Nucleic Acids Research, , Vol. , No. ***1***

A matter of size: how to deal with samples from a large gene expression dataset to construct a robust co-expression network.

Franziska Liesecke [1], Johan-Owen de Craene [1], Sébastien Besseau [1], Vincent Courdavault [1], Marc Clastre [1], Valentin Vergès [1], Nathalie Giglioli-Guivarc'h [1], Gaelle Glévarec [1], Olivier Pichon [1] and Thomas Dugé de Bernonville [1,*]

[1]Université de Tours, EA2106 Biomolécules et Biotechnologies Végétales, 31 avenue Monge, Tours, F-37200, France

ABSTRACT

Large-scale gene co-expression networks are an effective methodology to analyze sets of co-expressed genes and discover new gene functions or associations. Distances between genes are estimated according to their expression profile and are visualized in networks that may be further partitioned to reveal communities of co-expressed genes. Creating expression profiles is now eased by the large amounts of publicly available expression data (microarrays and RNA-seq). Although many distance calculation methods have been intensively compared and reviewed in the past, it is unclear how to proceed when many samples reflecting different conditions are available. Should as many samples as possible be integrated into network construction or be partitioned into smaller sets of more related samples? Previous studies have indicated a saturation in network performance to capture known associations once a certain number of sample is included in distance calculations. Here, we examined the influence of sample size on co-expression network construction from microarray and RNA-seq expression data using three plant species. We tested different down-sampling methods and compared network performance in recovering known gene associations to networks obtained from full datasets. We further examined how aggregating networks may help increase this performance by testing four aggregation methods.

INTRODUCTION

Co-expression networks are proven to be efficient approaches to uncover biologically relevant gene pair associations. The starting expression matrix from which distance measurements are calculated to determine correlations between genes, is expected to drive the final shape and content of the network. In relevant co-expression networks, the ratio of known and experimentally proven associations (True Positives) over the total number of captured links is expected to be high. Evaluating the True Positive Rate (TPR) and False Positive Rate (FPR) of a given network with respect to a reference annotation set remains challenging but simple machine learning algorithms have been shown to efficiently calculate these rates and assess a network quality (1).

Combining individual datasets (from independent studies, SRP /ERP numbers in RNA-seq or GSE in microarrays) is expected to increase biological situation range and help capture transient associations. Real gene pair associations will be found in the network only if their common expression is detected in the starting dataset. Including more datasets in co-expression analyses should therefore add biological situations where such co-expression occurs. Contrastingly, increasing the sample number in an expression dataset may also result in increased noise together with decreased capacity to detect transient associations, following the garbage in garbage out principle. An open question remains about the number of expression datasets needed to build the most biologically relevant networks, *i.e.* capturing as many real associations as possible while keeping the number of false negatives low.

If including more samples to construct a network improves its quality, it is still unclear how many datasets are sufficient to capture relevant gene associations. For model species with many available expression datasets (*e.g.* more than 1,000 samples), global networks can be constructed from one expression matrix combining every available sample but does it capture more efficiently biological associations than networks obtained from smaller or down-sampled datasets? How adding or removing samples alters the network composition is not clear.

A pioneer study(2) used an *Escherichia coli* microarray data compendium. They analyzed dependencies among samples and found that compendium subsets perform better than the full one in transcriptional regulatory network inference. The low efficiency of the global network was attributed to sample redundancies but could be circumvented by calculating an optimal effective number of samples. In their work, the full compendium (376 samples) could be down-sampled to 50% without decreasing network quality.

Using a larger *E. coli* expression compendium (524 samples) as well as synthetic datasets (with up to 2,000 samples), Altay et al(3) tested several information theory based inference methods as well as the sample size effect. In this case, both simulated and real datasets showed that *ca.* 100 samples were sufficient to capture transcriptional genome-wide regulations. These previous reports(2, 3) took advantage of the well known *E.coli* regulatory network to evaluate their

*To whom correspondence should be addressed. Tel: +33 247 367023; Email: thomas.duge@univ-tours.fr

co-expression networks. In another study(4), co-expression networks were obtained after applying a Random Matrix Theory process to threshold similarity matrices calculated with Pearson Correlation Coefficients (PCC). The effects of both gene number and sample size were analyzed for 3 species: Human, Rice and Yeast. The authors have shown a high edge conservation between full and down-sampled networks. However, new edges appeared with smaller datasets (down to 25% of the initial size) while other edges were lost. This indicates that conserved associations between genes are easy to uncover while revealing more transient association typically depends on the nature of samples in the dataset. Conserved associations corresponded to functional associations, suggesting that genes added or removed in the down-sampled networks mostly were found in already densely connected modules and were weakly connected to others (*i.e.* they were not hub genes).

Supervised down-sampling of a large dataset by finding the most appropriate dataset has been proposed to improve pathway reconstruction. Using a set of query genes, Hibbs et al(5) calculated correlations among this set on SVD-transformed individual datasets of *S. cerevisiae*. Each dataset was weighted according to its relevance, *i.e.* datasets maximizing PCC are given more weight. These weights are used next to calculate PCC of every gene with each query gene. This procedure is known as the SPELL algorithm (Serial Pattern of Expression Levels Locator). It has also been reported that creating subsets of related samples using a k-means approach improves feature detection(6). However, it remains to be determined whether individual networks resulting from partitions of down-sampled datasets could be more efficient in capturing associations than a network calculated from all initial samples. Aggregated networks have also been shown to improve the recovery of biologically relevant associations(7, 8). The underlying idea is that conserved coexpression links between 2 genes over several datasets reinforce the existence of a true association between these 2 genes (7). A web-based tool named MEM was designed to merge co-expression lists obtained from individual datasets(11).This multi-species microarray-based tool allows users to find the best coexpressed genes with an input query gene and manually excludes less relevant datasets considering the input. In MEM, gene ranks calculated with a query gene are aggregated over the selected datasets by using a binomial distribution hypothesis to attribute p-values to ranks and by taking the minimum value of all *p*-value. Ballouz et al(1) have constructed individual networks for different experiments and subsequently aggregated them, either taking all datasets or only the most significant (as indicated by AUROC of GO terms). They have revealed a clear improvement over individual experiments, probably because it has the advantage of combining moderately significant or condition-specific relationships.

Our aims are: (i) to establish the impact of the sample size on the recovery of relevant associations at both global and targeted levels and (ii) get more insights on the way individual and smaller networks may be aggregated to generate stronger networks. The latter implies a trade-off between capturing the maximum number of known associations and limiting the total association number.

MATERIALS AND METHODS

Dataset preparation

Microarray data were obtained from signal intensities in .CEL files downloaded from ArrayExpress in R(12, 41). Raw signals were quantile normalized with the RMA procedure in R(14). Outlier arrays were detected by monitoring quartile distribution together with Kolmogorov-Smirnov testing against an empirical cumulative distribution curve(6). RNA-seq data were obtained by downloading raw .fastq files from the EBI ENA. Reads were quasi-mapped with Salmon(15) on reference transcript assemblies obtained from Ensembl Plant to quantify transcript abundance. Assemblies were TAIR10 for *Arabidopsis thaliana*, SL2.50.31 for *Solanum lycopersicum* and AGPv3.31 for *Zea mays*. All accessions are indicated in Supplementary Table 1. Dataset sizes are indicated in Table 1.

Dataset partitioning

Initial expression matrices were down-sampled in three different ways (Figure 1). In the first, k-means partitioning was performed in R using with 200 random starts. The optimal value of k was graphically determined for each dataset using the elbow method and varying k between 10 and 100 by increments of 10. In the second, datasets were partitioned by grouping arrays or runs from the same study/project (GSE for microarrays, SRP, ERP or DRP number for RNA-seq). In the last, we down-sampled datasets by randomly selecting a given number of samples. For each sample size, the random sampling was performed many times until the resulting down-sampled datasets represented at least 90% of the samples contained in the full expression matrix.

Highest reciprocal ranking of Pearson Correlation Coefficients

Distances between transcript expression profiles were determined by calculating PCCs for each transcript pair. PCCs were next ranked so that for each gene, the rank value ranges from 0 (the gene itself) to N (the total number of genes), and the final rank value for a given gene pair was the highest of the two, *e.g.* for a gene pair A and B, HRR(A,B)=max(rank(cor(A,B)), rank(cor(B,A)). Computations were done in parallel with an MPI program written in C(16). The resulting distance matrices were arbitrary thresholded at a HRR$<$600 to obtain a large list of best co-expressed genes and these lists were further thresholded at different cut-off values to compare network performance at different sizes. As threshold choice has been shown to strongly influence network topology(18), the performance of networks obtained from different datasets

Table 1. Dataset sizes. Number of genes x number of samples.

	Microarrays	RNA-seq
Arabidopsis thaliana	22,178 x 10,138	33,602 x 1,676
Solanum lycopersicum	6,284 x 627	32,419 x 1,046
Zea mays	10,309 x 680	61,581 x 2,516

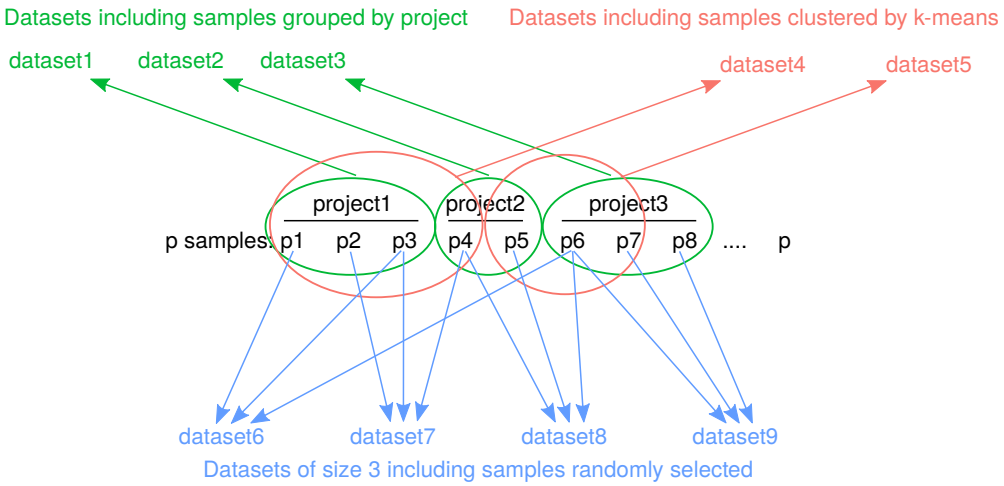

Figure 1. Down-sampling strategies. Starting from a given number of p samples retrieved from publicly available accessions, three methods were used to partition samples. Samples were grouped by their project accession, clustered by k-means or randomly selected (with possible redundancy between datasets).

(differing in size and the partitioning method) was evaluated for different network sizes.

