



UNIVERSITÀ  
DEGLI STUDI  
DI UDINE

## Università degli studi di Udine

### Gene co-expression analyses: an overview from microarray collections in *Arabidopsis thaliana*

*Original*

*Availability:*

This version is available <http://hdl.handle.net/11390/1104454> since 2018-01-09T16:28:52Z

*Publisher:*

*Published*

DOI:10.1093/bib/bbw002

*Terms of use:*

The institutional repository of the University of Udine (<http://air.uniud.it>) is provided by ARIC services. The aim is to enable open access to all the world.

*Publisher copyright*

(Article begins on next page)



# Gene co-expression analyses: an overview from microarray collections in *Arabidopsis thaliana*

Pasquale Di Salle\*, Guido Incerti\*, Chiara Colantuono and Maria Luisa Chiusano

Corresponding author. Maria Luisa Chiusano. Department of Agricultural Sciences, University of Naples "Federico II", via Università 100, 80055 Portici (NA), Italy. Tel.: +39 081 2539492. E-mail: [chiusano@unina.it](mailto:chiusano@unina.it)

\*These authors contributed equally to this work.

## Abstract

Bioinformatics web-based resources and databases are precious references for most biological laboratories worldwide. However, the quality and reliability of the information they provide depends on them being used in an appropriate way that takes into account their specific features. Huge collections of gene expression data are currently publicly available, ready to support the understanding of gene and genome functionalities. In this context, tools and resources for gene co-expression analyses have flourished to exploit the 'guilty by association' principle, which assumes that genes with correlated expression profiles are functionally related. In the case of *Arabidopsis thaliana*, the reference species in plant biology, the resources available mainly consist of microarray results. After a general overview of such resources, we tested and compared the results they offer for gene co-expression analysis. We also discuss the effect on the results when using different data sets, as well as different data normalization approaches and parameter settings, which often consider different metrics for establishing co-expression. A dedicated example analysis of different gene pools, implemented by including/excluding mutant samples in a reference data set, showed significant variation of gene co-expression occurrence, magnitude and direction. We conclude that, as the heterogeneity of the resources and methods may produce different results for the same query genes, the exploration of more than one of the available resources is strongly recommended. The aim of this article is to show how best to integrate data sources and/or merge outputs to achieve robust analyses and reliable interpretations, thereby making use of diverse data resources an opportunity for added value.

**Key words:** gene correlation; biological database; data heterogeneity; bioinformatics; CESA7; AT5G12080

## Background

Thirty years after the first implementation with antibodies, 20 after the first application with DNA and since the first genome-wide report, microarray technology still remains one of the least expensive and most powerful approaches used to explore the transcriptional landscape of a biological sample, whether it is represented by a tissue, a group of cells or a mixture, in physiological, stress or pathological conditions [1–4]. Though with

some well-known technical limits, their use enables the detection of differentially expressed genes from comparative experiments and the description of expression patterns in different tissues/conditions, or in time course experiments [2, 5]. The variability of the expression of a multitude of genes from a genome can be traced using this technology [6]. Moreover, as defined by the guilty by association (GbA) principle, genes sharing the same expression patterns in several experiments may be studied as candidates involved in the same functional

Pasquale Di Salle was a PhD in Computational Biology and Bioinformatics in the Department of Agraria, University of Naples 'Federico II', Italy.

Guido Incerti is a contract professor in Biostatistics at University of Trieste and postdoc researcher in Biostatistics at Department of Agraria, University of Naples 'Federico II', Italy.

Chiara Colantuono is a postdoc researcher in Computational Biology and Bioinformatics in the Department of Agraria, University of Naples 'Federico II', Italy.

Maria Luisa Chiusano is a permanent staff scientist and assistant professor in Molecular Biology and Bioinformatics in the Department of Agraria, University of Naples 'Federico II', Italy.

Submitted: 2 October 2015; Received (in revised form): 28 December 2015

© The Author 2016. Published by Oxford University Press. For Permissions, please email: [journals.permissions@oup.com](mailto:journals.permissions@oup.com)

network [7, 8]. Gene expression profile analysis is therefore used for the detection of co-expressed genes, i.e. genes with either positively or negatively correlated profiles [9]. This type of analysis requires the exploitation of a sufficiently high number of homogenous experiments, to ensure acceptable levels of statistical power. However, experimental homogeneity, in terms of the biological and technical aspects of the experimental design, is required to permit different experiments to be compared and minimize the occurrence of potential biases and confounding effects. Despite the increasing accessibility of microarray technology and the availability of gene expression profiles for many different biological species and samples, the different collections need to be appropriately selected for assessing gene co-expression. This is mainly owing to the extensive heterogeneity of the experimental efforts, often involving different genotypes or cultivars; the fast evolution of technologies, usually offered from different companies with different platforms; and the availability of different approaches for data analysis. The analysis of gene co-expression requires a suitable design [10]. As an example, gene co-expression under stress or pathological conditions should be assessed on enough samples within the same context and, possibly, be compared with an adequate set of controls in physiological conditions. It is therefore paradoxical that, although microarray-based transcriptome analyses are widespread, few collections, even from reference species, may be useful for dedicated co-expression analyses. Moreover, even fewer collections have been found to be representative for comparable approaches, often rendering the data not useful for more advanced analyses [11–14]. However, the interest in co-expression analysis is increasing and the experimental designs are becoming more suitable for providing source data useful to this end. The availability of alternative high-throughput transcriptome approaches, such as RNA-seq, is growing and further enhancing the feasibility of gene co-expression analyses in molecular biology [15–17].

Because *Arabidopsis thaliana* is well established as a reference in plant biology, with a genome completely sequenced since the year 2000, its transcriptome has been extensively studied using different microarray technologies [18–21], offering useful collections for gene co-expression analyses. Indeed, an overwhelming number of experiments comes from the Affymetrix platform (ATH1). All the related results were collected at the NASCArrays Web site. Though the direct web-based service for data distribution on the NASC's International Affymetrix Web site was closed in 2013, all the data are still accessible from the Gene Expression Omnibus repository of the US National Center for Biotechnology Information [22]. Being the reference for all the public Affymetrix ATH1 and AG 'GeneChip' microarrays for *A. thaliana*, the NASCArrays collection stores hundreds of different experiments and thousands of slides. All data are described following the MIAME guidelines, and the metadata include information on the sample and on the approaches exploited for hybridization, scanning and normalization, this latter generally based on the MAS5.0 protocol [23–25].

The availability of this and other *Arabidopsis* microarray collections has encouraged the development of web-based dedicated resources to collect the data and to offer tools for co-expression analyses (e.g. [26, 27]). Among these, we here consider AtCOECiS [28], ATTED II [29], BAR [30], CoP [31], CORNET [32], Cress Express [33], CSB.DB [34], GeneCAT [35], GeneMania [36], Genevestigator [37], PlaNet [38]. They show similarities and differences related to gene co-expression analysis. Moreover, they assess the effect of particular features of these resources on the co-expression results and identify important issues that users should consider when approaching similar analyses. Then, by

using a comparative meta-analysis of the results attainable using different resources queried for co-expression results with different genes, we provide an example of the variability of co-expression analysis results. Because almost all platforms considered in this work contain many different microarray experiments, we also tested, as an example, the hypothesis that experimental heterogeneity affects the stability of gene co-expression results. To this end, as we assessed the different potential sources of heterogeneity, we also considered the inclusion/exclusion of mutant-based experiments with a reference data set from physiological conditions, and tested the stability of co-expression occurrence, magnitude and direction in selected gene pools.

## Methods

### Reference data set for gene co-expression assessment

A reference collection of experiments from the *A. thaliana* ATH1 Affymetrix chipset, designed with 22 810 probes was organized as a framework for our analyses. The gene expression values from 63 experiments on samples including several tissues and organs in physiological conditions, and repeated in triplicate, for a total number of 189 microarray slides (Additional Table 1 in the Supplementary Material) of AtGenExpress [39] developmental series, were downloaded from [http://affymetrix.arabidopsis.info/link\\_to\\_ipiant.shtml](http://affymetrix.arabidopsis.info/link_to_ipiant.shtml), on January 2013. In addition, gene expression values from 16 experiments involving mutants were considered (Additional Table 1 in the Supplementary Material), bringing the final number of downloaded experiments to 79, with a total of 237 slides. In the following analysis, the full data set including that form mutant genotypes is referred to as the mut+ set, and that without mutant genotypes, the mut–set. Microarray results were normalized using the MAS 5.0 protocol (for each experiment the highest and lowest 2% of each signal were removed, and then all values were transformed to an average of 100). Only 21 769 probes had signals; from these, we removed 802 ambiguous probes (i.e. shared between genes, mapping gene families or noncoding genes). Moreover, we filtered out all the probes with an expression level under the 5th percentile in each sample of all the experiments, bringing the final number of gene-specific probes to 20 957. Because all experiments were repeated in triplicate, the signals of each gene-specific probe in each experiment were calculated as the average of the three replicates. Finally, a  $\log_2$  transformation was applied to all the signals.

