Cross species comparative transcriptomics using co-expression networks

Lars Grønvold[1] and Torgeir R. Hvidsten[1,2]

[1]Faculty of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, NO-1432, Ås, Norway.
[2]Umeå Plant Science Centre, Department of Plant Physiology, Umeå University, SE-90187, Umeå, Sweden.

Introduction

Comparing protein sequence is very powerful and is used routinely when annotating newly sequenced genomes. This relies on the observation that similar/homologous sequences have the same or similar molecular functions which makes it possible to transfer the knowledge acquired from experiments on model species over to other species. However, the function of a gene is also determined by its transcriptional program. Changes in gene regulation may underlie much of the differences we observe between species (Chan *et al.* 2010), however, we know very little about the dynamics of gene regulatory evolution compared to what we know about sequence evolution. One way to learn more is to compare gene expression from several species.

Because of the dynamic nature of gene expression, direct comparison between species requires sampling from the same types of tissues at the same developmental stage and under the same conditions, something that can be difficult or impossible for distantly related species. To avoid this limitation, it is possible to use co-expression to indirectly compare the gene expression patterns between species (Tirosh *et al.* 2007). The principle is that if a gene has conserved regulation it will have retained the same co-expression partners. As this method does not require directly comparable samples, it can take advantage of the ever increasing amount of expression data available in databases.

Another prerequisite for comparing gene expression in different species is to identify the orthologous genes, i.e. genes in each species that descend from the same single gene in the most recent common ancestor. Because genes often are duplicated and/or lost during evolution, there will be orthologs that do not have a simple one-to-one relationship (1:1 orthologs), but instead have a one-to-many relationship (1:N), if the gene is duplicated in one of the species after speciation, or many-to-many relationship (N:N), if there are duplications in both species.

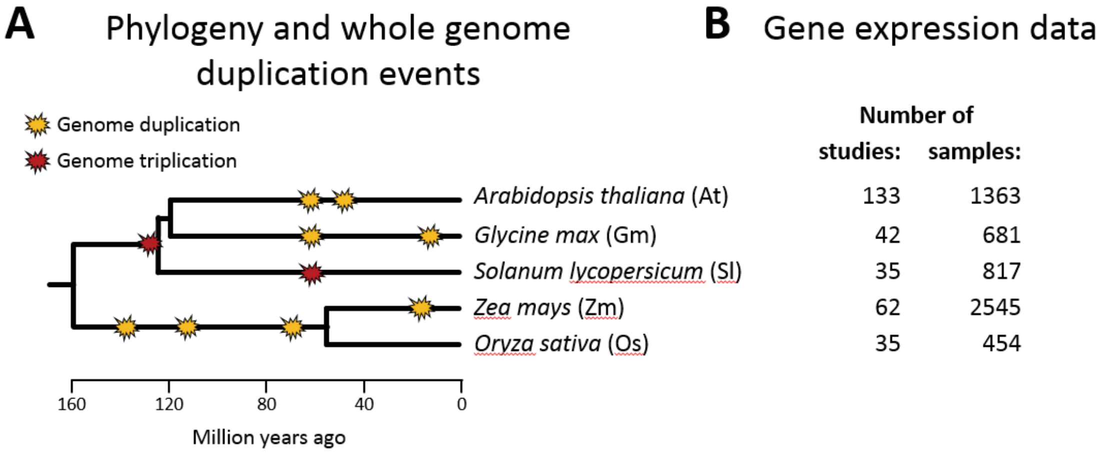
Gene-duplication is thought to be an important driver of the evolution of gene regulation. As most duplicated genes are lost in the long run, the ancient duplicates that are still retained must either have acquired a new function (neo-functionalization) or perform complementary sub-functions of the original gene function (sub-functionalization)(Ohno 1970). In terms of gene expression, sub-functionalization might imply that each copy is expressed in different tissues, thus allowing them to specialize, while neo-functionalization implies that one copy retains the ancestral expression pattern and the other copy acquires a novel expression pattern.

There are several different methods that compare co-expression across several species. Some of these methods mainly identify a cross-species consensus network or gene modules (Stuart *et al.* 2003; Oti *et al.* 2008; Mutwil *et al.* 2011; Zarrineh *et al.* 2011; Gerstein *et al.* 2014) while others specifically quantify the co-expression similarity between a pair of orthologous genes (Dutilh *et al.* 2006; Tirosh & Barkai 2007; Netotea *et al.* 2014; Monaco *et al.* 2015). There are two main approaches to comparing co-expression: overlap of co-expressed genes or correlation of co-expression values.