Network evaluation

Networks were evaluated for their ability to recover known or expected relationships between genes. Gene relationships obtained in a network were compared to gene associations described in Gene Ontology (GO) terms or the Reactome database(19). GO annotation files were downloaded from the Agrigo database v2.0(20). Network ability to capture gene pairs associated with identical GO terms was evaluated by applying a neighbor voting algorithm which measures how well connections in the co-expression network predict a gene annotation. A three fold cross validation was performed and used to calculate an Area Under the Receiver Operating Characteristic (AUROC) for each GO term with the EGAD R package(21). Network performance was estimated with the GO AUROC which corresponded to the average of all GO term AUROCs. GO AUROCs of 0.5 and 1 respectively indicate random and perfect predictability. We also tested significant enrichment of networks with GO terms using an hypergeometric test. To test gene association into biological pathways, global networks with a 1 million edges were queried with guide gene (GG) sets described in the Reactome database. Genes co-expressed with the guides were captured to construct a sub-network, a process known as Pathway Level Coexpression(23). A total of 11 biological pathways were analyzed for each species. Subnetworks were evaluated in their ability to capture biologically relevant GO terms (GO AUROCs and significantly enriched GO terms) as well as to partition guide genes into their exact groups/pathway by measuring network modularity and a normalized Chi-squared metric(16). Networks were constructed and analyzed with the R igraph package(17), including several topological metrics such as transitivity (the probability that adjacent nodes of a given node are connected) and the log likelihood of node degree to fit a power law distribution.

RESULTS

Individual down-sampled matrices

Biologists aiming at constructing co-expression networks for their model species may face a large number of samples and experiments in public databases. A largely unresolved question is the way to process such data. Should the data be combined or processed as data subsets? We evaluated both methods by generating down-sampled expression datasets from large compendia (Figure 1) in three different ways (Figure 1). Samples were grouped according to their project accession number, according to their similarity using a k-means clustering or by random sampling at different fixed sample sizes. Because we allowed replacement in the sampling procedure, some samples were found in several matrices (5 matrices on average and rarely more than 10) (Supplementary Figure 1). A large variety of data subsets were therefore investigated (Figure 2A; 3,673 for*A. thaliana*, 947 for *Z. mays* and 913 for *S. lycopersicum*). The randomized sampling procedure generated many small data subsets (a minimum of 25 samples) and fewer large ones (Figure 2B). Data subsets obtained from project-grouped or k-means clustered samples rarely contained more than 100 samples (Figure 2B). Each data subset was used to construct a distance matrix using HRR ranked PCC thresholded at HRR>600 to get a large list of the best co-expressed gene pairs. These lists were next cut off at different confidence levels to construct networks and evaluate their performance in capturing GO terms. First, many more edges were retained for network construction with less stringent confidence thresholds (Figure 2C; Spearmans *rho*>0.95, *p*-value<2e-16). A clear positive relationship was observed between edge number and GO AUROC (Spearmans *rho*>0.7, *p*-value<2e-16) revealing that more true than false positive edges were included in networks at the considered network sizes. At a given confidence threshold (*e.g.* 0.1), sample size tended to be negatively correlated with edge number (average Spearmans *rho*=-0.33, *p*-value<2e-16) (Figure 2D). Positive significant correlations between sample size and GO AUROC were

observed for the random down-sampling and all 3 species (Spearmans *rho*>0.5, *p*-value<5e-14), but for *A. thaliana* and *Z. mays* networks only with k-means down-sampled matrices (Spearmans *rho*>0.4, *p*-value<0.05) (Figure 2E). In the other cases, there was no significant correlation between sample size and GO AUROC. These results suggested that networks from larger datasets potentially captured more biologically relevant GO terms. Although including more edges in networks (at less stringent confidence thresholds) clearly increased GO AUROCs, networks constructed with larger datasets required less edges to reach a similar GO AUROC. Taking the randomly sampled *A. thaliana* microarray dataset as an example, a high GO AUROC of 0.65 was obtained on average with 781,482 edges for matrices with 25 samples and 367,830 for matrices with 100 samples (Supplementary Fig2). This trend was less contrasted for *S. lycopersicum* (for a GO AUROC of 0.60, 547,494 edges with 25 samples vs 269,550 with 200 samples) and *Z. mays* (for a GO AUROC of 0.60, 192,261 edges with 25 samples vs 84,795 with 400 samples) datasets. In every case, strong significant effects of sample size and edge number on GO AUROC were observed (Supplementary Figure 2), indicating that smaller datasets might generate GO AUROCs as high as larger datasets by increasing the edge number. This likely indicated that the best associations found in smaller datasets were either false positives or new and transient associations which did not correspond to known GO associations.

We next compared GO AUROCs of networks constructed from datasets with more than 20 and less than 75 samples to evaluate the 3 down-sampling methods. Within this size range, we ensured that networks deriving from each method were comparable in terms of initial dataset size (Figure 2B). It revealed significantly higher GO AUROCs for randomly sampled networks but no difference between project-grouped or k-means clustered samples (Figure 3A). This indicated that using PCC-HRR with randomized matrices may be more informative than using thematically related samples. One would have expected that reducing complexity in expression matrices by combining related or similar samples might improve correlations between genes. To verify this hypothesis, we calculated Spearmans *rho* correlations between samples for each matrix. We found a weak negative correlation between sample correlations and GO AUROC of the resulting networks (Figure 3B). Samples clustered per project or by k-means were in average significantly more correlated (0.92 and 0.85 respectively, calculated all data combined) than those selected randomly (0.76; Wilcoxon rank sum test, *p*-value<2.2e-16). Taken together, these results indicated that calculating a simple correlation between samples of a given dataset can be useful to partially predict performance of the resulting network. Data presented in Figure 3B also suggest that datasets with very weakly correlated samples (*e.g.*, with a Spearmans *rho*<0.6) should be associated with lower GO AUROC but this remains to be demonstrated.

At a 1 million edges, microarray based networks globally performed better than those based on RNA-seq (Figure 4). This trend was less clear for counts of significantly enriched GO terms (Supplementary Figure 3). For *S. lycopersicum* and *Z. mays*, networks derived from RNA-seq had significantly more enriched GO terms than those derived from microarrays while GO AUROCs were higher for microarray datasets. This was probably due to the incomplete transcriptome representation in microarrays for these two species (Table 1). While GO AUROCs were high for genes analyzed by microarrays, these genes inherently represent less different processes than those measured by RNA-seq. It also suggests that genes pairs with a same GO term are more likely to be direct neighbors in the microarray than in the RNA-seq networks. A weaker predictability for RNA-seq derived networks was surprising. Because networks were obtained at a same number of edges, it indicated that many edges were not represented into the GO reference annotation. These edges could be considered either as false positive or true interactions not captured in the current GO reference annotation. It is possible that the more exhaustive view with RNA-seq encompasses genes which associations could be false positives or yet unaccounted true associations resulting in a decreasing in RNA-seq network predictability.

Aggregating networks

Networks obtained from smaller datasets generally had lower GO AUROCs than those with larger ones, but they could theoretically capture transient or local associations hidden in larger datasets. To allow networks obtained from small datasets to highlight both transient and more conserved gene associations, we analyzed the performance of their aggregation. Each aggregate generates a new network also named aggregated network. We compared aggregated networks from smaller datasets with those obtained from full datasets. Aggregation was expected to increase the global GO AUROC when using small datasets to construct networks. To aggregate individual networks, edge lists are combined and redundant edges are collapsed giving them a new score (Figure 5). This score is based on either the number of occurrences (therefore depending on the number of networks in the aggregate) or on the lowest HRR value. The first method is based on edge co-occurrence (CO) and considers most represented gene pairs as more robust than those found in only one network. Using the lowest HRR as a weight considers gene pairs with low HRR as significant, even if it is found in only one network. In this case, local associations are expected to be captured more efficiently. During the aggregation process, we asked whether all networks or only some of them should be considered. Inspired by a previous study(1), we tested complete and partial aggregations containing either the 50% highest GO AUROCs or the 50% lowest GO AUROCs. For these 2 types of aggregation, best edges were retained according to their lowest HRR value. A total of 4 aggregation methods were thus investigated: co-occurrence (CO), HRR-based (H), HRR-based 50% highest GO AUROCs (H-HGA, HRR-based aggregation Highest GO AUROC) and HRR-based 50% lowest GO AUROCs (H-LGA, HRR-based aggregation Lowest GO AUROC). Once redundant edges are collapsed, the 1M best gene pairs (either by their CO or by their HRR value) are retained for further network characterization.

Aggregating networks from project or K-means partitioned data subsets We first investigated edge conservation among aggregation methods. Correlation between HRR values and number of co-occurrences of each gene pair was low and

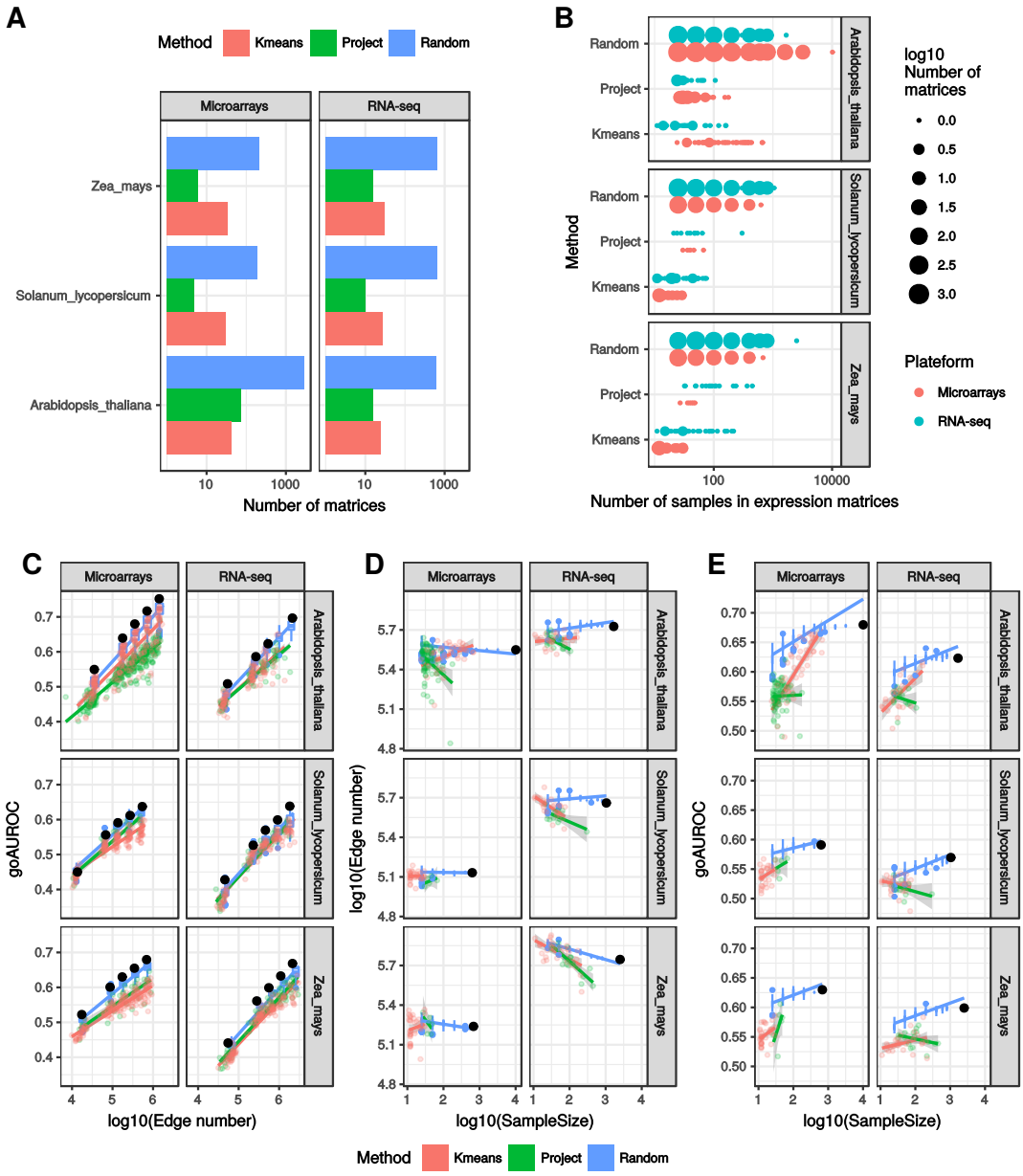

Figure 2. Performance of down-sampled expression matrices in capturing GO associations. Microarray and RNA-seq expression matrices from three species, *Arabidopsis thaliana*, *Solanum lycopersicum* and *Zea mays*, were prepared by combining all available RNA-seq data and subsequently down-sampled in three different ways, (i) random sampling, (ii) by project and (iii) by sample clustered by a k-means analysis, to construct networks at different confidence thresholds (1, 5, 10, 20 and 40% of the best co-expressed transcripts). The total number of down-sampled expression matrices is shown in A and the number of samples per table in B. In B, the highest number of samples in the Random category correspond to the full matrix. Pairwise relationships between GO AUROC, edge number and sample size are shown in C for all thresholds and in D and E at a threshold of 0.1. For networks obtained from randomly sampled expression matrices, data are summarized as boxplots (in blue). Lines correspond to regression lines with their 95% confidence interval in grey areas. Black dots correspond to data for networks inferred from full datasets.