Co-expression between any two *X* and *Y* genes was established after calculating the Pearson product-moment correlation coefficient (Equation (1)), considering a threshold value of  $|r| \geq 0.7$ , commonly used in gene co-expression networks (e.g. [40]):

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

where  $X_i$  and  $\bar{X}$  are the gene expression value in the *i*-th experiment and the average of the expression values in all the experiments for the gene *X*, respectively, while  $Y_i$  and  $\bar{Y}$  are the same for the gene *Y*.

### Testing the effects of sample heterogeneity on gene co-expression assessment

To evaluate the effects of sample heterogeneity on the stability of co-expression analysis results, we used two pair-wise

correlation matrices calculated for the 20957 gene-specific probes, either on the full data set of 79 experiments (i.e. the mut+ data set) and on the smaller data set of 63 experiments with nonmutant samples (i.e. the mut- data set). The correlation matrices were calculated using an in-house method. Because the inclusion of mutant data could lead to different co-expression outcomes for different genes, we purposely considered two different pools of genes in the following analyses. First, to selectively assess the effect of mutant inclusion on the most co-expressed genes, we selected the topmost 1% of gene-specific probes ( $N=200$ ) showing the highest number of co-expression occurrences in the mut+ data set. Second, to study the effect on the most unstably co-expressed genes, we selected the topmost 1% of gene probes showing the highest co-expression variation owing to mutant inclusion. In addition, for each pair of genes we calculated the correlation in the mut+ and mut- data sets (i.e.  $r_{\text{mut}+}$  and  $r_{\text{mut}-}$ , respectively) and their absolute difference (i.e.  $d = |r_{\text{mut}+} - r_{\text{mut}-}|$ ). Then, we ranked the gene pairs by decreasing  $d$ . Finally, unstable genes ( $N=200$ ) were identified from the gene pair ranking as those showing the lowest sum of ranks (i.e. the highest occurrence of large variations in correlation scores).

In a preliminary analysis of the most co-expressed and unstably co-expressed gene pools, we extensively tested the existence of a relationship between the frequency of gene co-expression and presence/absence of mutants in the data set. We used the Chi-square test for independence on  $2 \times 2$  contingency tables, for each gene separately and for all genes pooled, reporting the observed occurrences of either co-expressed ( $|r| \geq 0.7$ ) or non-co-expressed ( $|r| < 0.7$ ) genes, for either mut+ or mut- data sets (a total of 398 pair-wise comparisons for each tested gene, 199 in each data set). A significant Chi-square statistic indicated the dependence of the observed gene co-expression pattern on the inclusion or exclusion of mutants in the reference data set.

In a more specific approach, we assessed the effect of mutant inclusion on the correlation between each pair of genes, testing the significance of the difference between the correlation calculated with and without mutants (previously defined as  $d$ ) by calculating the probability  $P(r_{\text{mut}+} = r_{\text{mut}-})$ . We assumed that the distribution of correlations follows a t-distribution with  $n - 2$  degrees of freedom, when considering the null hypothesis of no correlation, with  $n$  number of experiments in the data set. Therefore, we used a two-tailed  $t$ -test on the Fisher's  $z$  transformed values of  $r_{\text{mut}+}$  and  $r_{\text{mut}-}$  according to [41]. We controlled for multiple comparisons using Bonferroni correction for the number of genes tested, such that the family-wise error rate was controlled at  $\alpha = 0.05/200 = 0.00025$  per gene pool. Then,  $P$ -values in the range of 0.00025 to 0.05 were considered only marginally significant. For a given gene pair, a significant value of  $d$  indicates that gene co-expression significantly changed after mutant inclusion in the data set. Occurrences of significant values of  $d$  within each data set were calculated both separately for each tested gene and for all pooled genes. In the case of gene pairs showing significant values of  $d$ , the type of mutant-dependent effect on co-expression change and the relevance of its occurrences in the gene pools were assessed. Types of effects were defined based on the values of  $r_{\text{mut}+}$  and  $r_{\text{mut}-}$  as follows: (i) gene co-expression inhibition (from statistically significant  $r_{\text{mut}-}$  to not significant  $r_{\text{mut}+}$ ), (ii) induction (from not statistically significant  $r_{\text{mut}-}$  to significant  $r_{\text{mut}+}$ ), (iii) inversion (from positive to negative correlation or vice versa) and (iv) changes of magnitude not affecting  $r$  sign and significance. For each type of effect, the mean and the 95% confidence interval of occurrences

within the gene pairs tested for each gene ( $N=199$ ) were calculated across all genes of each pool. To assess the relevance of the effects in the gene pools, we tested for significant deviations from zero using  $t$ -tests for single sample means. To evaluate the relative prevalence of different types of effect, occurrences were expressed as percentage of the total number of gene pairs significantly affected by all types of effect. Finally, for all tested gene pairs ( $2$  gene pools  $\times 200!/(200 - 2)!$  gene pairs per pools, corresponding to a total number of 39800 gene pairs), the co-expression variability owing to mutant effects was visually assessed by a scatterplot of Pearson correlations observed in the mut+ versus mut- data sets.

### Public co-expression analysis resources

Eleven web-based and publicly available resources, offering databases and facilities for gene co-expression analysis in *A. thaliana*, were consulted between February and June 2014 (Table 1).

All information was directly retrieved from the platform Web sites, and from reference publications indicated therein. The resources were surveyed for general functionality, including main features, tools and data sources. The total number of experiments, slides, treatments and conditions, the data normalization method, metrics and tools for gene co-expression analyses are also described in Additional Methods in the Supplementary Material.

### Meta-analysis of queries from different platforms

We investigated the performance of the surveyed web-based resources when queried using four different sample genes as probes, here indicated using the Arabidopsis Genome Initiative code. We selected the gene AT5G17420 (CESA7), which is well known to be co-expressed with AT5G44030 and AT4G18780 (genes CESA4 and CESA8, respectively) under physiological conditions, as the three different genes are components of a complex involved in the cell wall synthesis [18, 42]. The gene AT5G06680, implied in the gamma-tubulin complex, has been considered because it showed variability in co-expressed genes when analyzed within the mut- and mut+ collections (data not shown). AT1G20580, a small nuclear ribonucleoprotein, and AT1G01290, a cofactor of nitrate reductase and xanthine dehydrogenase, were selected among the most co-expressed and unstable gene pools from the mut- and mut+ collections. For all the four query genes, the analyses were performed on each platform (Table 1). When possible we selected the Pearson coefficient ( $r$ ) or the coefficient of determination ( $r^2$ ), to base the analysis on comparable metrics. The default metrics were used in other cases (see Additional Methods in the Supplementary Material). The default normalization methods were used in each platform. In the case of Cress Express, we performed the analysis changing the normalization methods among the three available. For each query, we recorded the topmost 20 co-expressed genes, resulting from each platform. The resulting genes were ranked according to the value of the specific metric proposed by each platform, when available.