The overlap based methods first identifies a set of genes co-expressed with the gene of interest in each species. This co-expression set can either be the neighbors in an un-weighted co-expression network, typically generated by applying a threshold to the co-expression matrix, or can be based on clustering (e.g. Mutwil *et al.* 2011). The two sets of co-expressed genes, one set from each species, are then compared by first determining the number of orthologous genes they share and then by evaluating the statistical significance of this overlap (using e.g. the hypergeometric test). The methods also differ in how the orthologs are defined. Some methods analyze large gene families with ancient paralogs (e.g. Pfam), some use ortholog groups while others only consider 1:1 orthologs.

Correlation based methods have the advantage of being threshold-independent. They calculate the correlation between two co-expression vectors, one from each species, that contain the co-expression values of the compared orthologs to a set of 1:1 reference orthologs. The idea is that these reference orthologs are unduplicated genes performing the ancestral function and that they display conserved expression patterns. Compared to the overlap-methods, the correlation-based methods are relatively less studied, and all studies that use this method have relied on calculating the co-expression matrix using the Pearson correlation coefficient (PCC)(Dutilh *et al.* 2006; Tirosh & Barkai 2007; Wang *et al.* 2011). The overlap-based methods, on the other hand, also utilize alternative correlation measures, such as mutual information (MI), and often perform an additional normalization step, such as context likelihood of relatedness (CLR), highest reciprocal rank or mutual rank (MR) before applying a co-expression threshold.

In this study, we evaluate the correlation-based method by testing various methods for calculating the co-expression matrix. We use “co-expression correlation score” (CCS) to refer to the cross-species correlation value we obtain from comparing the co-expression patterns of two orthologs. To evaluate the performance of different methods, we apply a novel method that rank the score of the 1:1 orthologs among the scores obtained between one of the orthologs and all the genes in the other species. The idea is that the orthologous gene would tend to be the one with the most similar expression compared to all non-orthologous genes. This ortholog rank score (ORS) can



	Number of studies:	Number of samples:
Arabidopsis thaliana (At)	133	1363
Glycine max (Gm)	42	681
Solanum lycopersicum (Sl)	35	817
Zea mays (Zm)	62	2545
Oryza sativa (Os)	35	454

Figure 1: (A) Phylogeny of the included species and known whole genome duplication events (Vanneste et al. 2014). (B) Amount of public gene expression data used in this study.

also be considered a significance measure for the CCS and may be used as an indicator of conserved co-expression instead of using the CCS directly.

Using RNAseq data from five plant species, we identified Pearson correlation followed by mutual rank (PCC+MR) as the best method for comparing co-expression across species. We also investigated the effect that the number of samples has on the results and find that using samples from many diverse studies gives the best results. Although both the CCS and ORS measures are inevitably biased by the samples that are available for each species, we demonstrate that it can reliably detect trends for groups of genes as we can clearly observe a relationship between gene duplication and conservation of expression patterns.

Results

Method overview

Coexpression matrices were generated from public gene expression data for five plant species (Figure 1). For each pair of species, the co-expression correlation score (CCS) was then calculated between each pair of species by using 1:1 orthologs as a common reference (Figure 2A-D). This gives a measure of the co-expression similarity between orthologs i.e. to what degree orthologs are co-expressed with genes that are also orthologs.

As a measure of significance for the CCS values, we calculate the ortholog rank score (ORS) which indicate the fraction of all genes (not only orthologs) with a lower or equal CCS (Figure 2E). By taking $1 - ORS$, the resulting value can be interpreted as a P-value where the null hypothesis is that the CCS of the ortholog is no different than the CCS of a random gene and the alternative hypothesis is that it is higher. The ORS for orthologs tend to have a distribution heavily skewed towards 1, so we use a log transformed ORS (logORS) where a value >1 means the ortholog is among the top 10%, >2 means top 1%, >3 top 0.1% and 4 is the highest score.