generally negative (Figure 6A, maximum *rho*=0.15 , average *rho*<-0.089), indicating that most frequent gene pairs had uncorrelated HRR values. This explained that only partial overlaps were found between network aggregates obtained with the two methods (H and CO) (Figure 6B). For microarray based-networks, aggregates obtained by co-occurring edges shared between 30 and 50% of edges with those obtained by the minimal HRR value. For RNA-seq based networks, common edges represented generally between 10 and 20%. Contrastingly, there was a higher number of common edges between HRR-based and H-HGA or L-LGA aggregated network while less than 10% were conserved between H-HGA and L-LGA based aggregates (Figure 6B). This indicated that HRR-based aggregates contained best edges from both best performing and worst performing networks.

We next measured aggregate performance in capturing GO associations. Networks derived from full matrices and CO aggregates had statistically higher GO AUROCs than single

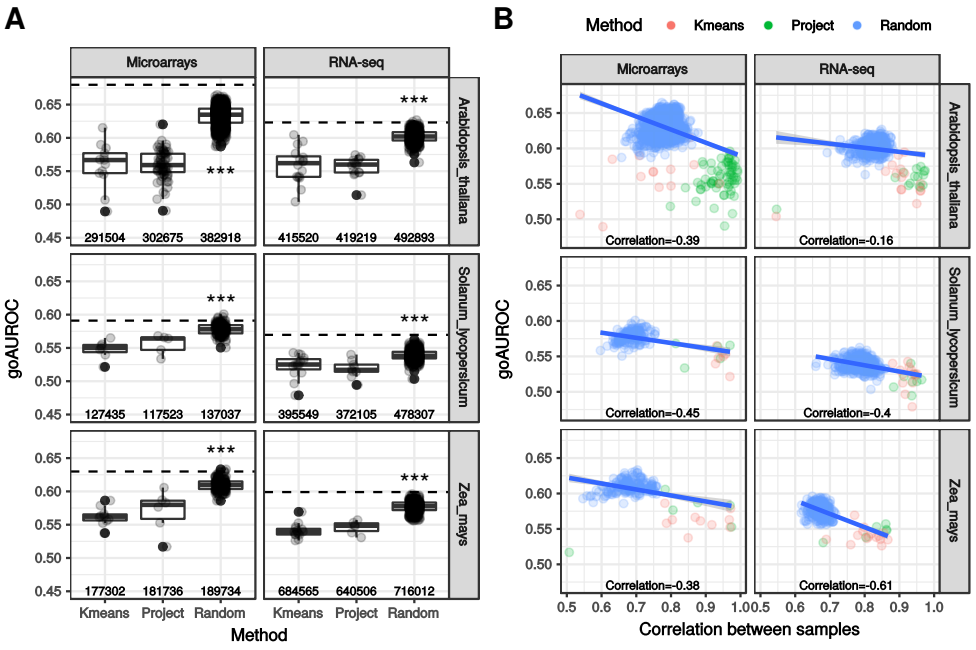

Figure 3. Comparison between down-sampling methods. (A) GO AUROCs of networks obtained at a threshold of 0.1 from datasets with more than 20 and less than 75 samples are summarized as boxplots for each down-sampling method. Asterisks show significant difference in median between randomly sampled datasets and the two other methods (Wilcoxon rank sum test, p<0.001). Average edge numbers are indicated below boxplots. (B) For each down-sampled matrix, correlation between samples was calculated (Spearmans *rho*) and plotted against the GO AUROC of the resulting network. Correlation between the two variables was calculated with the Pearson coefficient. Blue lines correspond to regression lines with their 95% confidence interval in grey areas. Dashed lines correspond to data for networks inferred from full datasets.

networks (Wilcoxon rank sum test, *p*-value<0.01) (Figure 6C). Although statistical differences in GO AUROCs between co-occurrence aggregates and individual networks were not confirmed by counts of significantly enriched GO terms (Supplementary Figure 4), it was likely that CO aggregates displayed the highest performance and were at least as efficient as networks derived from full matrices for both performance measures. All data combined, only networks aggregated by co-occurrence and 50% lowest GO AUROC (H-LGA) resulted in significantly different GO AUROCs (Figure 6D). We observed a substantial but not significant improvement in GO AUROCs between aggregates containing networks with the 50% best GO AUROCs or the 50% lowest GO AUROCs suggesting that a prior selection of networks to combine could substantially improve GO term recovery. No significant difference in performance was observed between aggregates of project or k-means clustered datasets, although in almost all cases the GO AUROC of the CO aggregates of k-means networks was higher than that of project aggregates.

To further characterize the different aggregates, we subsequently analyzed 13 biological pathways from the Reactome database in Pathway Level Coexpression (PLCs(23)) subnetworks. Pathway reconstruction was evaluated according to guide gene partitioning into specific communities and biological processes associated within these communities (Supplementary Table 2). PLC size was set at a 1,000 vertices (1,000 genes best co-expressed with the query guide genes) obtained from the 1 million best edges. Over the 13 biological pathways, PLC from the full dataset networks or aggregated networks according to edge co-occurrence appeared to better capture GO terms as reflected by high GO AUROC and counts of significantly enriched GO terms (Wilcoxon rank sum test, *p*-value<2e-16)(Supplementary Figure 5). No significant difference was observed among the three remaining aggregation methods (HRR-based, H-HGA and H-LGA). It is noteworthy that all aggregated networks from a same species x platform combination had the same number of edges (set at a 1 million) and almost the same number of vertices. Although PLC size was set at a 1,000 vertices, PLC networks had considerably variable edge numbers. In particular, PLCs based on co-occurrence aggregates had many more edges resulting in a higher mean node degree. As indicated by their higher GO AUROCs, these edges are likely to correspond to biologically relevant gene associations.

By contrast, networks aggregated according to edge HRR values had a lower mean node degree and the 1,000 vertices needed to construct the PLC were reached with a few edges, revealing that each node is connected to a few other ones. Concerning the pathway reconstruction quality, guide gene distribution into communities better matched the expected partition for the three best HRR based methods than for the co-occurrence one as revealed by their higher normalized Chi-squared values (Wilcoxon rank sum test, *p*-value<¡0.01)(Supplementary Figure 5). All data combined, normalized Chi-squared values were indeed the lowest in the co-occurrence aggregate (*p*-value<1e-05). This highlighted a trade-off between GO capture and pathway reconstruction.

It was likely that capturing gene associations annotated with GO terms was optimal with the most represented edges,

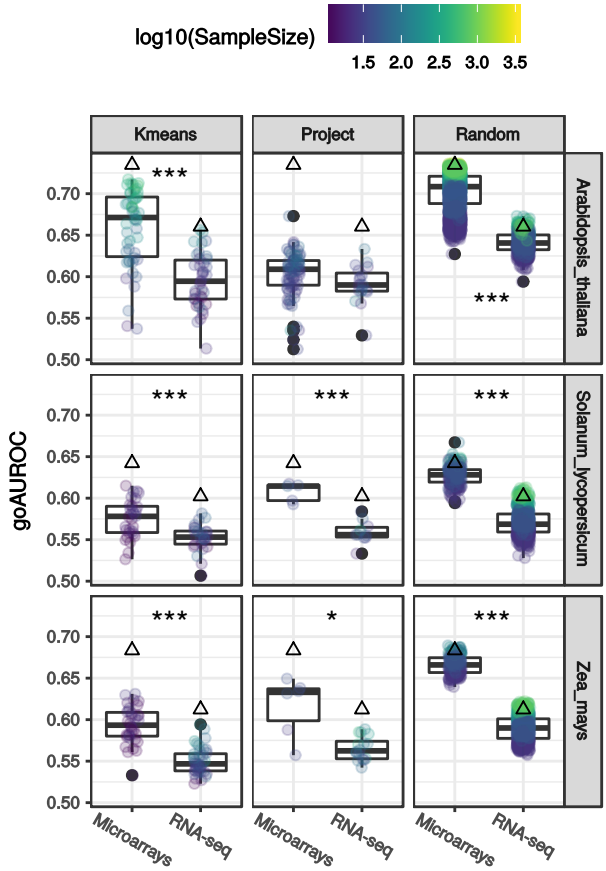

Figure 4. Performance comparison between microarray and RNA-seq. The performance of networks with a 1 million edge to capture GO terms was measured with GO AUROC. Asterisks denote a significant difference between the two platforms (Students *t* test, *, *p*-value<0.05, **, *p*-value<0.01, ***, *p*-value<0.001). Each point represent one individual network and boxplots summarize data all sample sizes confounded. White triangles correspond to data for networks inferred from full datasets.

while correctly associating genes from a same pathway requires transient and/or unknown associations with other genes. PLC network topology was evaluated by calculating the clustering coefficient which measures a probability that a given node is connected to other nodes in the network. This coefficient indicates the module structure of the network(24). It clearly appeared that the HRR-based aggregation procedure resulted in significantly lower clustering coefficients than co-occurrence aggregated and full dataset derived PLC networks (Wilcoxon rank sum test, *p*-value<1e-05)(Supplementary Figure 5). In accordance with this, degree distribution of these two kinds of PLC networks also had a better fit to the power law as revealed by the calculated log likelihood (Supplementary Figure 5). Taken together, these results suggest that for the full dataset or edge co-occurrence aggregates, partitions of guide genes into communities did not strongly reflect the expected partitions in pathways but their higher clustering coefficients revealed a more modular structure (see Supplementary Figure 6 for an example with the secondary metabolite Reactome pathway). This might be in turn explained by interdependencies between sub-pathways (guide genes that could be theoretically simultaneously found in several communities) as well as incomplete pathways in the database. Concerning HRR value based aggregation methods (H, H-HGA and H-LGA), project network aggregates had globally better metrics than those using datasets with k-means clustered samples (Supplementary Figure 5). However, these aggregates never outperformed CO aggregates or full dataset networks. On the whole, PLC networks from full dataset had very good features in terms of GO AUROC, modularity, normalized Chi-squared and clustering coefficient.

