*, **10**(11), 791–803.

Yilmaz, A., Mejia-Guerra, M., Kurz, K., Liang, X., Welch, L., and Grotewold, E. (2011). Agris: the arabidopsis gene regulatory information server, an update. *Nucleic acids research*, **39**(suppl 1), D1118–D1122.

Zampieri, M., Soranzo, N., and Altafini, C. (2008). Discerning static and causal interactions in genome-wide reverse engineering problems. *Bioinformatics*, **24**(13), 1510–1515.

Zhang, K., Li, J. B., Gao, Y., Egli, D., Xie, B., Deng, J., Li, Z., Lee, J.-H., Aach, J., Leproust, E. M., Eggan, K., and Church, G. M. (2009). Digital rna allelotyping reveals tissue-specific and allele-specific gene expression in human. *Nat Methods*, **6**(8), 613–618.