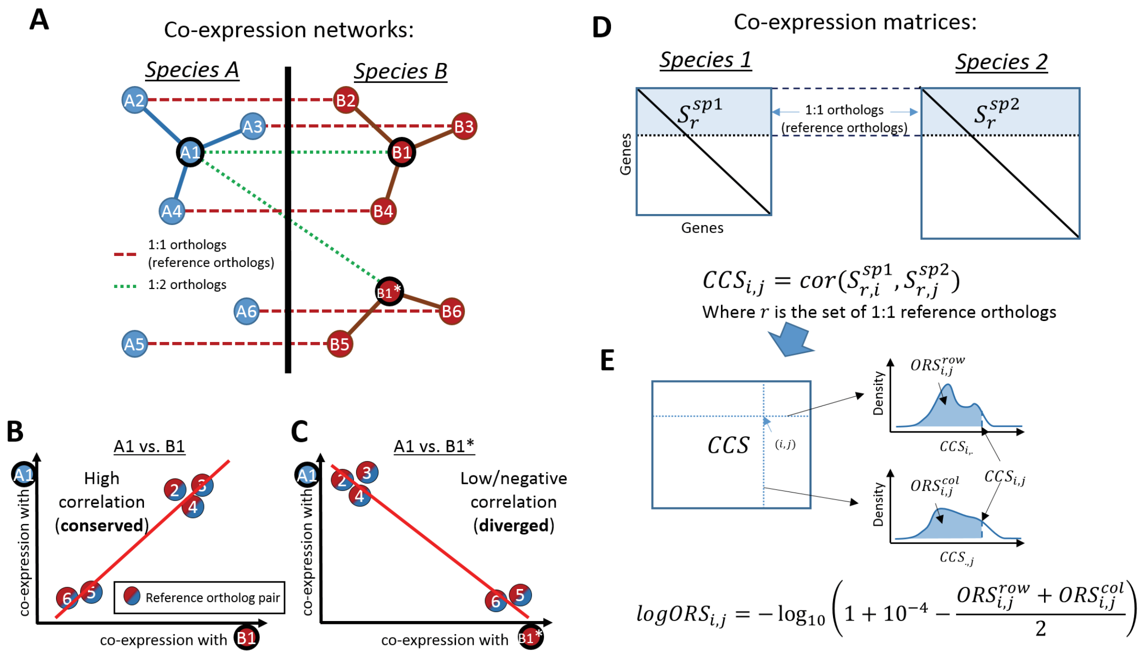


Figure 2: Co-expression correlation score (CCS) and ortholog rank score (ORS). (A) Cartoon example of two co-expression networks in two species aligned by their orthologs. (B) Cartoon example of how CCS is calculated for the ortholog pair A1 and B1. Correlation between co-expression with the reference orthologs in the respective species is high, indicating that the ortholog pair has a conserved gene expression pattern. (C) For the ortholog pair A1 and B1 the correlation of co-expression is negative as the B1* gene is co-expressed with a completely different set of genes, i.e. the expression pattern has diverged. (D) The rows of the co-expression matrices for two species are aligned by their 1:1 reference orthologs. Only these rows are used when the CCS is calculated by pearson correlation between all combinations of columns in the two co-expression matrices. (E) The ORS for an ortholog pair i and j is the proportion of values in the corresponding row or column in the CCS matrix which is lower or equal to the CCS of the ortholog pair itself, here illustrated as the area under the density curve. The mean of the column and row ORS is subtracted from 1.0001 and log transformed to get the logORS which is more suitable for visualization.*

Testing alternative co-expression methods

There are many approaches to calculating co-expression network, and here we evaluated several of the most common methods. We tested three correlation measures, Pearson correlation (PCC), Spearman correlation (SCC) and mutual information (MI), as well as two methods for normalizing the correlations, mutual rank (MR) or context likelihood of relatedness (CLR). These methods were applied to all pairs of species and evaluated using the median logORS of 1:1 orthologs (Figure 3).

With the exception of the *Gm-Os* and *Gm-Zm* comparisons, the PCC+MR method performed best and was therefore used in all subsequent analyses. Among the correlation methods, SCC performed slightly worse than PCC, while MI performed differently depending on the species