## Results and discussion

### Effects of sample heterogeneity by mutant inclusion on gene co-expression profiles

We tested the hypothesis that sample heterogeneity in the reference data set could affect the stability of gene co-expression

**Table 1.** Web-based platforms offering facilities for gene co-expression analysis on *Arabidopsis*

| Platform       | Web site  | Release | Slides | Normalization       | Metrics         |
|----------------|---|---------|--------|---------------------|-----------------|
| AtCOECiS       | <a href="http://bioinformatics.psb.ugent.be/ATCOECiS">http://bioinformatics.psb.ugent.be/ATCOECiS</a>                   | 2009    | 322    | RMA                 | $r$             |
| ATTED II       | <a href="http://atted.jp/">http://atted.jp/</a>   | 2007    | 11 171 | RMA                 | MR, $r$         |
| BAR            | <a href="http://bar.utoronto.ca/welcome.htm">http://bar.utoronto.ca/welcome.htm</a>                                     | 2005    | NA     | MAS 5.0             | $r$             |
| CoP            | <a href="http://webs2.kazusa.or.jp/kagiana/cop0911/">http://webs2.kazusa.or.jp/kagiana/cop0911/</a>                     | 2010    | 5272   | MAS 5.0             | CC, VF          |
| CORNET         | <a href="https://cornet.psb.ugent.be/">https://cornet.psb.ugent.be/</a>   | 2009    | NA     | RMA                 | $r, \rho$       |
| Cress Express  | <a href="http://cressexpress.org">http://cressexpress.org</a>   | 2008    | 1799   | RMA, MAS 5.0, GCRMA | $r^2$ , slope   |
| CSB.DB         | <a href="http://csbdb.mpimp-golm.mpg.de/csbdb/dbcors/ath.html">http://csbdb.mpimp-golm.mpg.de/csbdb/dbcors/ath.html</a> | 2004    | NA     | MAS 5.0 (GCOS)      | $r, \rho, \tau$ |
| GeneCAT        | <a href="http://genecat.mpg.de/cgi-bin/Ainitiator.py">http://genecat.mpg.de/cgi-bin/Ainitiator.py</a>                   | 2008    | 351    | RMA                 | $r$             |
| GeneMania      | <a href="http://www.genemania.org/">http://www.genemania.org/</a>   | 2008    | NA     | NA                  | W               |
| Genevestigator | <a href="https://www.genevestigator.com/gv/">https://www.genevestigator.com/gv/</a>                                     | 2004    | 9848   | RMA, MAS 5.0        | $r$             |
| PlaNet         | <a href="http://aranet.mpimp-golm.mpg.de/">http://aranet.mpimp-golm.mpg.de/</a>   | 2011    | 1074   | NA                  | $r$             |

Note. For each platform, web address, year of first release, total number of slides, methods used for signal normalization and metrics used to establish gene co-expression are reported. For further details of databases and microarray collections, as well as parameter settings used for querying each platform, see Additional Methods in the [Supplementary Material](#).

Metric symbols indicate: mutual ranking (MR), Pearson's product-moment correlation coefficient ( $r$ ) and associated  $P$ -value, cosine correlation (CC), vertex F-measure (VF), coefficient of determination ( $r^2$ ), slope of least squares best fitting regression line (slope), Spearman's rank correlation coefficient ( $\rho$ ), Kendall's rank correlation coefficient ( $\tau$ ) and weight of the relation (W). NA = information not available.

analysis. To this aim, among different factors possibly producing sample heterogeneity, we considered the presence/absence of mutant samples in the data set. Then, after calculating pair-wise correlations among all gene probes, either including or excluding data from mutant samples, we selected two gene pools for further analysis. These corresponded to either the most co-expressed or the most unstably co-expressed genes (i.e. genes showing the highest co-expression variation owing to mutant inclusion/exclusion).

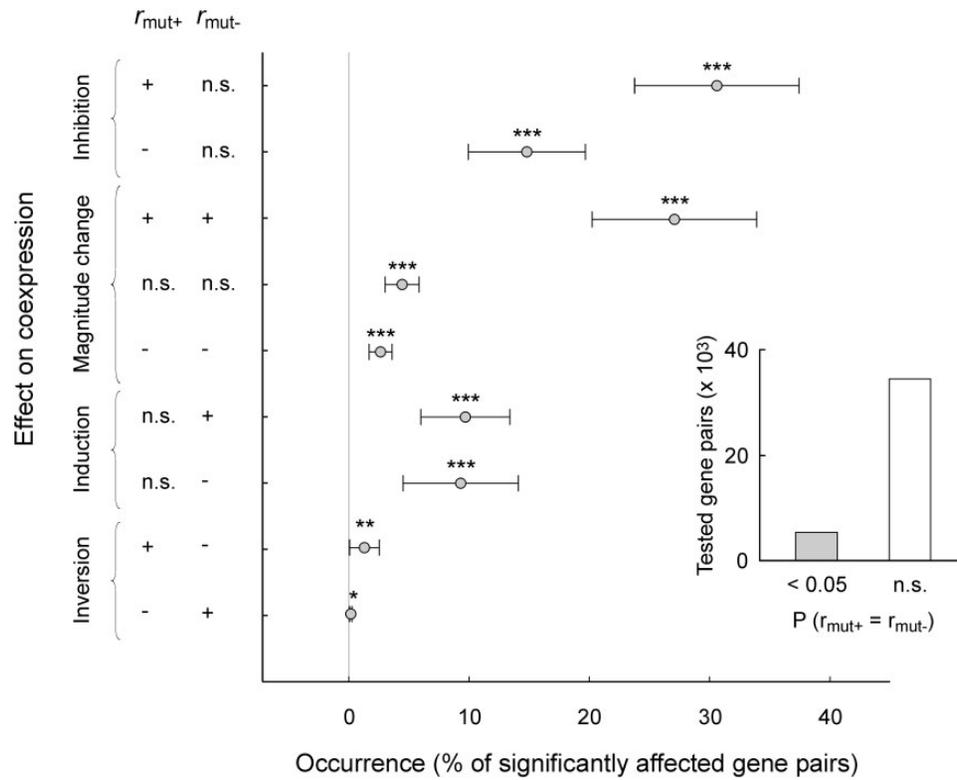
In a first analysis, for each single gene of both pools, we tested the independence of co-expression occurrences from the presence/absence of mutants in the reference data set, using the independence Chi-square test. The analysis showed a clear pattern of interdependence between the two variables, with mutant inclusion affecting the co-expression pattern of most genes in both pools, and for data pooled for all genes (Additional Tables 2 and 3 in the [Supplementary Material](#)). In particular, Chi-square tests for the topmost 200 co-expressed genes (Additional Table 2 in the [Supplementary Material](#)) were significant (at  $P < 0.00025$ ) for 136 genes, marginally significant ( $0.00025 < P < 0.05$ ) in 39 cases and not significant ( $P > 0.05$ ) in 19 cases. Six genes were always co-expressed with all other remaining 199 genes of the pool, either considering the mut+ or the mut- data sets and, therefore, the Chi-squared test did not apply. These results were similar to what was observed for the unstably co-expressed genes (Additional Table 3 in the [Supplementary Material](#)). Indeed, the genes from this pool showed 101 significant, 30 marginally significant and 31 nonsignificant cases of mutant presence/absence effect on co-expression occurrence. Twenty-five genes were never co-expressed with the other genes of the pool, whether considering the mut+ or the mut- data sets, and therefore, the Chi-squared test was not calculated.

In a more detailed approach, focused on the specific effect of sample heterogeneity on the magnitude and direction of gene co-expression, we assessed whether mutant inclusion/exclusion in the reference data set produced significant variation of pair-wise correlation, considering all pair-wise comparisons for all genes of the two pools. For each gene pair, we calculated the correlations from the data sets including or excluding mutants and tested the significance of the difference between the two resulting values ( $d$ ) using a two-tailed t-test on the Fisher's  $z$  transformed correlation values.

Considering unstably co-expressed genes (insert in [Figure 1](#)), 14% of the tested gene pairs (5300 out of 39800) showed a significant value of  $d$ , indicating that the presence of mutants affected the co-expression magnitude. In the cases of mutant-dependent significant effects on co-expression magnitude (i.e. significant  $d$  values), further investigations provided an insight on the type and relevance of this effect ([Figure 1](#)). For each pair of genes showing a significant  $d$  value, the type of effect was defined based on the values of pair-wise correlation observed either in the presence ( $r_{mut+}$ ) or in the absence ( $r_{mut-}$ ) of mutants. On the other hand, the relevance of each type of effect was assessed by testing its occurrence in the gene pool for significant deviations from zero, using t-tests for single sample mean.