During the aggregation process, we were also interested in aggregating microarray and RNA-seq based networks. Following the co-occurrence principle, we considered that gene pairs highlighted in networks deriving from different technologies may be more robust. We therefore determined co-occurring genes between PLCs obtained with the two technologies and analyzed intersection networks for each aggregation method. We found that PLCs from co-occurrence aggregates displayed the highest degree of conservation between microarrays and RNA-seq as revealed by the number of co-occurring edges and vertices (Figure 7A). In fact, the other aggregation methods did not allow to correctly align microarray and RNA-seq PLC networks. Microarray and RNA-seq intersection network of HRR aggregates had very few edges and vertices as exemplified with the secondary

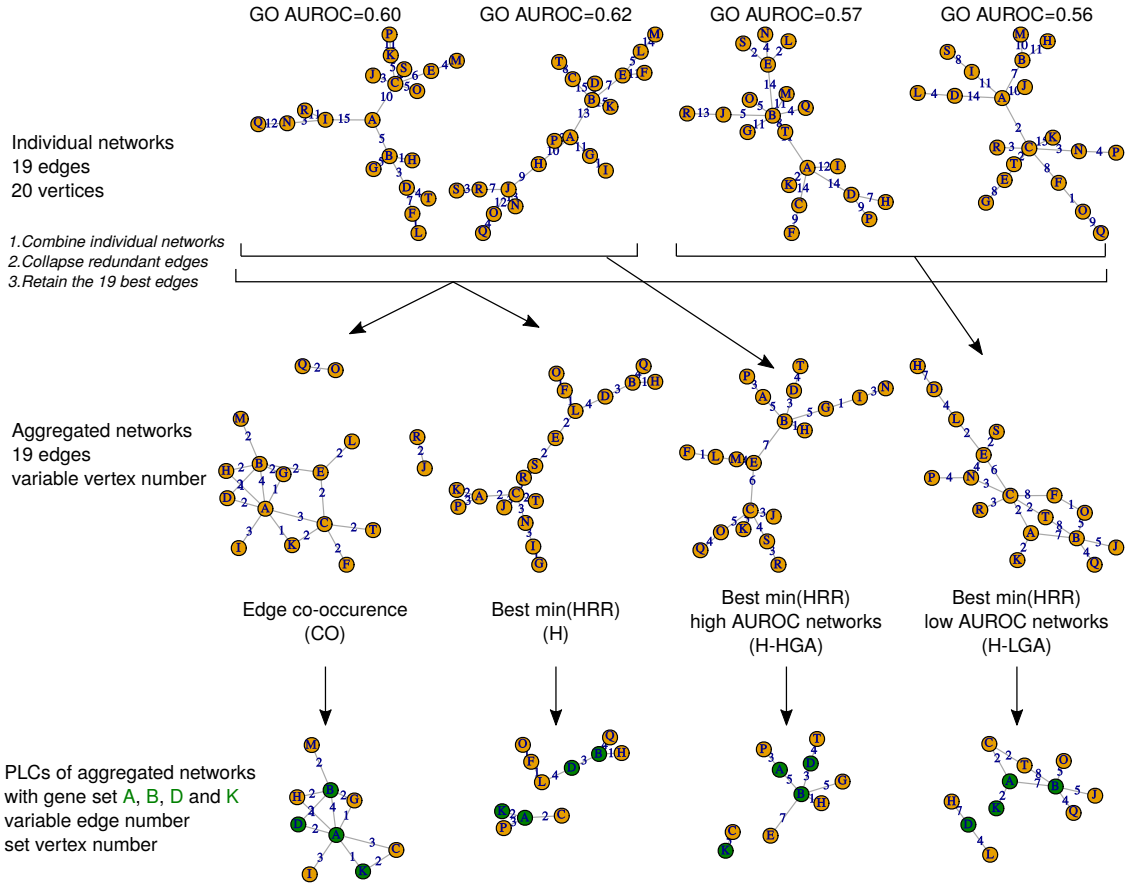

Figure 5. Network aggregation procedures. In this example, single networks had 20 vertices (named with A to T) and 19 edges (numbers indicate a hypothetical HRR value sampled with replacement between 1 and 15) and were generated according to the model of Barabasi and Alberts(38). Aggregation was either total or partial. Total aggregation resulted in 20 vertices and 57 non redundant edges. The 19 best pairs were retrieved either by taking the best HRR (H) or the most co-occurring (CO) edges. For partial aggregation, we combined either the 50% of networks with the highest GO AUROCs (HGA) or 50% with the lowest GO AUROCs (LGA). For these two partial aggregations, the 19 best pairs were retrieved by taking the best HRR (H-HGA or H-LGA).

metabolite pathway in *S. lycopersicum*(Figure 7B). This reinforces the observation that gene pairs with lowest HRR did not frequently co-occur. Combining microarray and RNA-seq aggregates resulted in a substantial drop of GO AUROC in contrast to technology specific aggregates (an average of 0.54 vs 0.57 respectively, p-value$<$0.001). This was explained by the loss of vertices and edges during the combination of aggregates. Concerning data partitioning methods, no significant difference was observed between k-means and project on GO AUROC.

Taken together, these results suggest that, according the reference annotation sets used here, aggregating networks using the edge co-occurrence method is useful to improve the capture of relevant gene pairs. Final networks containing edges found in both microarray and RNA-seq based networks are likely to provide an appropriate transcriptional map of a given pathway at the expense of a slightly lower GO AUROC.

Aggregating networks from randomly sampled data subsets. Data subsets obtained by random combinations of samples resulted in networks performing better than those originating from subsets corresponding to project or k-means partitions of the initial dataset (Figure 3A). Given the large sample size in each initial matrix (Table 1), aggregation possibilities were numerous. We evaluated different aggregate sizes with multiple sample combinations for each and we aggregated networks according to the size of the datasets they originated from. Aggregate performance was evaluated with GO AUROC and the number of significantly enriched GO terms. Aggregates differed by both the number of networks they contained (x-axis on Figure 8), the combination of networks included (error bar on y-axis on Figure 8 and by the sample size of the initial dataset used to generate individual networks (different colors on Figure 8). Plotting mean GO AUROCs of aggregates obtained from different sample combinations revealed an overall low variation whatever the aggregate size or the initial dataset sample size (Figure 8). Largest variations in GO AUROC were observed for aggregates of networks obtained with datasets having 25 and 50 samples, especially in *S. lycopersicum* microarray networks. This suggests that edge content differed substantially between different combinations of a same aggregate size. This was relatively unexpected for *S. lycopersicum* microarray based aggregates because the corresponding arrays contain less than 7,000 genes which

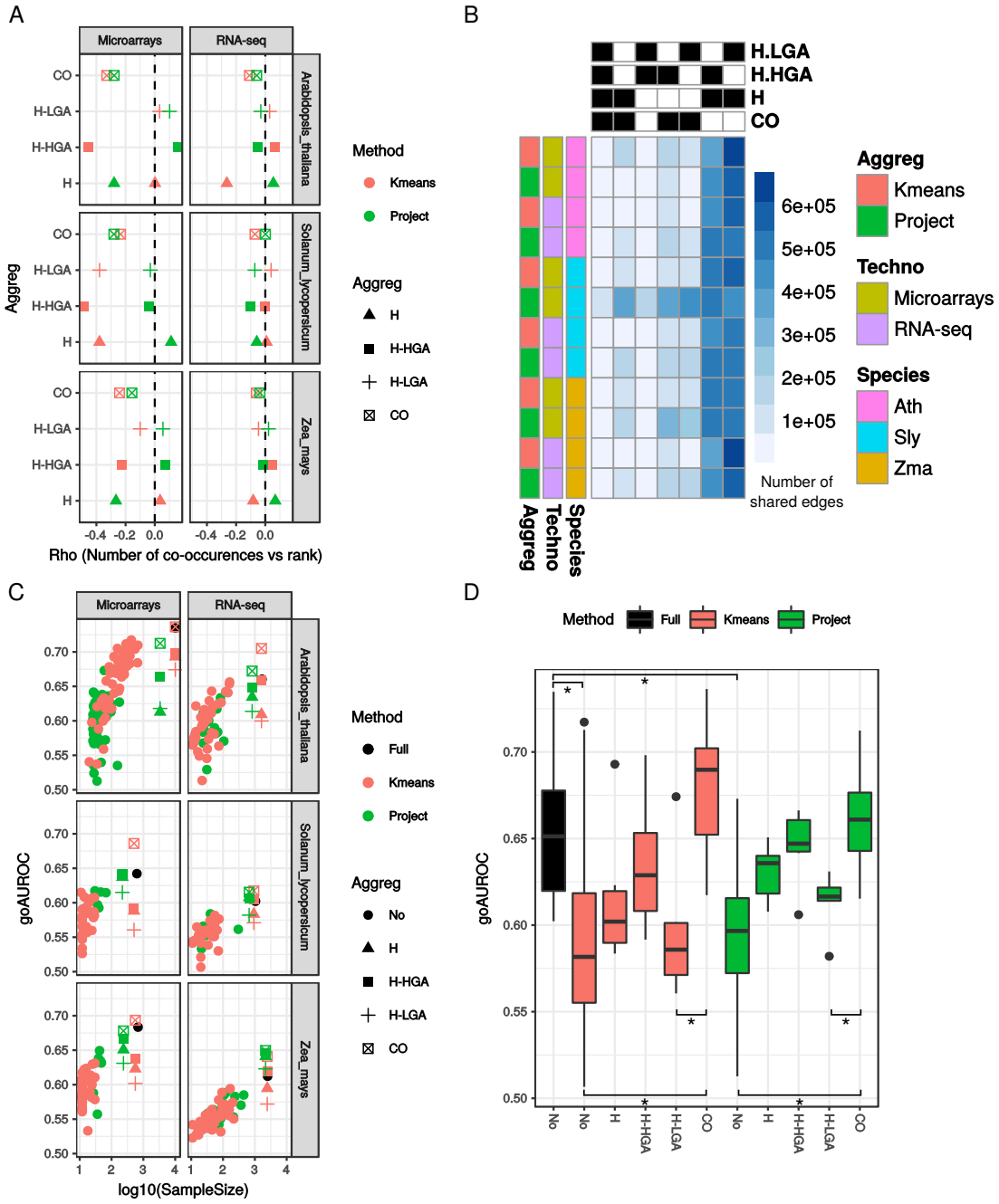

Figure 6. Aggregated networks (1 M edges) from expression matrices down-sampled by grouping samples by their project or by clustering them by k-means. Aggregation was either total or partial. For total aggregation, the 1 M best pairs were retrieved either by taking the best HRR (H) or the most co-occurring (CO) edges. For partial aggregation, we combined either the 50% of networks with the highest GO AUROCs (HGA) or 50% with the lowest GO AUROCs (LGA). For the two partial aggregations, the 1 million best pairs were retrieved by taking the best HRR (H-HGA or H-LGA). (A) For each 4 aggregation methods and each dataset, correlation (Spearmans *rho*) between gene pair HRR value and number of co-occurrence was calculated. (B) Number of edges found among aggregates obtained by different methods. Black and white boxes for columns respectively indicate that an aggregate is included or not in the comparison. (C) Network performance measured by GO AUROC. Single non aggregated (No) networks with 1 million edges are also reported. (D) Data from all species and platforms summarized in boxplots. Asterisks denotes significant differences between two procedures (Wilcoxon rank sum test, p-value<0.05).

might have led to more homogeneous edges. This higher heterogeneity shows that correlations are largely impacted by the starting dataset when using partially represented transcriptomes.