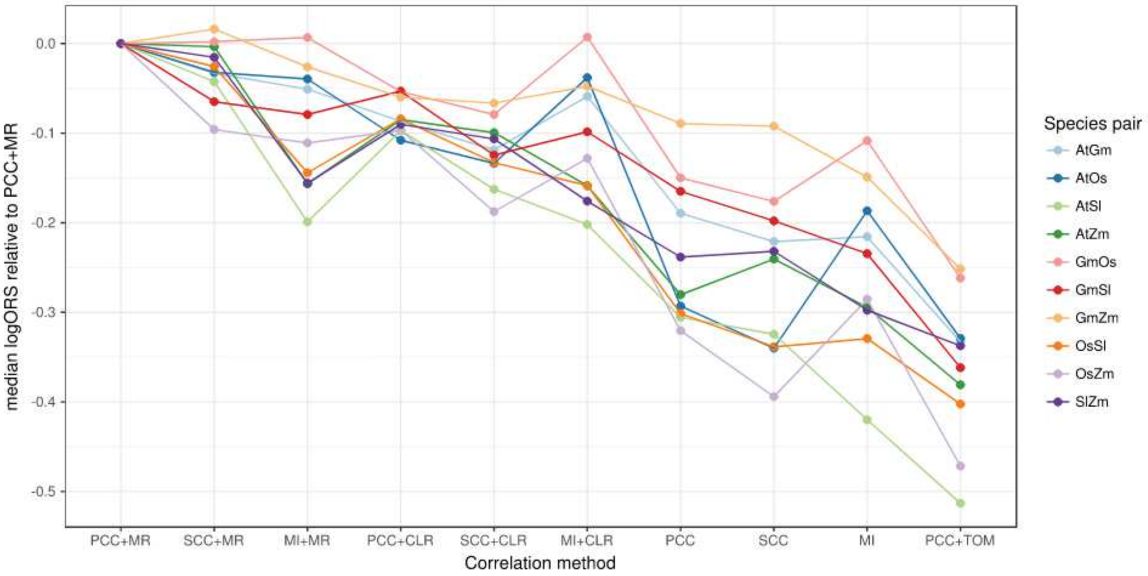


Figure 3: Evaluation of different methods for calculating co-expression matrices. The median logORS of all 1:1 orthologs between all pairs of species for each co-expression measure relative to the median logORS for PCC+MR.

pair. MR performed better than CLR, and both perform clearly better than using the correlation matrix directly. We also tested the topological overlap measure (TOM) from the commonly used WGCNA R package, which seemed to not be suitable for calculating CCS.

More samples gives higher ORS

The number of samples needed for robust inference of co-expression networks, and hence robust estimates of the conservation of networks in network comparison applications, is still debated. To test the effect of the number of samples on the conservation score, we picked subsets of various sizes from the 1363 *At* samples available and calculated ORS (PCC+MR) against the full set of *Os* samples (454 samples). To save computation time, only 1:1 orthologs were included (~5k genes). Samples were selected either randomly among all samples (individual samples) or by selecting all samples from the same study before selecting samples from another study (studies).

When selecting studies, adding more samples will, with few exceptions, result in a higher median ORS (Figure 4). This shows that in general it is a good idea to include as many samples as possible. On the other hand, when selecting individual samples, the ORS scores seem to reach the maximum and stay levelled after around one tenth of the samples have been included. This indicates that the studies tend to contain redundant samples (replicates or similar conditions) and illustrates the importance of diversity in the sampled conditions in order to get good ORSs.

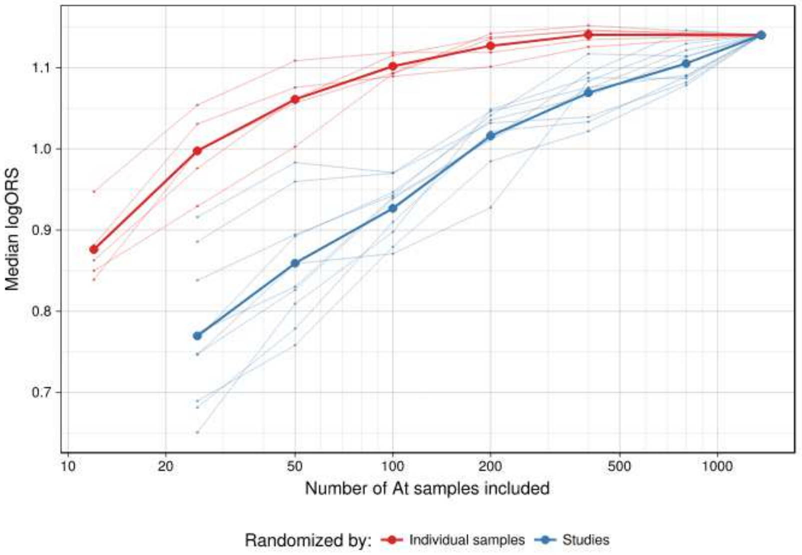


Figure 4: logORS between At and Os using various numbers of At samples. Each of the thin lines represent one selection of samples, where each point represents the same samples as the previous point plus addition samples. The thick lines show the mean of the permutations.

Sample subsets within species

The expression similarity between two species as measured by CCS or ORS will be affected by both the biological differences and by the bias caused by differences in the experimental conditions of the included samples (i.e. sampling bias) from each species. It is impossible to disentangle these factors when comparing different species, however, by comparing networks inferred from different sets of samples from within the same species it possible to study the effect of sampling bias alone (if we discount the biological differences within the sampled population which could contain mutants and variants).

The variation occurring from comparing different networks from the same species can also be considered to give rise to the background distribution for conserved gene expression (Meysman *et al.* 2013), i.e. the expected distribution of scores for an ortholog that has conserved expression pattern.

For each species, the samples were randomly split into two sets making sure that each half did not share any samples from the same study. The ORS was then calculated within species between networks inferred from the two halves of the samples. and this was repeated 10 times. When comparing networks within a species, any gene could be used as reference “orthologs”, however, to make these intra-species comparisons more similar to cross-species comparison, we used the genes which have 1:1 orthologs to the other species.