All types of mutant-related effects were relevant, being observed with significant occurrence among the tested genes ([Figure 1](#)), though with important differences for different types of effect. In particular ([Figure 1](#), [Table 2](#)), inhibition of co-expression highly prevailed, with 1622 and 784 total cases of positive and negative correlations (i.e. 30.6% and 14.8% of all the significant observed effects) becoming not significant after exclusion of mutants from the data set. Significant changes of magnitude, not affecting co-expression direction, were also frequently observed, mostly in the case of positive correlations (1434 cases, 27.1% of all the significant effects), while for negative correlations such effect was still relevant, but more rarely observed (140 cases, i.e. 2.6% of all the significant effects). Induction of gene co-expression (i.e. nonsignificant correlations in presence of mutants that turn into significance, either positive or negative, after mutant exclusion) were relatively frequent (512 and 492 cases, corresponding to 2.3% and 2.2% of all the significant effects for positive and negative correlations). Co-expression inversion (i.e. positive correlation in presence of mutants turning into negative correlation after mutant exclusion, and vice versa) was also recorded, although rarely, with 70 (1.3% of all the significant effects) and 10 cases (0.2%), respectively.

When such analysis was performed for the most co-expressed genes, only 0.04% of the tested gene pairs (14 out of 39800) were significantly affected by mutant inclusion/exclusion in the data set ([Table 2](#)). No occurrence of co-expression inhibition, induction or inversion was recorded. All the 14 cases of significant effects resulted in changes of



**Figure 1.** Effect of mutants on pair-wise co-expression of unstable gene pools. Data in main graphs refer to occurrences (mean and 95% confidence interval) of different types of effect, expressed as percentage of the total number of significantly affected gene pairs. Asterisks above symbols refer to statistically significant occurrences (t-tests for single mean; significant deviation from zero indicated by the gray vertical line; \*\*\* $P < 0.001$ ; \*\* $P < 0.01$ ; \* $P < 0.05$ ). The insert shows overall occurrences of significant ( $P < 0.05$ ) and not significant (NS) co-expression changes owing to mutants (two-tailed t-tests on Fisher’s z transformed values of  $r_{mut+}$  and  $r_{mut-}$  for the null hypothesis that gene pair-wise co-expression is not affected by mutant exclusion or inclusion, that is,  $r_{mut+} = r_{mut-}$ ).

**Table 2.** Significance and occurrence of different mutant-related effects on gene co-expression

| Gene pool                   | Effect type      | $r$   |       | $d$ | Occurrences |      |        |             | t-test |        |
|-----------------------------|------------------|-------|-------|-----|-------------|------|--------|-------------|--------|--------|
|                             |                  | mut + | mut - |     | N           | %    | Mean   | 95% CI      | t      | P      |
| Most co-expressed genes     | Inhibition       | +     | NS    | *   | 0           | 0    | 0      | -           | 0      | 1      |
|                             |                  | -     | NS    | *   | 0           | 0    | 0      | -           | 0      | 1      |
|                             | Magnitude change | +     | +     | *   | 14          | 100  | 0.07   | 0-0.14      | 1.88   | 0.061  |
|                             |                  | NS    | NS    | *   | 0           | 0    | 0      | -           | 0      | 1      |
|                             |                  | -     | -     | *   | 0           | 0    | 0      | -           | 0      | 1      |
|                             | Induction        | NS    | +     | *   | 0           | 0    | 0      | -           | 0      | 1      |
|                             |                  | NS    | -     | *   | 0           | 0    | 0      | -           | 0      | 1      |
| Inversion                   | +                | -     | *     | 0   | 0           | 0    | -      | 0           | 1      |        |
|                             | -                | +     | *     | 0   | 0           | 0    | -      | 0           | 1      |        |
| Unstably co-expressed genes | Not significant  | All   | All   | NS  | 39 786      | -    | 198.93 | 198.9-199   | 5350   | <0.001 |
|                             | Inhibition       | +     | NS    | *   | 1622        | 30.6 | 8.11   | 6.3-9.92    | 8.84   | <0.005 |
|                             |                  | -     | NS    | *   | 784         | 14.8 | 3.92   | 2.63-5.21   | 6.01   | <0.004 |
|                             | Magnitude change | +     | +     | *   | 1434        | 27.1 | 7.17   | 5.36-8.98   | 7.80   | <0.003 |
|                             |                  | NS    | NS    | *   | 236         | 4.5  | 1.18   | 0.81-1.55   | 6.25   | <0.002 |
|                             |                  | -     | -     | *   | 140         | 2.6  | 0.7    | 0.45-0.95   | 5.48   | <0.001 |
|                             | Induction        | NS    | +     | *   | 512         | 9.7  | 2.56   | 1.58-3.54   | 5.16   | <0.001 |
|                             |                  | NS    | -     | *   | 492         | 9.3  | 2.46   | 1.19-3.73   | 3.82   | <0.001 |
|                             | Inversion        | +     | -     | *   | 70          | 1.3  | 0.35   | 0.02-0.68   | 2.10   | 0.037  |
|                             |                  | -     | +     | *   | 10          | 0.2  | 0.05   | 0.02-0.08   | 3.24   | 0.001  |
|                             | Not significant  | All   | All   | NS  | 34 500      | -    | 172.5  | 168.2-176.7 | 80.39  | <0.001 |

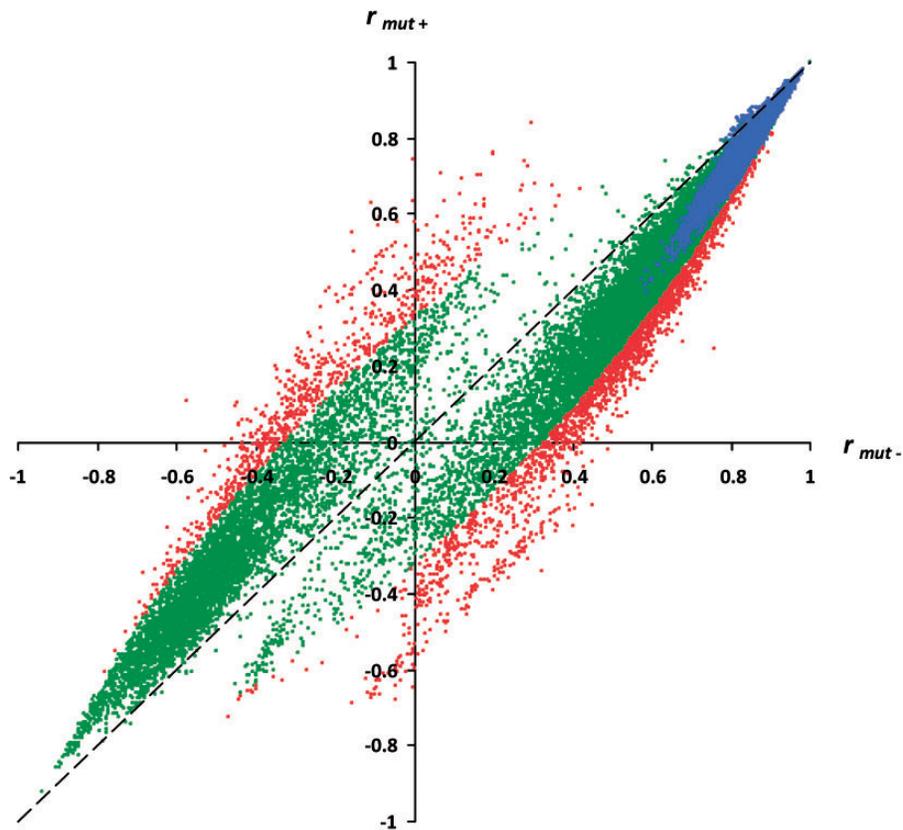
*Note.* For most co-expressed and unstably co-expressed gene pools, effect type depends on Pearson’s correlation scores (+, positive and significant; -, negative and significant; NS = not significant) as resulting from the reference data sets including (mut+) and excluding (mut-) experiments with mutants, and on the statistical significance of their difference (i.e.  $d = |r_{mut+} - r_{mut-}|$ ; \*, significant; NS = not significant). Occurrences for each gene pool are reported as total counts (N), percentage of total significant effects (%), mean and 95% confidence interval in the gene pairs tested for each gene (N = 199). Results of single sample t-tests on each mean for significant deviation from zero are also reported.

magnitude in gene co-expression, all corresponding to positive correlations both in presence and absence of mutants.