Networks aggregated according to edge HRR values rarely outperformed the network from the full dataset. Edge CO aggregation generally improved GO AUROC for RNA-seq based networks but the effect was less visible for microarray data. Larger dataset sizes generally decreased differences

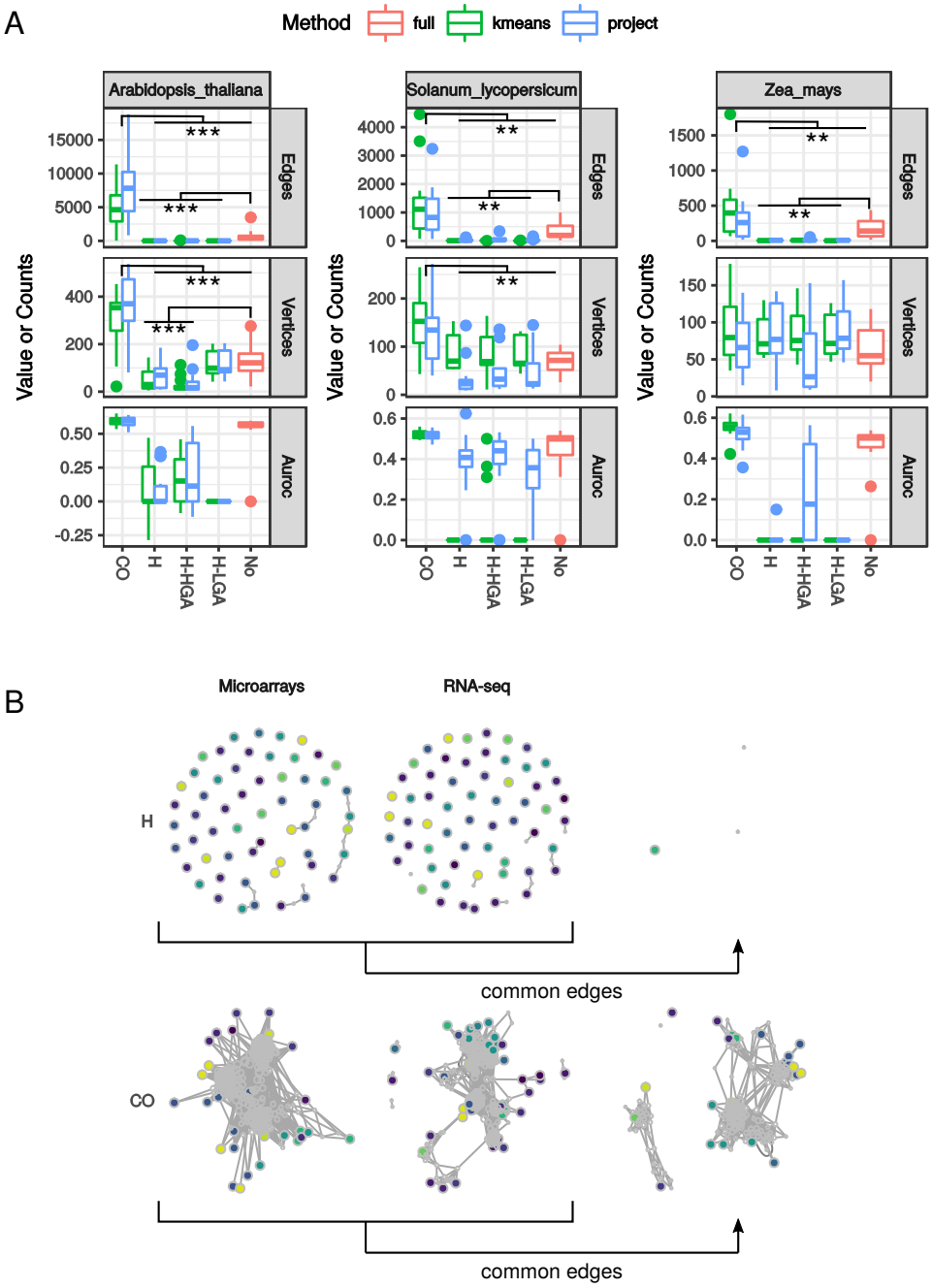

Figure 7. Characteristics of merged Pathway Level Coexpression (PLC) obtained with microarray and RNA-seq data. For total aggregation, the 1 M best pairs were retrieved either by taking the best HRR (H) or the most co-occurring (CO) edges. For partial aggregation, we combined either the 50% of networks with the highest GO AUROCs (HGA) or 50% with the lowest GO AUROCs (LGA). For these two partial aggregations, the 1 million best pairs were retrieved by taking the best HRR (H-HGA or H-LGA). No aggregation indicates that networks from the full datasets were compared. Network size and performance measured with GO AUROC are presented in (A). Asterisks indicate significant differences (Wilcoxon rank sum test, **, p-value$<$0.01; ***, p-value$<$0.001). In (B) are shown two examples with *Solanum lycopersicum* and the secondary metabolite pathway. Microarray and RNA-seq networks were obtained by aggregating networks from k-means partitioned data subsets with either the CO (edge co-occurrence) or the H (best HRR) method. The third row displays the edges and vertices common to microarray and RNA-seq deriving PLCs.

between aggregation methods, especially when few networks ($<$10) are combined, suggesting that networks constructed from large dataset sizes (*e.g.* $>$50) are more robust. GO AUROC was increased with an average of 0.02 and a maximum of 0.05 for *A. thaliana* RNA-seq when using the CO method (Supplementary Figure 7). Aggregation of networks from randomly selected samples increased the number of significantly enriched GO terms on average in 21 and 182 for

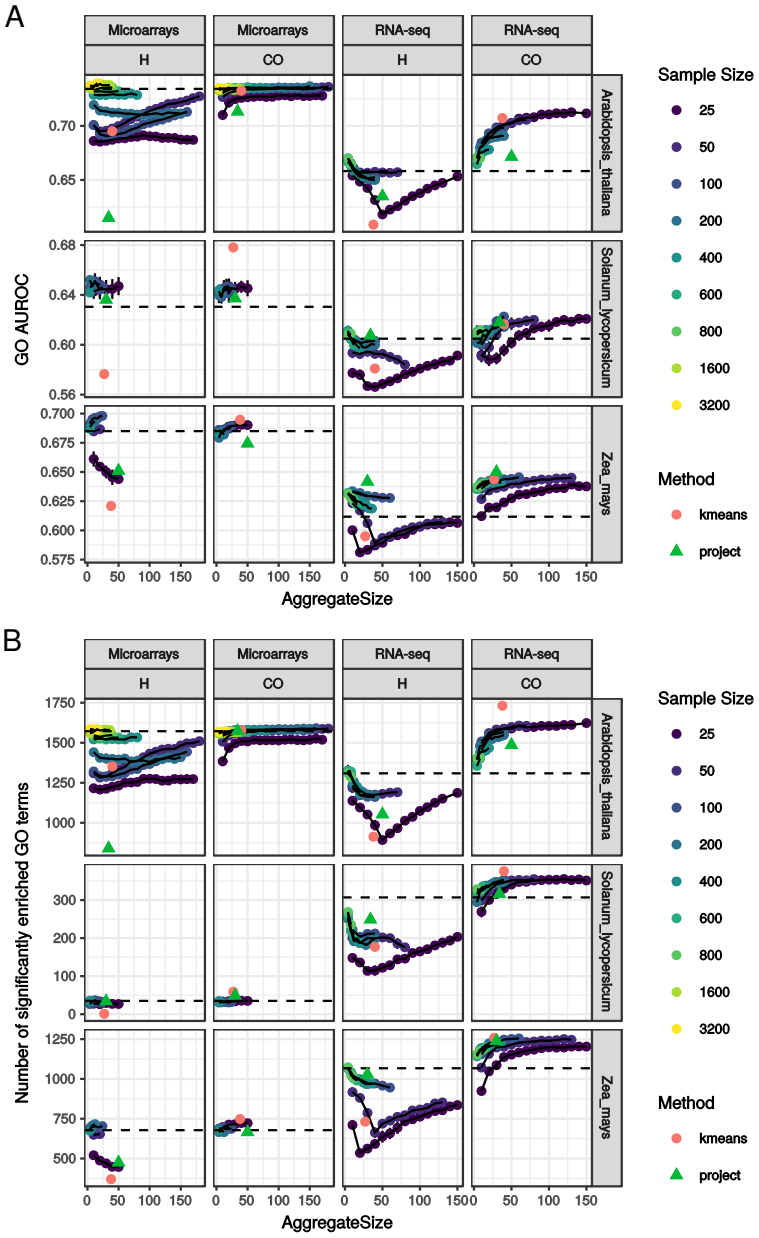

Figure 8. Aggregate performance (A, GO AUROC; B, counts of significantly enriched GO terms) with networks obtained from randomly sampled data subsets. Aggregate size was set at a 1 million edges. Only two aggregation methods are shown. Redundant edges in the aggregate were collapsed according to their co-occurrence (CO) in several individual networks or according to their best HRR value (H). Horizontal dashed lines indicate the number of GO terms for networks obtained with the full datasets. All points correspond to average measures over replicates and vertical black bars to standard deviation. Pink points and green triangles indicate GO AUROCs of networks aggregates using partitioned datasets with k-means or by project respectively.

microarrays and RNA-seq respectively, with a minimum of 2 for *S. lycopersicum* microarrays and a maximum of 314 for *A. thaliana* RNA-seq in comparison to the full dataset derived networks (Figure 8, Supplementary Figure 7).

As a general trend, including more networks with the CO aggregation method helped to increase GO AUROC (Spearmans *rho*>0.5, *p*-value<0.05) whatever the initial sample size until a plateau was reached (Figure 8). Using larger sample sizes had contrasting effects on GO AUROCs according to the initial dataset considered, increasing GO AUROCs for *Z. mays* and *S. lycopersicum* RNA-seq aggregates (*rho*>0.75, *p*-value<1e-05) but decreasing it for *A. thaliana* RNA-seq aggregates (*rho*=-0.80, *p*-value<1e-09). When looking at aggregate size x dataset size combination maximizing GO AUROC, the CO method performed better with more aggregates of networks from small datasets (Supplementary Figure 7). Contrastingly, the HRR method required less aggregates but larger dataset sizes.

The aggregates of networks inferred from k-mean partitioned datasets performed generally at least as good as the best random aggregates (Figure 8). This trend was even more pronounced when considering the number of significantly

enriched GO terms. We examined whether it was also the case for PLC (Supplementary Figure 8). PLC performance varied considerably according to the pathway considered as shown by min and max pathway metrics in Supplementary Figure 8. This was probably due to the important differences in Reactome pathway sizes and compositions (Supplementary Table 2). The CO aggregation method was the best to improve GO term recovery in comparison to full dataset networks (+0.014 vs -0.03 for the HRR method for GO AUROC; +23 vs -54 for the number of significantly enriched GO terms, Welch two sample *t*-test *p*-value<2e-16 for both parameters) but modularity and normalized Chi-squared statistic were higher when the aggregation was based on HRR (modularity: 0.58 for CO vs 0.66 for HRR; normalized Chi-squared statistic: 0.17 vs 0.25, Welch two sample *t*-test *p*-value<2e-16 for both parameters) (Supplementary Figure 8). PLC GO AUROC was only significantly influenced by aggregate size, not by the initial sample size (ANOVA, *p*-value<0.00573, expected for *Z. mays* microarray based networks), while counts of significantly enriched GO terms were influenced by aggregate size for RNA-seq based networks only (ANOVA, *p*-value<0.0206). GO capture in PLC with aggregates of networks from randomized datasets could therefore be maximized by using small dataset and large aggregate sizes. Differences between randomly sampled datasets or partitioned datasets in GO term capture varied according to the species and the platform. On the whole, aggregates of networks constructed from k-mean partitions had a higher number of significantly enriched GO terms for RNA-seq as well as a higher average GO AUROC (excepted for Zea mays) (pairwise *t*-tests *p*-value<0.05). This was true for small aggregate sizes, because in all cases, larger aggregate sizes (obtained with smaller dataset sizes, *e.g.* 25 or 50) generated PLCs with GO capture performance similar to k-means partitioned dataset network aggregates.