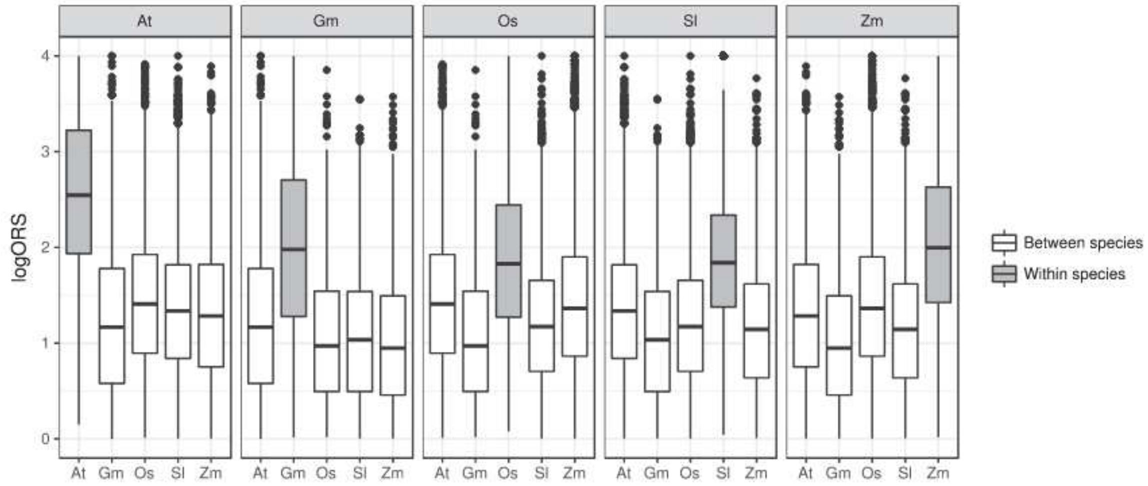


Figure 5: Distribution of ORSs of 1:1 orthologs between all pairs of species and between subsets of samples within species. Note that the results for each species pair is displayed twice for easier comparison.

The resulting ORS distribution within species are consistently higher than the ORS between species (Figure 5). Note that the between species ORS only partially reflect the phylogeny, e.g. the highest median ORS is between distantly related At and Os. One reason for this could be sample bias, e.g. the high within-species ORS for At is reflected in generally higher than expected ORS between any species and At. Another reason that the between-species ORS vary could be that the set of 1:1 orthologs also varies. This could be the reason for the low ORS for Gm, as it has relatively few 1:1 orthologs because of the recent whole genome duplication (WGD).

Expression divergence in duplicated genes

It has been shown before that duplicated genes have higher rate of expression divergence than single-copy genes (Gu *et al.* 2004; Huminiecki & Wolfe 2004; Assis & Bachtrog 2015). We therefore expect orthologs with duplications to have lower expression similarity than 1:1 orthologs. Indeed, when comparing the median logORS of 1:1 orthologs with one-to-many orthologs (i.e. 1:2, 1:3 and 1:4), there is a clear downwards trend as the number of duplicates increase (Figure 6).

An interesting exception is observed in *Gm* where the duplicates (i.e. 1:2 orthologs) seem to have an equally conserved expression pattern as the 1:1 orthologs. A possible explanation is that most of the duplicates in *Gm* originate from the relatively recent whole genome duplication (WGD) about 13 million years ago (Schmutz *et al.* 2010) and as such has not had enough time for gene expression to diverge. Note that, although much less distinct, a similar trend can be observed in *Zm*, which also experienced a WGD in about the same time frame. For comparison, the three other species haven’t experienced a WGD event in about 50-70 million years (Figure 1).

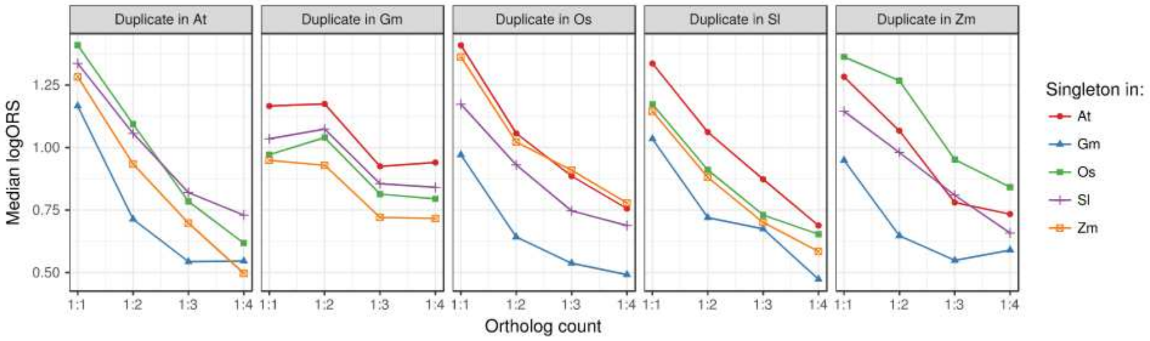


Figure 6: Median expression conservation of orthologs with 1:1-4 relationship for all pairs of species.