Our analysis of co-expression profiles, represented by 400 genes subdivided into two pools out of a total of 20957 gene-specific probes in the reference data set, showed a clear-cut effect from data set heterogeneity, here represented by mutant inclusion/exclusion, on co-expression profiles. Interestingly, such effects can be extremely variable for different gene pools, in terms of overall occurrence, magnitude and direction of co-expression changes. In particular, our assessment of *mut-* and *mut+* collections, exploited to assess a possible effect of data set heterogeneity on the results of co-expression analysis, showed a different pattern for the two gene pools considered in our analysis. Indeed, when pair-wise gene correlation values observed in the absence of mutants were compared against corresponding values calculated in presence of mutants, relevant differences between the most co-expressed and the unstably co-expressed gene pools were observed (Figure 2). In particular, the genes of the first pool shared high correlation values irrespective of the mutant inclusion/exclusion, indicating that the most co-expressed genes can be considered stably co-expressed because they are negligibly affected by the sample heterogeneity in the reference data set (Figure 2). Conversely, the pool of genes selected as the most unstably co-expressed ones clearly showed relevant deviations in their pair-wise correlation profiles related to the presence/absence of mutant samples in the

data set (Figure 2), hence indicating an overall relevant effect of sample heterogeneity on gene co-expression assessment.

In general, our first analysis based on Chi-squared testing of co-expression occurrences excluded the independence of co-expression from mutant presence/absence for most genes of both pools, apparently indicating the existence of a relationship in both cases (Additional Tables 2 and 3 in the Supplementary Material). However, the evaluation of significance, type and occurrence of this relationship highlighted many important differences between the two pools. In particular, the total occurrence of significant effects among unstable genes was >370-fold higher (i.e. 5300/14) than among the most co-expressed genes, which, as a consequence, can be defined as stably co-expressed genes. Interestingly, all the 14 gene pairs affected by mutant inclusion in the reference data set showed a gene in common, namely AT5G12080. This gene encodes a mechanically sensitive (or stretch-activated) ion channel in the plasma membrane with a moderate preference for anions, and has been reported as involved in anion transport, detection of mechanical stimulus, leaf senescence and programmed cell death in response to reactive oxygen species [43, 44]. Such functions could be related to the differential expression levels of AT5G12080 among plant organs and tissues. Indeed, the gene is significantly more expressed in shoots and stems, both in wild-type genotype and in mutant samples [one-way analysis of variance (ANOVA),  $F_{5,183} = 18.1$ ,  $P < 0.001$  and  $F_{2,45} = 299.8$ ,  $P < 0.001$ , respectively].



**Figure 2.** Choice of experimental conditions used for co-expression analysis differently affects co-expression scores of different gene pools. Data refer to Pearson's correlation coefficient scores ( $r$ ) for 39800 gene pairs, calculated either excluding or including experiments with mutants from the reference data set. Each point in the graph represents for a given gene pair the correlation calculated for the *mut-* data set on the x-axis and its corresponding value from the *mut+* data set on the y-axis. Blue and green dots indicate pairs of stable and unstable gene pairs, respectively (200 genes for each pool, 199 gene pairs for each gene); red dots indicate gene pairs significantly affected by the presence of mutants (i.e. significant  $d$  values, see Table 2). Points above the diagonal dotted line correspond to gene pairs with co-expression increased (or anti-co-expression decreased, for negative  $r$  scores) in presence of mutants (i.e.  $r_{mut+}$  higher than  $r_{mut-}$ ) and vice versa for points below the line. For details on the reference data set and gene pools definition see main text.

AT5G12080 was significantly more expressed in experiments with mutants compared with the mut- data set, both considering different plant organs separately (one-way ANOVA,  $F_{1,64} = 26.9$ ,  $P < 0.001$ ;  $F_{1,58} = 4.0$ ,  $P < 0.05$ ;  $F_{1,40} = 40.6$ ,  $P < 0.001$  for flowers, leaves and shoots and stems, respectively) and all pooled data (one-way ANOVA,  $F_{1,235} = 148.5$ ,  $P < 0.001$ ). Such results suggest potential interesting characteristics of AT5G12080. First, we have shown that, within the pool of the most co-expressed genes, the gene AT5G12080, responsible for all the significant variations of co-expression owing to mutants, is over-expressed in mutant samples. Second, AT5G12080 was included in the pools of most co-expressed genes, which also were found to be the most stably co-expressed genes in the presence of mutants. These observations suggest that the only deviations from a stable co-expression in presence of mutants could be owing to mutations directly involving AT5G12080 and/or pathways enhancing its expression level. In other words, AT5G12080 could be the most stably co-expressed among the mutated genes, and/or the only mutated among the most stably co-expressed genes, because no other gene pair in that pool was affected by the presence of mutants. Independent of the particular results on the characteristics of AT5G12080, this case study can be regarded as a general example of insights attainable with our approach. In particular, our results showed that by the analysis of co-expression variation between two partially overlapping data sets (i.e. mut+ and mut-), it is possible to make inferences on the functional behavior of a single gene.

Finally, considering the pool of the most co-expressed genes, we observed further 39786 gene pairs showing no significant variations of co-expression (Table 2), while maintaining significant positive correlations in both data sets, mut+ and mut-, respectively.

### Arabidopsis thaliana resources for gene co-expression analyses

Despite the multitude of resources available for *A. thaliana* gene expression data, and the heterogeneity of the associated tools, our survey of the most referenced web-based platforms (Table 1) illustrates the variety of the specific features that could possibly affect the results of gene co-expression analysis, including the type of statistics implemented for assessing pairwise gene co-expression. A total of nine different co-expression metrics are available in the 11 tested platforms (Table 1), including Pearson product-moment correlation coefficient, common in eight of the platforms, and lacking in Cress Express, where  $r^2$  is used instead, CoP and GeneMania; the slope of least squares best-fitting regression line (exclusively used in Cress Express); the cosine correlation and the vertex F-measure (both exclusively used in CoP); the mutual ranking (used in ATTED II); the Spearman and Kendall rank correlation coefficients (both used in CSB.DB) and weight of the relation (in GeneMania). However, the pre-processing methods for microarray data normalization, though changing, were less variable among the platforms (Table 1). In particular, the RMA method is proposed by six platforms (AtCOEGIS, ATTED II, CORNET, Cress Express, GeneCAT, Genevestigator), the MAS 5.0 protocol was implemented in five resources (BAR, CoP, Cress Express, CSB.DB, Genevestigator), while the GCRMA method is proposed only in Cress Express. Interestingly, Cress Express offers data normalized by all the three different methods [45, 46]. Finally, information on the normalization method implemented was not available from GeneMania and PlaNet platforms. Moreover, all the platforms exploit different data and different number of slides.

In addition, it is not always possible to select the preferred data set or to get information on its content (Table 1 and Additional Table 4 in the Supplementary Material).

### Comparative meta-analysis of results from different platforms

The comparative meta-analysis of gene co-expression results from the tested databases, based on the four query genes AT5G17420, AT5G06680, AT1G20580 and AT1G01290, and on the default settings of the databases provided different outcomes for the different queries (Figures 3 and 4). The figures summarize the number of genes in common when comparing lists of results between two different platforms.

In the case of AT5G17420 (CESA7), which is well known for being involved in cell-wall synthesis, the lists of the topmost 20 co-expressed genes produced by the different platforms were relatively coherent (see also Additional Table 4 in the Supplementary Material), despite the relevant differences of data sets, co-expression metrics and normalization methods proposed by each database. In particular, ATTED II, CoP, BAR, Cress Express, GeneCAT and Genevestigator shared about 50% of the genes in the resulting lists, with >70% of shared genes when considering pair-wise comparisons including BAR, CoP and GeneCAT results (Figure 3) [47]. In contrast, the PlaNet and GeneMania output lists shared at most 25% with the results found in the lists provided by the other databases, and only 1 gene of 20 when compared with each other. This can be explained by the fact that these two platforms do not offer ranked lists of co-expressed genes, as they are more focused on defining co-expressed gene modules. Therefore, their results are not appropriately sorted and the selection of the first 20 genes in