Concerning topological metrics, modularity was not significantly influenced by either aggregate size or sample size in *S. lycopersicum* and *Z. mays* microarray based aggregates (ANOVA, *p*-value>0.05) (Supplementary Figure 8). Strongest effects were observed for RNA-seq based aggregates, with larger sample sizes correlating with higher modularity (Spearmans *rho*>0.61, *p*-value<2e-16). A very similar observation was made for the clustering coefficient, although CO based aggregates had higher values than HRR based aggregates (average of 0.58 for CO, 0.22 for HRR; *t*-test, *p*-value<2e-16). For all species x platform combinations, PLCs had a significantly higher log likelihood of fitting to a power law when aggregated with the CO method (*t*-test, *p*-value<0.03). These topological metrics suggest that CO based aggregates have a more clustered structure with hubs corresponding to few genes having many connections with other, as exemplified with the Fatty Acid pathway in Supplementary Figure 9. Partitioning guide genes into communities was similar to expected partitions when using the HRR aggregation method (average of 0.25 for HRR and 0.16 for CO; *t*-test *p*-value<2e-16). This was mainly due to the higher number of guide genes represented in HRR based aggregates (average of 51 for HRR aggregates vs 38 for CO aggregates; *t*-test *p*-value<2e-16). However, the lower clustering coefficients of HRR-based aggregates showed a less evident interpretation of the communities (Supplementary Figure 9). With the exception of *A. thaliana* microarray based networks, the clustering coefficient was higher in aggregates of randomly selected samples than in aggregates of partitioned datasets (Supplementary Figure 8; *t*-test *p*-value<0.014). Although aggregates of networks constructed with k-means partitioned datasets had a very good performance in GO term capture, the PLC topology may be less biologically relevant, such as exemplified in *Z. mays* RNA-seq based k-means aggregates vs random aggregate (Supplementary Figure 9).

Application to jasmonic acid (JA) biosynthesis in plants

To further demonstrate why aggregating networks is useful to study biological pathways, we extracted PLC from RNA-seq networks containing a 1 million edges with genes involved in the biosynthesis of JA (9). For this example, we focused on three networks for each species: (i) a network inferred from the full RNA-seq dataset, (ii) a co-occurrence (CO) aggregate of individual networks inferred from the k-means partitioned full dataset and (iii) a HRR-value based aggregate of individual networks inferred from the k-means partitioned full dataset. Each 1 million edge network (a total of 9, 3 construction modes in 3 species) was queried with guide genes (GG) obtained from the Plant Metabolic Network databases(10) (26 for *A. thaliana*, 44 for *S. lycopersicum* and 33 for *Z. mays*) and included lipoxygenases (LOX), allene oxide synthases (AOS) and cyclases (AOC) among others (Supplementary Table 3). Among the 1 million edge networks of *A. thaliana*, *S. lycopersicum* and *Z. mays*, 2,050, 575 and 1,935 edges on average contained one guide gene respectively. It was surprising that although more guide genes were considered for *S. lycopersicum*, the edge number was lower than in the two other species, the maximum edge number (863) being observed for the full dataset derived network. The n best edges involving at least one guide gene (having the lowest HRR values or the highest CO) in each network were retained so that the final PLCs contain *ca.* 30 vertices (Figure 9A; Supplementary Table 4). This cut-off value was chosen for three reasons: (i) to allow comparisons among networks and among species, (ii) it represented no more than 10% of the vertices found in PLCs performed on the 1 million edge networks and (iii) in CO or H aggregates, it retained edges with a CO weight >2 (*i.e.*, each edges was found in at least two individual network before aggregation) or a HRR value <10 respectively. This HRR value is a stringent threshold(39). Hence, the 9 resulting PLCs (3 construction modes for 3 species) with 30 vertices are expected to be high confidence closely focused on JA metabolism and signaling.

Genes with evident relationships such as TIFY transcription factors, known modulators of the JA signaling pathway(22), were classified as associated genes (AG) while those with no previously described relationship to this pathway were classified as other genes (OG) (Figure 9B). Within each species, the number of GG retained in the 30-vertices PLCs was similar but they differed among networks. In fact, gene content strongly differed between PLCs because less than 20% of the genes were conserved in a same species between each pairwise PLC comparison, showing the strong impact of the construction procedure on the resulting PLCs. Although differences between numbers of AG among aggregation methods were not statistically significant, CO

aggregates of networks inferred from k-means partitioned datasets appeared to contain more AG (11.3 in average over the 3 species) than networks inferred from full datasets (3.6 in average) or from H aggregates (3.3 in average). Mapping gene accessions to the Uniprot databases clearly revealed that CO aggregated networks contained more genes related to response to wounding for *A. thaliana* and *S. lycopersicum* (Supplementary Table 4).

In *S. lycopersicum* and *Z. mays*, CO aggregates contained JA biosynthesis related genes not included as guide genes and not found in the other networks, suggesting that these aggregates were likely to give a more exhaustive picture of transcriptional relationship within the JA biosynthesis and signaling than the other construction methods. It was the case for Solyc07g007870.2 encoding a 12-oxophytodieonic reductase (OPR3; found in the community containing Solyc04g079730.1 corresponding to an AOS) and GRMZM2G136857 encoding a putative JA methyltransferase (in the community containing GRMZM2G104843 corresponding to a LOX). For the two species, patatin-like proteins (Solyc04g079250.2 and GRMZM2G154523) potentially involved in releasing linolenic acid from membrane phospholipids as a JA biosynthetic precursor were detected in the CO aggregates(9). This highlights CO aggregate ability to capture close relationships from biological pathways and to prioritize candidate genes to be functionally tested. The present CO aggregates also indicate that unexpected functions might be associated with the JA biosynthesis or signaling. For example, different hydrolases were found in the transcriptional neighborhood of the guide genes (AT1G31550 encoding a GDSL esterase/lipase, Solyc03g083010.2 encoding a alpha/beta fold hydrolase and GRMZM2G032160 encoding a glycoside hydrolase). Together with the presence of amino-acid metabolism related genes (such as AT4G08870 encoding an arginase and Solyc03g013160.2 encoding an amino-acid transporter), these candidate genes may reveal unexpected but key connections of the JA pathway with primary metabolism.

In *A. thaliana*, the CO aggregate contained 5 TIFY, 2 WRKY as well as the basic helix-loop-helix (bHLH) AtMYC2 transcription factors, families containing known JA signaling regulating factors(26, 30). Similarly, the CO aggregate of *S. lycopersicum* contained 4 TIFY, 2 bHLH and 1 WRKY. For *Z. mays*, the community containing a lipoxygenase encoded by GRMZM2G104843 did not display significant GO term enrichment but included several transcription factors with bHLH (2), TIFY (2) or WRKY (1) domains suggesting they could be the functional orthologs of the two other species. For example, transcripts for WRKY40 (AT1G80840 and Solyc03g116890.2) were also found in *Z. mays* (GRMZM2G120320), suggesting this would be the Z. mays functional WRKY40 ortholog. This was confirmed by BlastP analysis against nr NCBI database in which Solyc03g116890.2 and GRMZM2G120320 had best similarities with AT1G80840 (respective max score 261 and 188), although AT1G80840 had better homologies to Solyc06g068460.3 than to Solyc03g116890.2 (score 261 vs 241 with and Solyc03g116890.2) and to GRMZM2G111711 than to GRMZM2G120320 (score 185 vs 181 with GRMZM2G120320). This revealed that direct orthology would have probably failed at finding the correct functional WRKY factors whereas our network approach succeeded. These transcripts in *S. lycopersicum* and *Z. mays* are currently annotated as putative WRKY factor or uncharacterized protein in Uniprot, but their presence in the JA PLCs obtained here strongly suggests their involvement in JA signaling. Also of interest were transcripts encoding putative bHLH transcriptions factors found in CO aggregates of *S. lycopersicum* (Solyc09g083360.2) and *Z. mays* (GRMZM2G301089). Both of them had best similarities with *A. thaliana* bHLH92 (AT5G43650). The only bHLH transcription factor identified in CO aggregate of *A. thaliana* was MYC2 (AT1G32640). Whether orthologs of bHLH92 may play a similar role than MYC2 in the two other species deserve more investigations. On the whole, our networks clearly indicate an involvement of a few specific members from large families of transcription factors (on average: 231 bHLH, 25 TIFY and 111 WRKY for the 3 species as indicated in the PlantTFDB(13)), that would have not been as efficiently identified by other approaches.

DISCUSSION

Using gene co-expression is an efficient approach to predict a gene function or find candidate genes involved in a given pathway(25, 27, 28, 29, 31, 32). In many cases, the authors generate their own data and infer co-expression from them. It is also possible to re-use published data to visualize transcriptional relationships in large expression compendia, such as those available on ATTED II(33) and PlaNet(40) databases. In any case, it is not clear how many samples should be included and/or how related they should be to calculate distances between genes. Specific datasets may highlight specific correlations while sharing many transcriptional relationships(34). Whether a limited number of datasets is more appropriate than maximal datasets (containing all available samples) to capture gene transcriptional relationships remains to be determined.