Even considering that a majority of the duplicates in *Gm* are recent, it would still be expected for the singleton genes to have a more conserved expression pattern. Since there is almost no difference between the ORS of the singletons and duplicates in *Gm*, there must be an additional mechanism at work. To explain the high ORS of *Gm* duplicates we hypothesize that there has been a preferential retention of genes that tend to have high ORS. To test this, we use the ORS distribution of 1:1 orthologs between two other species (*At* and *Os*) and compare the subset for which the corresponding Gm orthologs are duplicates (At1:Gm2) with the subset for which the corresponding Gm ortholog is a singleton (At1:Gm1). As predicted, the ORS tend to be higher (Wilcox rank-sum test, P=3.2e-7) when the Gm orthologs are duplicates (Figure 7).

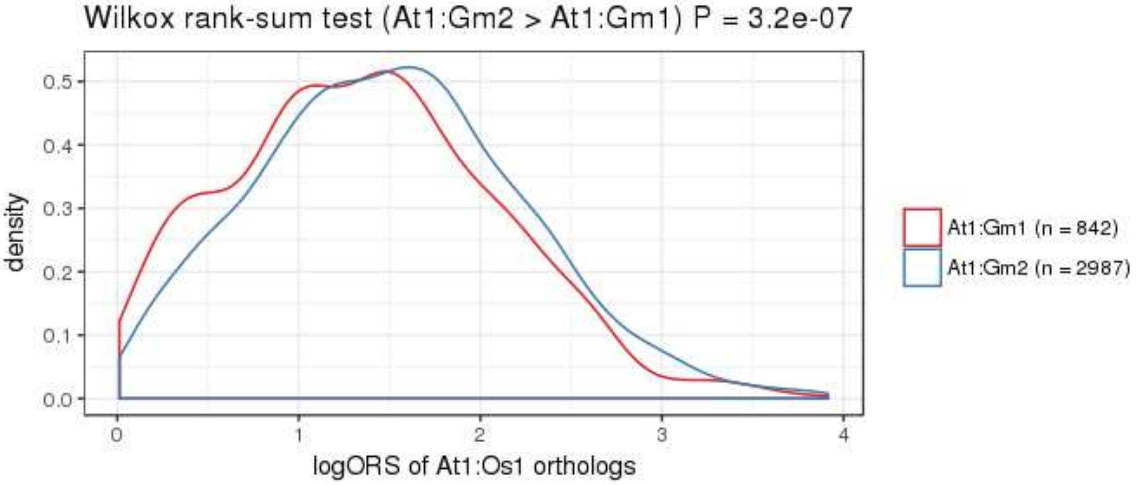


Figure 7: Expression similarity between At1:Os1 orthologs tend to be higher if the corresponding Gm ortholog is duplicate than if it is singleton. The density of logORS for two subsets of 1:1 orthologs between At and Os (At1:Os1). The red line denotes the subset where the At ortholog has a 1:1 relationship with a Gm ortholog, i.e. no duplicates in Gm. The blue line denotes the subset where the At ortholog has a 1:2 relationship with Gm, i.e.duplicates in Gm.

Discussion

Regulation of transcription is an essential function in all organisms and it is a central goal of biology to make sense of regulatory mechanisms and pathways. However, there is a lack of knowledge regarding how regulation differ between species and the evolutionary dynamics of regulatory pathways. Better methods for cross-species comparative transcriptomics would help increase our understanding of gene regulation and how it evolves. Using co-expression to indirectly compare gene transcription between species is a compelling method as it makes it possible to take advantage of the vast amount of expression data collected in public databases.

In this study we investigated several variations of co-expression measures as a basis for cross-species comparison and applied a novel scoring scheme that we termed ortholog rank score (ORS) to determine the best method. We find that by applying a normalization procedure to the co-expression matrix, such as mutual rank (MR), there is a clear improvement (figure 3) over using correlation only.