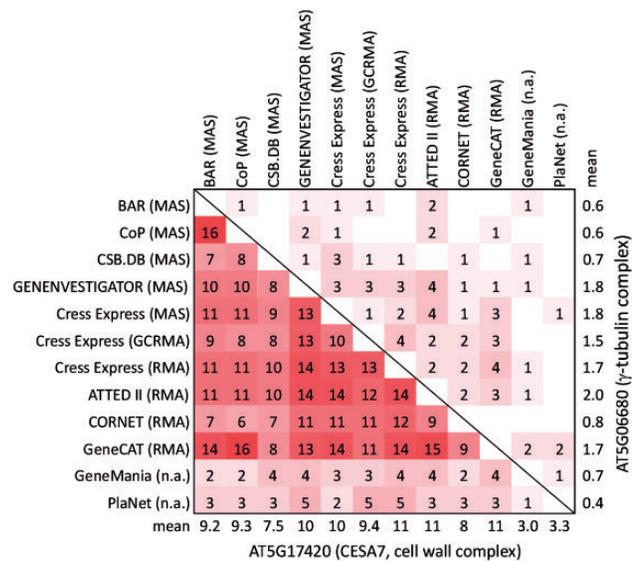
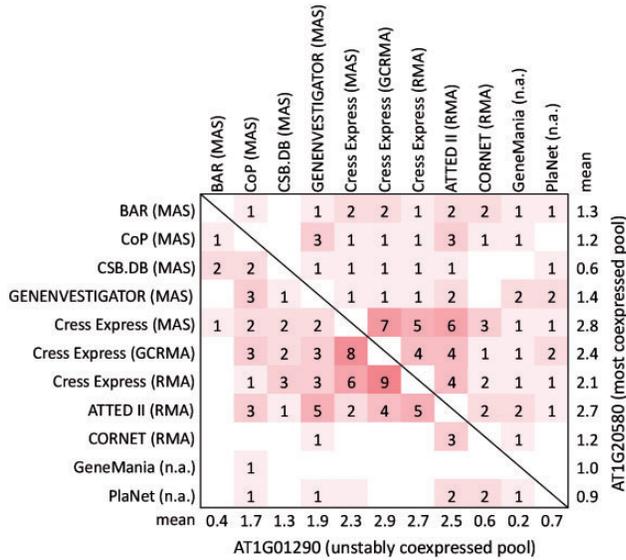


Figure 3. Pair-wise comparison of co-expression analysis results from different public platforms, when querying AT5G06680 (upper right values with respect to the matrix diagonal) and AT5G17420 (CESA7, lower left values). For each matrix cell, data refer to the number of items shared between the two lists of topmost 20 co-expressed genes, as proposed by the two platforms corresponding to cell row and column. Zero scores are omitted to improve readability. Color shading highlights the cell values, according to a color scale ranging between 0 (white) and 20 (full red). Data normalization methods used for the analysis are reported in brackets. For each resource and for each query gene, the mean number of shared items among the databases is also reported, excluding the self-matching values along the matrix diagonal. A colour version of this figure is available online at BIB online: <http://bib.oxfordjournals.org>.



**Figure 4.** Pair-wise comparison of co-expression analysis results from different public platforms, when querying AT1G20580 (upper right values with respect to the matrix diagonal) and AT1G01290 (lower left values). For each matrix cell, data refer to the number of items shared between the two lists of topmost 20 co-expressed genes, as proposed by the two platforms corresponding to cell row and column. Zero scores are omitted to improve readability. Color shading highlights the cell values, according to a color scale ranging between 0 (white) and 20 (full red). Data normalization methods used for the analysis are reported in brackets. For each resource and for each query gene, the mean number of shared items among the databases is also reported, excluding the self-matching values along the matrix diagonal. A colour version of this figure is available online at BIB online: <http://bib.oxfordjournals.org>.

the lists cannot be compared with the ones from the other platforms. Finally, CSB.DB and CORNET showed an intermediate outcome, with concordance with other lists below the 50%, though still relatively high.

Considering previous knowledge about co-expression patterns of CESA7, the presence of AT5G44030 (CESA4) and AT4G18780 (CESA8) in the results from the different platforms can be used as an indicator of the output reliability (Additional Table 5 in the Supplementary Material) [47]. Indeed, most of the platforms confirm the co-expression of the CESA4-7-8 complex. ATTED II, CORNET, Cress Express (three different lists for each normalization method), GeneCAT, GeneMania and Genevestigator, where the ranking of co-expressed genes is defined by the *P*-value of the co-expression metric, listed CESA4 and CESA8 within the top three genes co-expressed with CESA7. Conversely, CESA4 and CESA8 were not found in the list of results from CSB.DB and PlaNet. In the case of CSB.DB, the probe set of CESA8 was missing in the data set exploited for the current analysis, while PlaNet did not show CESA4 in the top 20 output list, because of the lack of ranking when reporting the list. This gene, however, is identified when considering the whole cluster of genes co-expressed with CESA7. The nearly coherent results from CESA7, which has fewer exceptions, mainly owing to the platform set up (lacking probes or visualization limits), are not confirmed when considering the results of co-expression analyses for AT5G06680, implied in the  $\gamma$ -tubulin complex [48]. Remarkably, the average number of shared items among the output lists did not exceed 10% in all cases (Additional Table 6 in the Supplementary Material). Surprisingly, considering pair-wise comparisons between different output lists, in 23 cases of 66 a complete discordance between the results provided by different tools was observed, with

most of the lists pairs showing no items in common (Figure 3). In 21 other cases only one gene was shared, while the maximum number of shared items was 4 (Figure 3), which is even more surprising considering that, among the tested resources, the Cress Express platform offered three identical collections that were queried only by changing the normalization methods (Figure 3).

Finally, querying the platforms with AT1G20580 and AT1G01290, selected among the most stable and unstable co-expressed gene pools, respectively, detected from our in-house presented analysis, produced contrasting outcomes among the tested resources (Figure 4). In particular, for both genes the highest concordance among the resulting lists was observed for the three results from the Cress Express platform, which vary only for the parameter settings. However, the number of shared items among these lists was on average 5 and 7 for AT1G20580 and AT1G01290, respectively (Figure 4). Comparing the results from other platforms, the mean number of shared items among the different lists were consistently low for both query genes, rarely exceeding 10% of the list items (i.e. 2 of 20), with the exception of the pair-wise comparisons involving the ATTED II platform, which showed slightly higher concordances (Figure 4).

Our meta-analysis showed that different platforms for gene co-expression analysis, when queried with a single gene, can produce contrasting results as a consequence of different experimental collections and specific technical features of the tools used for the analysis. The outcome, typically a list of the most co-expressed genes, can be totally different for different tools, depending on the genes queried, as observed when comparing the results for AT5G17420 (CESA7) and AT5G06680. Indeed, differences in co-expression results between specific tools can be highly variable, as shown, for example, when comparing BAR and CoP output rankings. These two platforms shared 16 and 1 items when queried with AT5G17420 (CESA7) and AT5G06680, respectively (Figure 3). This occurs even when the analysis is carried out on the same query gene using the same experimental data sets and correlation metrics. This was evident in the three versions of the results from the Cress Express platform, where discordances are produced by data pre-processing approaches, which often cannot be controlled by the user, such as the normalization method. In this respect, our results are consistent with previous findings on the important role of the normalization methods in determining the results of co-expression analysis. Our studies, however, provide further evidence, as they are based on a higher number of tested platforms and also consider other features [49].

The possible bias associated with the specifics of the resource tools can sometimes be relevant, being most important in the case of genes with highly sensitive co-expression profiles, such as AT5G06680, while being less relevant for genes like CESA7 (cf. cell color shading in Figure 3 versus Figure 4). In this latter case, it is interesting to note that the co-expression results are less variable, despite some huge difference in the data set size and experiment content among the platforms (e.g. passing from 11 171 slides of Genevestigator to 351 of GeneCAT). This consensus result supports more reliable assessments of a possible functional relationship among the listed genes. However, the assessment of the co-expression profile of AT5G06680 may be harder to establish, probably owing to the higher variability of the possible expressed 'partners' according to the conditions of expression [Additional Table 1 in the Supplementary Material and changes in number of correlated genes detected in the mut+ and mut- data sets (data not shown)].

In addition to the data set composition, our evidence highlights that the normalization method had a remarkable influence on the results, as we showed for the Cress Express platform. In this case, the same data source collection, queried by three versions (3.0, 3.1 and 3.2) and only differing by normalization algorithm (RMA, GCRMA and MAS5.0, respectively), produced highly variable output lists. In particular, the number of shared genes between the three pairs of lists produced by the three approaches were 4, 2 and 1 for queries with AT5G06680, and 13, 13 and 10 for queries with CESA7. Such numbers, although different between the two query genes, are consistent with the range of items that the Cress Express tools shared with other platforms based on different data sets. In other words, the variability of results related to normalization methods was of the same order of magnitude of data variability associated mainly with the reference experimental data set. In this respect, the results of our meta-analysis showed a strong influence of several technical issues on the outcomes of gene co-expression analyses, also suggesting a role of gene expression variation on the stability of co-expression profiles across different experimental collections.