Our results clearly showed that individual networks from down-sampled datasets (Figure 1) do not outperform networks obtained from the largest datasets (Figure 2), in accordance with previous studies(1, 2, 3, 4). For both microarray and RNA-seq, the more samples contained in the initial dataset, the higher the performance in capturing edges matching GO terms is expected to be. Single networks from randomly down-sampled data subsets had a higher GO AUROC than single networks obtained from project or k-means partitioned data subsets probably because samples in these latter were more correlated between themselves (Figure 3). In addition, it appeared that microarray data resulted in higher GO AUROC at equivalent sample sizes (Figure 4). Aggregation has already been used in co-expression analysis(7, 8), however without extensive testing of aggregation methods. We found a substantial performance gain when aggregating individual networks. Retaining edges co-occurring in several networks to construct the aggregate clearly improved recovering known gene pairs (Figure 6 and 8). This method allowed a good overlap between microarray and RNA-seq derived aggregates using project or k-means partitioned dataset networks (Figure 7). Such co-occurring edges in networks deriving from two technologies and differing initial sample sets could be considered as more robust. Taken altogether, these results

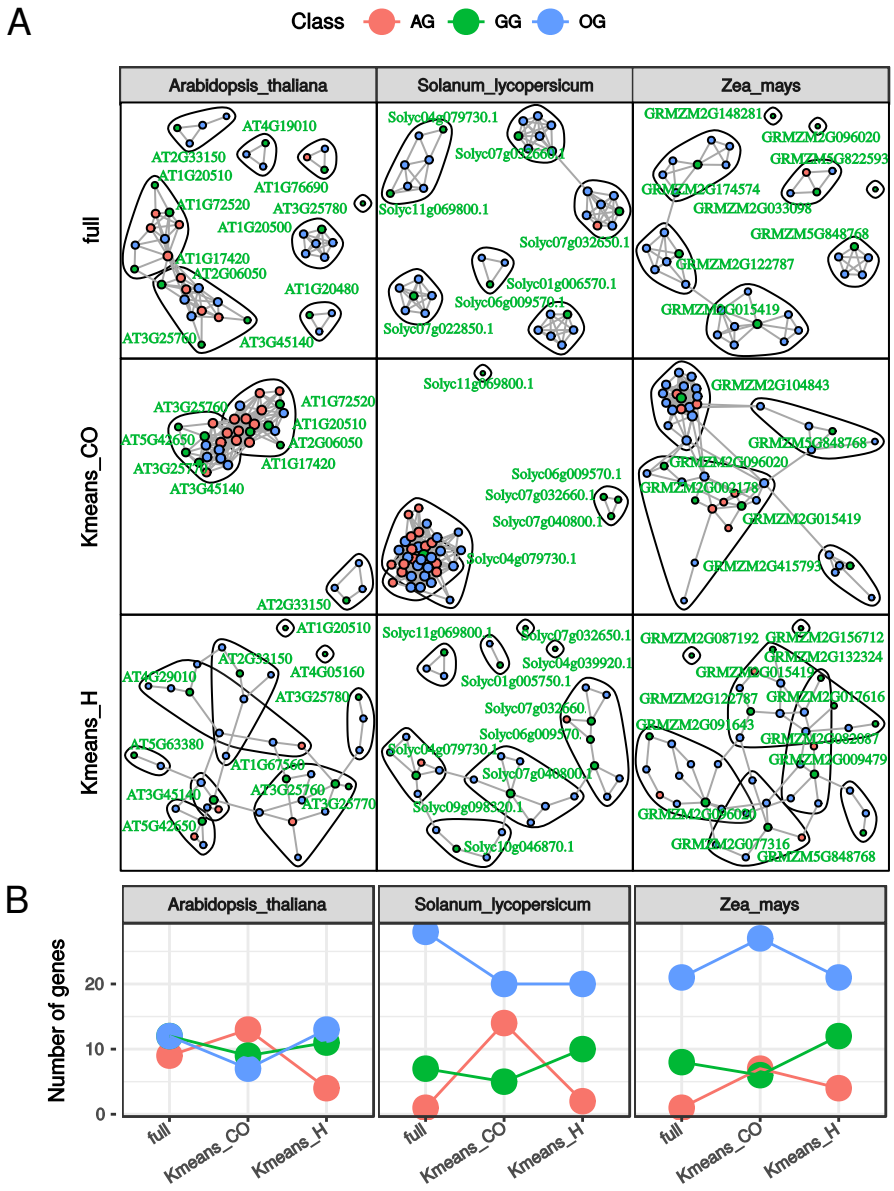

Figure 9. Pathway Level Coexpression (PLC) for the JA biosynthesis pathway in *Arabidopsis thaliana*, *Solanum lycopersicum* and *Zea mays*. Gene accessions were obtained from the Plant Metabolic Pathway Databases. PLCs were constructed by retrieving guide genes with their best co-expressed genes from large networks constructed from full RNA-seq datasets or from RNA-seq datasets partitioned with K-means and subsequently aggregated according to edge co-occurrence (CO) or edge HRR value (H). PLC size was set at 30 genes (A). Communities were detected with a fast greedy algorithm and are delimited by black polygons. Genes in PLC were classified as guide genes (GG, green, with their respective locus tags), associated genes (AG, non-guide genes with evident relationships to guide genes, red) or other genes (OG, non-guide genes without evident relationship to guide genes, blue) and counted in each PLC (B).

suggest that inferring networks from large compendia with PCC and HRR are biologically relevant for both microarray and RNA-seq data.

Aggregating individual networks constructed from randomly sampled datasets resulted in meta networks with satisfying performance, especially when using RNA-seq data where some aggregates performed much better than the network derived from the full dataset (Figure 8). However, aggregates of k-means partitioned datasets generally had a better GO AUROC than other aggregates. This indicates that the strategy proposed by Feltus et al(6) is very appropriate for both microarray and RNA-seq data. These networks appeared to highlight a wide range of transcriptional relationships (Figure 8). This was further confirmed by both GO AUROC and PLC topological metrics. In addition, results obtained for the k-means partitioned datasets appeared to be more robust than for randomly sampled datasets for which trends differed between species (Figure 8). Further experiments will be necessary to determine the minimal number of samples required for a k-means based partitioning.

In the present article, we did not test other distance calculation methods. Whether our results can be directly transposed to other more complex calculations (such as in supervised methods for regulatory network inference) remains to be determined. It is possible that the ranking procedure used here is tolerant to error and adapted to larger datasets. Mutual information based networks of *E. coli* expression data processed with different algorithms displayed similar trends during down-sampling(3). Down-sampling had also a similar impact on network build with Spearmans rank correlations(1). Although more investigations will be required when using other distance measurements, it is likely that similar results will be observed with other methods.

In our previous work(16), we showed that PLC quality could be monitored with the GO AUROC, normalized Chi-squared and modularity. This has previously been demonstrated using KEGG pathways on Arabidopsis. In the present article, we extended our set of validation pathway by considering the Reactome database. We found that it included more complete information especially for hormone signaling pathways and was more convenient to compare pathways between the three plant species investigated here. However, we found very low normalized Chi-squared values in contrast to the KEGG pathways. We impute this difference to the complexity of the Reactome pathways which sometimes included many different subpathways. To further characterize PLC quality, we showed that the clustering coefficient also called transitivity(24) was useful to measure how networks were marked by few hubs (high clustering coefficient value) or by transcripts homogeneously connected to each others (low clustering coefficient value).

The higher AUROCs obtained with microarrays may reveal that single color arrays are well adapted to co-expression networks or that our annotation sets were more appropriated for genes effectively represented on each array. As gene models evolve with genome annotation refinement, RNA-seq, which is based on a mapping of reads on a reference transcriptome, allows to quantify more comprehensively gene expression(34). For the three species investigated here, reference transcriptomes represented more genes than arrays. It is possible that the reference annotation sets used here reflected more array gene content than the more comprehensive RNA-seq based transcriptomes, suggesting that some associations in the RNA-seq network could be true positives but are considered as false positives because not found in the reference annotation sets.

CONCLUSION

Taken altogether, our results suggest that co-expression networks using PCCs ranked with HRR clearly benefit from increasing sample size of the initial expression dataset. Small sized datasets (with less than 100 samples) had variable performance which was probably due to the samples they contained. We observed that differences between networks decreased when constructed from datasets with more than 100 samples (Figure 2 and 4). As a consequence, any combination of more than 100 samples may generate robust networks. When more than 500 samples are available (as it was the case in our 6 combinations), more biologically relevant networks can even be obtained by creating single networks after partitioning the whole dataset with a k-means algorithm and aggregating them according to co-occurring edges.

SUPPORTING MATERIAL

Supplementary Figure1: Occurrence of samples in randomly down-sampled expression matrices. For each sample, we counted the number of matrices containing it.

Supplementary Figure 2: Significant interactions between edge number and sample size on GO term recovery. Statistical effects were analyzed by ANOVA, asterisks denoting a significant effect (*, p-value<0.05; ***, p-value<0.001). The 4 groups of points represent networks obtained at 4 different significance thresholds.

Supplementary Figure 3: Performance comparison between microarray and RNA-seq. The performance in capturing GO terms of networks with a 1 million edges was measured by the number of significantly enriched GO terms (hypergeometric test, q-value<0.05). Asterisks denote a significant difference between the two plateform (Students t test, *, p-value<0.05, **, p-value<0.01, ***, p-value<0.001). Each point represent one individual network and boxplots summarize data all sample sizes combined. White triangles correspond to data for networks inferred from full datasets.

Supplementary Figure 4: Performance of aggregated networks from expression matrices down-sampled by grouping samples by their project or by clustering them by k-means. Performance is evaluated by the number of significantly enriched GO terms in each network (hypergeometric test, q-value<0.05). Single non aggregated (No) networks with 1 million edges are also reported. Aggregation was either total or partial. For the total aggregation, the 1 M best pairs were retrieved either by taking the best HRR (H) or the most co-occurring (CO) edges. For the partial aggregation, we combined either the 50% of networks with the highest GO AUROCs (HGA) or 50% with the lowest GO AUROCs (LGA). For these two partial aggregations, the 1 million best pairs were retrieved by taking the best HRR (H-HGA or H-LGA).

Supplementary Figure 5: Application of PLC on aggregated networks from expression matrices down-sampled by grouping samples by their project or by clustering them by k-means. Boxplots summarize values obtained for 13 biological pathways from the Reactome Database. Performance in GO recovery is evaluated by GO AUROC and the number of significantly enriched GO terms (hypergeometric test, q-value<0.05). Aggregation was either total or partial. For the total aggregation, the 1 M best pairs were retrieved either by taking the best HRR (H) or the most co-occurring (CO) edges. For the partial aggregation, we combined either the 50% of networks with the highest GO AUROCs (HGA) or 50% with the lowest GO AUROCs (LGA). For these two partial aggregations, the 1 million best pairs were retrieved by taking the best HRR (H-HGA or H-LGA).

Supplementary Figure 6: PLC Networks obtained full datasets and aggregated networks for the secondary metabolites Reactome pathway.

Supplementary Figure 7: Maximum performance improvement of aggregated networks (A, GO AUROC; B, counts of significantly enriched GO terms) over full dataset. Aggregate size was set at a 1 million edges. Only

two aggregation methods are shown. Y values correspond to the mean GO AUROC or counts of significantly enriched GO terms from the best combination of sample and aggregate size (from Figure 7) to which the value of the full dataset network was subtracted. Redundant edges in the aggregate were collapsed according to their co-occurrence (CO) in several individual networks or according to their best HRR value (H).

Supplementary Figure 8: Application of PLC on aggregated networks from expression matrices randomly down-sampled. Individual networks (1M edges) from a same sample size were aggregated in various number (x axis, aggregate size) and PLCs were performed using 13 different gene sets from the Reactome database. Characteristics were averaged over all replicates from a same sample size x aggregate size combination. Each point corresponds to averaged measure of 13 pathways in the Reactome database and vertical bars range from min and max values. Performance in GO recovery is evaluated by GO AUROC and the number of significantly enriched GO terms (hypergeometric test, q-value<0.05) and both measures are expressed as the average of each pathway difference between aggregates and the corresponding full dataset network. Other measures, modularity, clustering coefficient and the log likelihood of a Power law fit are used to analyze PLC topologies. Normalized Chi-squared value measure the quality of guide gene partitioning into expected subgroups as depicted in the Reactome database. Pink and blue points respectively correspond to aggregates of networks obtained with k-means or project partitioned datasets.

Supplementary Figure 9: PLC with the Fatty acid metabolism gene set on aggregated networks. For clarity purposes, edges are not shown. Communities are delimited by black polygons and colored vertices represent guide genes from a same sub-pathway. Redundant edges in the aggregate were collapsed according to their co-occurrence (CO) in several individual networks or according to their best HRR value (H). All PLCs contain the best edges allowing the representation of no more than 300 vertices.

Supplementary Table 1: Sample and their corresponding study accession number.

Supplementary Table 2: Gene accessions for Reactome pathways.