While it is not obvious why normalization of the co-expression matrix improves the cross species comparison, one possible explanation is that it reduces the bias towards large clusters in the co-expression network. For example in plants, there is a significant number of genes that are mainly expressed in tissues performing photosynthesis. Because plant expression datasets usually include both leaf samples and non-photosynthetic tissues, these genes tend to have highly correlated expression profiles (related discussion in Kinoshita & Obayashi 2009). When calculating CCS, the genes from larger co-expression clusters will get a high score because they are supported by a large number of co-expressed reference orthologs. The CCS would therefore be biased by co-expression cluster size. Both MR and CLR considers the correlation between two genes relative to the correlation with all other genes, thereby reducing this bias.

When comparing ORS of 1:1 orthologs with 1:N orthologs we found that, as expected, there is a clear tendency for duplicated genes to have more diverged expression (Figure 6). However, soybean (*Gm*) which has undergone a whole genome duplication (WGD) about 13 million years ago does not show any difference in ORS between the duplicate (1:2) and singleton (1:1) orthologs. Duplicated genes are thought to experience relaxed selection pressure that allows their expression regulation to diverge faster (Ohno 1970). The singleton genes must have lost their duplicates at some point after the WGD event. If loss of duplicates occurred randomly then a possible explanation for the observed pattern could be that a majority of the lost duplicates were lost recently, and they are consequently indistinguishable from duplicates. On the other hand, it could be that some genes are more likely to be retained than others. Supporting this, we found that the orthologs of the retained *Gm* duplicates tend to have higher ORS when comparing between other species (figure 7). However, there must have been a selective advantage to retain the genes that usually tend to have high ORS. A plausible explanation is the dosage balance hypothesis, which states that genes coding proteins that act in protein complexes, tend to be sensitive to change in relative dosage between individual subunits (Papp *et al.* 2003). The hypothesis predicts that genes sensitive to dosage balance are preferentially retained after WGDs and preferentially lost after small scale duplications. This has been suggested as an explanation for why certain categories of genes, such as transcription factors and kinases, are preferentially retained after WGDs in *Arabidopsis* (Blanc & Wolfe; Maere *et al.* 2005).

The use of ORS as a measure of performance facilitates the evaluation of different variations of methods and it is likely that there are untried alternatives that outperform PCC+MR. For example, one reason why mutual information (MI) didn't perform so well could have been that it doesn't differentiate between positive and negative correlations, which could be resolved by combining MI with PCC to generate a signed MI. There might also be something to gain by applying a transformation function, such as the soft-threshold used in WGCNA (Langfelder & Horvath 2008), to put more weight on the strong correlations. There are also alternatives to using correlation when comparing the co-expression matrices across species, such as the topological overlap measure. As it is only the imagination that limits the alternatives, future work should aim to get some theoretical understanding of why one method outperforms another.

In other studies that generate/compare co-expression networks from mixed public expression data, a sample filtering step (Mutwil *et al.* 2011) or weighting scheme (Obayashi & Kinoshita 2009) is often included. The purpose of this is to reduce bias caused by the redundancy from samples with similar expression profiles. We did not include such a step in this analysis but it is something that should be considered in future work.

The median ORS between two species can be considered to be an indicator of the total divergence in expression between them. It is expected that the expression divergence between species corresponds to the phylogenetic distance. However the observed median ORS does not reflect the phylogeny. For example, we see that the dicot *At* and monocot *Os* have a median ORS that is higher than between *At* and other dicots (Figure 5). There are several factors that could bias the ORS, such as the number of samples, how many different conditions that are sampled and to what extent the same conditions are sampled in the compared species. These biases can complicate comparisons that include more than two species. It might however be possible to compensate for some of these effects. The number of samples seems to have a rather predictable effect on the ORS (Figure 4) - and can thus be accounted for. The within-species ORS (Figure 5) might also be used as a normalization factor. However, neither would account for the sample condition compatibility between pairs of species.

Co-expression based methods can leverage the diverse data gathered in public repositories to quantify similarity of gene expression patterns across species. With improved methods such as presented in this study, combined with the ever increasing amount of available data, we believe it will be possible to investigate hypotheses about the evolution of gene regulation that previously was out of reach.

Method

Gene orthology

Ortholog information was downloaded from Ensembl plants using biomart. These are based on EnsemblCompara gene trees (Vilella *et al.* 2009). A pair of genes are defined as orthologs if the ancestor node in the gene tree is a speciation event.

Expression data

FPKM expression values was downloaded from the PODC website (http://plantomics.mind.meiji.ac.jp/podc/, Ohyanagi *et al.* 2015). Isoform level expression values were converted to per gene expression by summing the FPKMs of all isoforms for each gene. The Gm and Zm expression data at PODC were mapped to a different reference genome then we used for orthology information (Ensembl plants). The corresponding samples were therefore downloaded from EBI using their RNAseq-er API (http://www.ebi.ac.uk/fg/rnaseq/api/). FPKM values were log transformed using log2(1+FPKM).