Hence, the evidence presented here suggests that before proceeding with interpretation of results from web-based tools, various tests, by different parameter setting, on different resources should be performed. Then, results should be carefully compared to assess their consistency. Otherwise, the result from a single resource should be regarded as limited not only by the specific array collections made available, but also by the particular data processing methods applied.

In this context, a relevant issue could be how to fuse data from different web-based resources to fully and suitably exploit their information content. The effect of heterogeneity of data sources and tools on co-expression results highlights the need for reconciliations at data and at methodology levels, and for tools that aid the selection of appropriate sources and methods. Uniformity in data sets would require data sharing for the creation of comprehensive collections. These should be accessible from the different sources at the database level. Tools for flexible data selection would permit the design of suitable co-expression analyses, before further assessments. Data collection in the form of a flat file exported from reference resources is widespread in bioinformatics, including for microarray data [22, 50]. Data sharing between different resources could also be solved using web services-based approaches [51]. This would overcome issues like elimination of redundancy by facilitating removal of duplicated experiments from different resources and overcoming relevant issues like semantics [52]. Indeed, a web services framework could favor the selection of appropriate data sets for successive analyses, although it would not overcome methodological issues such as differences in methods for raw data processing and normalization [23].

However, the possibility of exploiting dedicated resources including reference collections for a single or dedicated species taking advantage of online user-friendly tools and avoiding the burden of data management and software selection would be inestimable value for nonexpert users. Data fusion at results level has also been proposed, based on formal probabilistic approaches, capable of integrating heterogeneous outputs from different resources [53]. Computational implementations can include the representation of each input data set as a separate kernel and the weighted optimized combination of these kernels to reconstruct co-expression patterns [54], as well as Bayesian network-based functions [55], decision trees [56] and weighted rank aggregation [57]. In particular, we tested the

RankAggreg R package [58], which exploits the rank aggregation method. This R package takes different lists of ranked elements as input. Ranked lists of co-expressed genes can also be considered. As output, the method provides an optimal list, which is defined using Spearman or Kendall distances and Cross-Entropy Monte Carlo or Genetic algorithms. The results from the different combinations of metrics and methods for the lists of genes that are co-expressed with CESA 7 from the 11 platforms here considered are reported in the Additional Table 7 in the Supplementary Material. Although the methods presented support the merging of co-expression results from different resources, this problem has not yet been fully solved, as no generally robust method can be routinely applied to noisy results from heterogeneous resources [53]. Moreover, as gene co-expression results are highly context dependent, being variable for different genes and data sources, as we have shown here, a reliable data-fusion approach should always be consistent with the selection of appropriate experiments fitting the underlying working hypothesis.

In the context of gene co-expression analysis, the definition and the mining of co-expression networks is also worthy of mention. Indeed, matrices, reporting co-expression results for gene pairs, represent the source data for the definition of networks of co-expressed genes. This approach expands the view from lists of genes to graphs of genes, where connections are based on co-expression. This can be accomplished considering simple ranks and/or significance of pair-wise correlation values [59], but also exploiting more sophisticated algorithms [60]. Some of the platforms we have reviewed here provide embedded network generation tools (Additional Methods in the Supplementary Material [61]) or data formats suitable for network visualization by appropriate tools, among which Cytoscape is one of the most commonly used [62].

Although the instability of gene co-expression across different resources can be considered a major source of uncertainty inflating the confidence associated with the definition of co-expression networks, with the transfer of uncertainty from the pair-wise to the network level, the exploitation of networks can support the assessment of consensus among different results. Different results may indeed support the identification of key nodes, or conserved patterns, which may be cross confirmed in their co-expression by different approaches, supporting a result-based merging by suitable mining [63, 64] or visualization tools [65, 66].

## Conclusions

Our study consistently demonstrated that gene co-expression analysis of the transcriptome from the same target plant can be significantly affected by the heterogeneity of the reference data set, which we mimicked by the sample case of inclusion/exclusion of experiments from mutants. The results can vary significantly for different gene pools, but, most important, the mutant-related effect can be characterized in terms of occurrence, magnitude and direction. Interestingly, our results provided information even at the single-gene level, suggesting that a candidate gene (i.e. AT5G12080), known to be coding for relevant biological functions, can potentially discriminate between mutant and nonmutant samples. Moreover, our survey of different platforms available on the web also showed that they can produce remarkably variable results according to several technical and biological factors. Considering technical issues, the number and the heterogeneity of array collections among different platforms can be considered the major drivers. However,

as shown by our meta-analysis of the results attainable by single-gene query, different normalization methods, as well as metrics used for assessing gene co-expression, can also play an important role. The variability of results appears to be highly dependent on the single gene used for the query upon simple analyses, as in the case of stable gene networks, when the tested platforms are generally more prone to concordant outcomes.

All of our results, although they were produced by a general survey, a meta-analysis and an original assessment on a limited number of genes, highlight important issues and consequences for data processing in gene co-expression analyses. Our study suggests that data mining should include 'databases mining', i.e. it is necessary to move away from the idea of analyzing one single collection to the idea that it is necessary to compare all, or at least the majority, of the available resources, paying close attention to the bias the selection of a reference can bring. Moreover, the results obtained from different resources must be tested for consistency against the selected hypotheses. Finally, considering the high number of available resources, the diversity of databases should be fully exploited to get added value information and to increase the robustness of results that, otherwise, could be biased by an inappropriate usage of one single reference.

### Key Points

- *Arabidopsis thaliana* co-expression platforms produce remarkably different results because of several technical and biological issues.
- Gene co-expression analysis is significantly affected by the heterogeneity of the reference data set and methods, for example, inclusion/exclusion of mutants.
- Comparing and merging results from the manifold gene co-expression analysis resources is the strategy to get robust and reliable results.

### Supplementary data

Supplementary data are available online at <http://bib.oxfordjournals.org/>.

### Funding

This work was supported by the Genopom PRO (project code: PON02\_00395\_3215002) and GenHORT (PON02\_00395\_3215002) Projects, funded by MIUR (Ministero dell Istruzione, dell Università e della Ricerca, Italy) within the PON (Programma Operativo Nazionale) funding program.

### Acknowledgments

We thank three anonymous reviewers for their suggestions that improved the manuscript. Special thanks goes to Dianna Pickens for her support for the English revision.