Supplementary Table 3: Guide genes related to Jasmonic Acid biosynthesis. Gene accessions were retrieved from the Plant Metabolic Network database.

Supplementary Table 4: Gene content and functional enrichment of Pathway Level Coexpressions with Jasmonic Acid (JA) biosynthesis related genes. GG, guide genes; AG, associated genes (evident association with JA biosynthesis or signaling); OG, other genes (non-evident association with JA biosynthesis or signaling).

ACKNOWLEDGEMENTS

We deeply acknowledge the Fdration CaSciModOT (CCSC Orlans-Tours, France), Jean-Louis Rouet and Laurent Catherine for help and access to the Rgion Centre computing grid. We also thanks Yann Jullian for access and help on University computer resources.

Funding. Doctoral Fellow attributed to F.L. was funded by the Rgion Centre-Val de Loire, France and the Ministre de lEnseignement Suprieur et de la Recherche, France.

Conflict of interest statement. None declared.

REFERENCES

1. Ballouz, S. and Verleyen, W. and Gillis, J. (2015) Guidance for RNA-seq co-expression network construction and analysis: safety in numbers. *Bioinformatics*, **31**, 2123–2130.
2. Cosgrove, E.J. and Gardner, T.S. and Kolaczyk, E.D. (2010) On the choice and number of microarrays for transcriptional regulatory network inference. *BMC bioinformatics*, **11**, 454.
3. Altay, G. (2012) Empirically determining the sample size for large-scale gene network inference algorithms. *IET systems biology*, **6**, 35–43.
4. Gibson, S.M. and Ficklin, S.P. and Isaacson, S. and Luo, F. and Feltus, F.A. and Smith, M.C. (2013) Massive-scale gene co-expression network construction and robustness testing using random matrix theory. *PloS one*, **8**, e55871.
5. Hibbs, M.A. and Hess, D.C. and Myers, C.L. and Huttenhower, C. and Li, K. and Troyanskaya, O.G. (2007) Exploring the functional landscape of gene expression: directed search of large microarray compendia. *Bioinformatics*, **23**, 2692–2699.
6. Feltus, F.A. and Ficklin, S.P. and Gibson, S.M. and Smith, M.C. (2013) Maximizing capture of gene co-expression relationships through pre-clustering of input expression samples: an Arabidopsis case study. *BMC systems biology*, **7**, 44.
7. Lee, H.K. and Hsu, A.K. and Sajdak, J. and Qin, J. and Pavlidis, P. (2004) Coexpression analysis of human genes across many microarray data sets. *Genome research*, **14**, 1085–1094.
8. Gillis, J. and Pavlidis, P. (2011) The impact of multifunctional genes on" guilt by association" analysis. *PloS one*; **6**, e17258.
9. Wasternack, C. and Feussner, I. (2018). The oxylipin pathways: biochemistry and function. *Annual Review of Plant Biology*; **69**, 363–386.
10. Schlapfer, P. and Zhang, P. and Wang, C. and Kim, T. and Banf, M. and Chae, L. and Dreher, K. and Chavali, A. K. and Nilo-Poyanco, R. and Bernard, T. and Kahn, D. and Rhee, S.Y. (2017). Genome-Wide Prediction of Metabolic Enzymes, Pathways, and Gene Clusters in Plants. *Plant Physiology*; **173**:2041–2059.
11. Adler, P. and Kolde, R. and Kull, M.s and Tkachenko, A. and Peterson, H. and Reimand, J. and Vilo, J.K. (2009) Mining for coexpression across hundreds of datasets using novel rank aggregation and visualization methods. *Genome biology*; **10**, R139.
12. Kauffmann, A. and Gentleman, R. and Huber, W. (2008) arrayQualityMetricsa bioconductor package for quality assessment of microarray data. *Bioinformatics*; **25**, 415–416.
13. Jin, J.P. and Tian, F. and Yang, D.C. and Meng, Y.Q. and Kong, L. and Luo, J.C. and Gao, G. (2017) PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Research*; **45(D1)**:D1040–D1045.
14. Gautier, L.t and Cope, L. and Bolstad, B.M. and Irizarry, R.A. (2004) affyanalysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*; **20**, 307–315.
15. Patro, R. and Duggal, G. and Love, M.I. and Irizarry, R.A. and Kingsford, C. (2017) Salmon provides fast and bias-aware quantification of transcript expression. *Nature methods*; **14**, 417.
16. Liesecke, F. and Daudu, D. and Dug de Bernonville, R. and Besseau, S. and Clastre, M. and Courdavault, V.t and De Craene, J.-O. and Crche, J. and Giglioli-Guivarc'h, N. and Glvarec, G. and Pichon, O. and Dug de Bernonville, T. (2018) Ranking genome-wide correlation measurements improves microarray and RNA-seq based global and targeted co-expression networks. *Scientific Reports*; **8**, 10885.
17. Csardi, G., and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*, **1695**, 1–9.
18. Couto, C.M.V. and Comin, C.H.e and da Fontoura Costa, L. (2017) Effects of threshold on the topology of gene co-expression networks. *Molecular BioSystems*; **13**, 2024–2035.
19. Naithani, S. and Preece, J. and DEustachio, P. and Gupta P. and Amarasinghe, V. and Dharmawardhana, P.D. and Wu, G. and Fabregat, A. and Elser, J.L. and Weiser, J. and Keays, M. and Fuentes, A.M. and Petryszak, R. and Stein, L.D. and Ware, D. and Jaiswal, P. (2017) *Nucleic Acids Research*; **45(D1)**; D1029-D1039.
20. Tian, T. and Liu, Y. and Yan, H.u and You, Q. and Yi, X. and Du, Z. and Xu, W. and Su, Z. (2017) agriGO v2. 0: a GO analysis toolkit for the agricultural community, 2017 update. *Nucleic acids research*; **45**, W122–W129.
21. Ballouz, S. and Weber, M. and Pavlidis, P. and Gillis, J. (2016) EGAD: ultra-fast functional analysis of gene networks. *Bioinformatics*; **33**, 612–614.
22. Bai, Y. and Meng, Y. and Huang, D. and Qi, Y. and Chen, M. (2011) Origin and evolutionary analysis of the plant-specific TIFY transcription factor family. *Genomics*; **98**, 128–136.
23. Wei, H. and Persson, S. and Mehta, T. and Srinivasasainagendra, V. and Chen, L. and Page, G.P. and Somerville, C. and Loraine, A. (2006) Transcriptional coordination of the metabolic network in Arabidopsis. *Plant physiology*; **142**, 762–774.
24. Horvath, S. and Dong, J. (2008) Geometric interpretation of gene coexpression network analysis. *PLoS computational biology*; **4**, e1000117.
25. Righetti, K. and Vu, J.L. and Pelletier, S. and Vu, B.L. and Glaab, E. and Lalanne, D. and Pasha, A. and Patel, R.V. and Provart, N.J. and Verdier, J. and others. (2015) Inference of longevity-related genes from a robust coexpression network of seed maturation identifies regulators linking seed storability to biotic defense-related pathways. *The plant cell*; **27**, 2692–2708.
26. Hickman, R. and van Verk, M.C. and Van Dijken, A.J.H. and Pereira Mendes, M. and Vroegop-Vos, I.A. and Caarls, L. and Steenbergen, M. and Van Der Nagel, I. and Wesselink, G.J. and Jironkin, A. and Talbot, A. and Rhodes, J. and de Vries, M. and Schuurink, R.C. and Denby, K. and Pieterse, C.M.J. and Van Wees, S.C.M. (2017) Architecture and dynamics of the jasmonic acid gene regulatory network. *The Plant Cell*; **tpc-00958**.
27. Ruiz-Sola, M. and Coman, D. and Beck, G. and Barja, M.V. and Colinas, M. and Graf, A. and Welsch, R. and Rütimann, P. and Bühlmann, P. and Bigler, L. and others. (2016) Arabidopsis GERANYLGERANYL DIPHOSPHATE SYNTHASE 11 is a hub isozyme required for the production of most photosynthesis-related isoprenoids. *New Phytologist*; **209**, 252–264.
28. Guerin, C. and Joet, T. and Serret, J. and Lashermes, P. and Vaissayre, V. and Agbessi, M.D.T. and Beule, T. and Severac, D. and Amblard, P. and Tregear, J. and others. (2016) Gene coexpression network analysis of oil biosynthesis in an interspecific backcross of oil palm. *The Plant Journal*; **87**, 423–441.
29. Tantong, S. and Pringsulaka, O. and Weerawanich, K. and Meeprasert, A. and Rungrotmongkol, T. and Sarnthima, R. and Roytrakul, S. and Sirikantaramas, S. (2016) Two novel antimicrobial defensins from rice identified by gene coexpression network analyses. *Peptides*; **87**, 7–16.
30. Birkenbihl, R.P. and Liu, S. and Somssich, I.E. (2017) Transcriptional events defining plant immune responses. *Current opinion in plant biology*; **38**, 1–9.
31. Caputi, L. and Franke, J. and Farrow, S.C. and Chung, K. and Payne, R.M. E. and Nguyen, T.-D. and Dang, T.-T. T. and Soares Teto Carqueijeiro, I. and Koudounas, K. and Dugé de Bernonville, T. and Ameyaw, B. and Jones, D. M. and Vieira, I. J. C.o and Courdavault, V. and O'Connor, S. E. (2018) *Science*; **360**, 1235–1239.
32. Sibout, R. and Proost, S. and Hansen, B.O. and Vaid, N. and Giorgi, F.M. and Ho-Yue-Kuang, S. and Legée, F. and Cézart, L. and Bouchabké-Coussa, O. and Soulhat, C. and others. (2017) Expression atlas and comparative coexpression network analyses reveal important genes involved in the formation of lignified cell wall in Brachypodium distachyon. *New Phytologist*; **215**, 1009–1025.
33. Obayashi, Takeshi and Aoki, Yuichi and Tadaka, Shu and Kagaya, Yuki and Kinoshita, Kengo. (2017) ATTED-II in 2018: a plant coexpression database based on investigation of the statistical property of the mutual rank index. *Plant and Cell Physiology*; **59**, e3–e3.
34. Schaefer, R.J. and Michno, J.-M. and Myers, C.L. (2017) Unraveling gene function in agricultural species using gene co-expression networks. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*; **1860**, 53–63.
35. Barabási, A.-L. and Albert, R. (1999) Emergence of scaling in random networks. *science*; **286**, 509–512.
36. Bolger, A.M. and Lohse, M. and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*; **30**, 2114–2120.
37. Kauffmann, A. and Rayner, T.F. and Parkinson, H. and Kapushesky, M. and Lukk, M. and Brazma, A. and Huber, W. (2009) Importing arrayexpress datasets into r/bioconductor. *Bioinformatics*; **25**, 2092–2094.
38. Barabsi, A. L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*; **286**, 509–512.
39. Mutwil, M. and Klie, S. and Tohge, T. and Giorgi, F.M. and Wilkins, O. and Campbell, M.M. and Fernie, A.R. and Usadel, B. and Nikoloski, Z. and Persson, S. (2011). PlaNet: combined sequence and expression comparisons across plant networks derived from seven species. *The Plant Cell*: **tpc.111.083667**.
40. Proost, Sebastian and Mutwil, Marek. (2017) BPlaNet: comparative co-expression network analyses for plants. In Editor,A. and Editor,B. (eds),

Plant Genomics Databases, Springer, pp. 213–227.
41. R Core Team. (2018) R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*, **https://www.R-project.org/**.