### References

1. Chang TW. Binding of cells to matrixes of distinct antibodies coated on solid surface. *J Immunol Methods* 1983;65:217–23.
2. Schena M, Shalon D, Davis RW, et al. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 1995;270:467–70.
3. Lashkari DA, DeRisi JL, McCusker JH, et al. Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proc Natl Acad Sci USA* 1997;94:13057–62.
4. Slonim DK, Yanai I. Getting started in gene expression microarray analysis. *PLoS Computat Biol* 2009;5:e1000543.
5. Hoheisel JD. Microarray technology: beyond transcript profiling and genotype analysis. *Nat Rev Genet* 2006;73:200–10.
6. Schulze A, Downward J. Navigating gene expression using microarrays - a technology review. *Nat Cell Biol* 2001;38:E190–5.
7. Oliver S. Guilt-by-association goes global. *Nature* 2000;403:601–3.
8. Quackenbush J. Genomics. Microarrays - guilt by association. *Science* 2003;302:240–1.
9. Heyer LJ, Kruglyak S, Yooseph S. Exploring expression data: identification and analysis of coexpressed genes. *Genome Res* 1999;9:1106–15.
10. Yang Y, Speed T. *Design and Analysis of Comparative Microarray Experiments*. London: Chapman & Hall/CRC, 2003, 35–91.
11. Foss EJ, Radulovic D, Shaffer SA, et al. Genetic basis of proteome variation in yeast. *Nat Genet* 2007;39:1369–75.
12. Fu J, Keurentjes JJ, Bouwmeester H, et al. System-wide molecular evidence for phenotypic buffering in *Arabidopsis*. *Nat Genet* 2009;41:166–7.
13. Ghazalpour A, Bennett B, Petyuk VA, et al. Comparative analysis of proteome and transcriptome variation in mouse. *PLoS Genet* 2011;7:e1001393.
14. Gerstein MB, Rozowsky J, Yan KK, et al. Comparative analysis of the transcriptome across distant species. *Nature* 2014;512:445–8.
15. Bostan H, Chiusano ML. NexGenEx-Tom: a gene expression platform to investigate the functionalities of the tomato genome. *BMC Plant Biol* 2015;15:148.
16. Leinonen R, Sugawara H, Shumway M, et al. The sequence read archive. *Nucleic Acids Res* 2011;39:D19–21.
17. Giardine B, Riemer C, Hardison RC, et al. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res* 2005;15:1451–5.
18. Initiative AG. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 2000;408:796–815.
19. Honys DT, Well D. Transcriptome analysis of haploid male gametophyte development in *Arabidopsis*. *Genome Biol* 2004;5:R85.
20. Craigon DJ, James N, Okyere J, et al. NASCArrays: a repository for microarray data generated by NASC's transcriptomics service. *Nucleic Acids Res* 2004;32:D575–7.
21. Rehrauer H, Aquino C, Gruitsem W, et al. AGRONOMICS1: a new resource for *Arabidopsis* transcriptome profiling. *Plant Physiol* 2010;152:487–99.
22. Barrett T, Troup DB, Wilhite SE, et al. NCBI GEO: mining tens of millions of expression profiles—database and tools update. *Nucleic Acids Res* 2007;35:D760–5.
23. Brazma A, Hingamp P, Quackenbush J, et al. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nat Genet* 2001;29:365–71.
24. Brazma A, Parkinson H, Sarkans U, et al. ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res* 2003;31:68–71.
25. Pepper SD, Saunders EK, Edwards LE, et al. The utility of MASS expression summary and detection call algorithms. *BMC Bioinformatics* 2007;8:273.
26. Rhee SY. The *Arabidopsis* Information Resource (TAIR): a model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. *Nucleic Acids Res* 2003;31:224–8.
27. Yim WC, Yu Y, Song K, et al. PLANEX: the plant co-expression database. *BMC Plant Biol* 2013;13:83.

28. Vandepoele K, Quimbaya M, Casneuf T, et al. Unraveling transcriptional control in Arabidopsis using cis-regulatory elements and coexpression networks. *Plant Physiol* 2009;150:535–46.
29. Obayashi T, Hayashi S, Saeki M, et al. ATTED-II provides coexpressed gene networks for Arabidopsis. *Nucleic Acids Res* 2009;37:D987–921.
30. Toufighi K, Brady SM, Austin R, et al. The botany array resource: e-Northern, expression angling, and promoter analyses. *Plant J* 2005;43:153–63.
31. Ogata Y, Suzuki H, Sakurai N, et al. CoP: a database for characterizing co-expressed gene modules with biological information in plants. *Bioinformatics* 2010;26:1267–8.
32. De Bodt S, Carvajal D, Hollunder J, et al. CORNET: a user-friendly tool for data mining and integration. *Plant Physiol* 2010;152:1167–79.
33. Srinivasainagendra V, Page GP, Mehta T, et al. CressExpress: a tool for large-scale mining of expression data from Arabidopsis. *Plant Physiol* 2008;147:1004–16.
34. Steinhauser D, Usadel B, Luedemann A, et al. CSB.DB: a comprehensive systems-biology database. *Bioinformatics* 2004;20:3647–51.
35. Mutwil M, Obro J, Willats WG, et al. GeneCAT—novel webtools that combine BLAST and co-expression analyses. *Nucleic Acids Res* 2008;36:W320–6.
36. Mostafavi S, Ray D, Warde-Farley D, et al. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biol* 2008;9:S4.
37. Zimmermann P, Hirsch-Hoffmann M, Hennig L, et al. GENEVESTIGATOR. Arabidopsis microarray database and analysis toolbox. *Plant Physiol* 2004;136:2621–32.
38. Mutwil M, Klie S, Tohge T, et al. PlaNet: combined sequence and expression comparisons across plant networks derived from seven species. *Plant Cell* 2011;23:895–910.
39. Schmid M, Davison TS, Henz SR, et al. A gene expression map of Arabidopsis thaliana development. *Nat Genet* 2005;37:501–6.
40. Lee HK, Hsu AK, Sajdak J, et al. Coexpression analysis of human genes across many microarray data sets. *Genome Res* 2004;14:1085–94.
41. Weaver B, Wuensch KL. SPSS and SAS programs for comparing Pearson correlations and OLS regression coefficients. *Behav Res Methods* 2013;45:880–95.
42. Eckardt NA. Cellulose synthesis takes the CesA train. *Plant Cell* 2003;15:1685–7.
43. Haswell ES, Peyronnet R, Barbier-Brygoo H, et al. Two MscS homologs provide mechanosensitive channel activities in the Arabidopsis root. *Curr Biol* 2008;18:730–4.
44. Velez KM, Maskaev G, Frick EM, et al. Arabidopsis MSL10 has a regulated cell death signaling activity that is separable from its mechanosensitive ion channel activity. *Plant Cell* 2014;26:3115–31.
45. Hubbell E, Liu WM, Mei R. Robust estimators for expression analysis. *Bioinformatics* 2002;18:1585–92.
46. Wu Z, Irizarry RA, Gentleman R, et al. A model-based background adjustment for oligonucleotide expression arrays. *J Am Stat Assoc* 2004;99:909–17.
47. Eklund AC, Szallasi Z. Correction of technical bias in clinical microarray data improves concordance with known biological information. *Genome Biol* 2008;9:R26.
48. Soga K, Kotake T, Wakabayashi K, et al. Transient increase in the transcript levels of gamma-tubulin complex genes during reorientation of cortical microtubules by gravity in azuki bean (*Vigna angularis*) epicotyls. *J Plant Res* 2008;121:493–8.
49. Usadel B, Obayashi T, Mutwil M, et al. Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. *Plant Cell Environ* 2009;32:1633–51.
50. Kolesnikov N, Hastings E, Keays M, et al. ArrayExpress update—simplifying data submissions. *Nucleic Acids Res* 2015;43:D1113–16.
51. Pettifer S, Thorne D, McDermott P, et al. An active registry for bioinformatics web services. *Bioinformatics* 2009;25:2090–1.
52. Goble C, Stevens R. State of the nation in data integration for bioinformatics. *J Biomed Inform* 2008;41:687–93.
53. Troyanskaya OG. Putting microarrays in a context: integrated analysis of diverse biological data. *Brief Bioinform* 2005;6:34–43.
54. Lanckriet GR, De Bie T, Cristianini N, et al. A statistical framework for genomic data fusion. *Bioinformatics* 2004;20:2626–35.
55. Troyanskaya OG, Dolinski K, Owen AB, et al. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). *Proc Natl Acad Sci USA* 2003;100:8348–53.
56. Zhang LV, Wong SL, King OD, et al. Predicting co-complexed protein pairs using genomic and proteomic data integration. *BMC Bioinformatics* 2004;5:38.
57. DeConde RP, Robert P, et al. Combining results of microarray experiments: a rank aggregation approach. *Stat Appl Genet Mol Biol* 2006;5:15.
58. Pihur V, Datta S, Datta S. RankAggreg, an R package for weighted rank aggregation. *BMC Bioinformatics* 2009;10:62.
59. Obayashi T, Kinoshita K. Rank of correlation coefficient as a comparable measure for biological significance of gene coexpression. *DNA Res* 2009;16:249–60.
60. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 2008;9:559.
61. Warde-Farley D, Donaldson SL, Comes O, et al. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res* 2010;38:W214–20.
62. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003;13:2498–504.
63. Zhang J, Lu K, Xiang Y, et al. Weighted frequent gene co-expression network mining to identify genes involved in genome stability. *PLoS Comput Biol* 2012;8:e1002656.
64. Skinner J, Kotliarov Y, Varma S, et al. Construct and compare gene coexpression networks with DAPfinder and DAPview. *BMC Bioinformatics* 2011;12:286.
65. Provart N. Correlation networks visualization. *Front Plant Sci* 2012;3:240.
66. Enright AJ, Ouzounis CA. BioLayout—an automatic graph layout algorithm for similarity visualization. *Bioinformatics* 2001;17:853–4.