PCC+MR similarity matrix

Co-expression for each gene pair of each species was first calculated using Pearson correlation between gene expression vectors.

$$S_{i,j}^{PCC} = PCC(E_i, E_j)$$

where E_i is the expression value vector of gene i for all samples and PCC is the Pearson correlation coefficient function. The similarity matrix was then transformed using the log mutual rank:

$$S_{i,j}^{PCC+MR} = 1 - \frac{log\left(\sqrt{R_{i,j}R_{j,i}}\right)}{log(n)}$$

where n is the number of genes and $R_{i,j}$is the rank of $S_{i,j}^{PCC}$in row i of the similarity matrix ordered from highest to lowest value. Or in other words, if all genes were sorted from highest to lowest co-expression value with gene i, then $R_{i,j}$would be the position of gene j. Note that ATTED-II uses $\sqrt{R_{i,j}R_{j,i}}$ as the mutual rank measure (Obayashi *et al.* 2009). The log transformation puts a higher weight on the most similar genes. Subtracting from 1 and dividing by $log(n)$ scales the value to a range between 0 and 1 but has no effect on the downstream analysis.

Co-expression correlation score (CCS)

For two species, the CCS between gene i in the first species and gene j in the second species is an indirect measure of expression similarity calculated by taking the Pearson correlation between the corresponding two columns of the co-expression similarity matrices. As each species has a different set of genes, only the rows with the corresponding one-to-one orthologs are correlated:

$$CCS_{i,j} = PCC(S^{sp1}_{r,i}, S^{sp2}_{r,j})$$

where S^{sp1} and S^{sp2} are the S^{PCC+MR} similarity matrices for the two species (*sp1* and *sp2*) and r are all the one-two-one ortholog pairs in the two species except pairs containing genes i or j.

Ortholog rank score (ORS)

The ortholog rank score (ORS) is derived from the CCS and can be viewed as a measure of significance of the CCS or as an alternative to CCS as a co-expression similarity measure. $ORS_{i,j}$(ORS between gene i in the first species and gene j in a second species) is calculated as the proportion of all genes in the second species which has a CCS with gene i that is lower or equal to that of gene j. In other words, $ORS_{i,j}$is the normalised rank of the $CCS_{i,j}$in row i of the CCS matrix ordered from lowest to highest value. For example, if $ORS_{i,j} = 1$then $CCS_{i,j} \geq CCS_{i,g}$ for any gene $g$in the second species, or, if $ORS_{i,j} = 0.90$ then $CCS_{i,j} \geq CCS_{i,g}$for 90% of the genes in the second species. By taking one minus the ORS, the resulting value can be interpreted as an empirical P-value (i.e. $P = 1 - ORS$) for the hypothesis that the ortholog pair i,jhas a more similar expression than expected by chance. Because the ORS is directional, i.e. $ORS_{i,j} \neq ORS_{j,i}$, we calculate an undirected ORS by taking the mean of $ORS_{i,j}$ and $ORS_{j,i}$. Furthermore, as the distribution of ORS for ortholog pairs is skewed towards 1 we use a logarithmic transformation when plotting to make comparisons easier:

$$logORS_{i,j} = -log_{10}\left(1 + 10^{-4} - \frac{ORS_{i,j} + ORS_{j,i}}{2}\right)$$

Adding the value 10^{-4} before log transforming ensures that the score gets a value between 0 and 4.

Alternative co-expression methods

Mutual information (MI) was calculated with B-Spline smoothed bins (Daub *et al.* 2004). Number of bins was set to 7 and spline order to 3.

Context likelihood of relatedness (CLR) is a background correction step that aims to remove random and indirect correlation (Faith *et al.* 2007). It involves calculating the Z-score for each row in the similarity matrix, i.e subtract the mean and divide by the standard deviation. Only positive Z-scores were used, while negative values were replaced with 0 as in (Netotea *et al.*; Madar *et al.* 2010):

$$z_{i,j} = max\left\{0, \frac{S_{i,j} - \underline{S_{i,.}}}{\sigma_i}\right\}$$

Where $\sigma_i$is the standard deviation of row i in the co-expression matrix S, and $\underline{S_{i,.}}$is the mean of row i. The CLR is then calculated by combining the row-wise and column-wise Z-scores:

$$CLR_{i,j} = \sqrt{{z_{i,j}}^2 + {z_{j,i}}^2}$$

We also tested the topological overlap measure (TOM) using the TOMsimilarity function in the WGCNA R package (Langfelder & Horvath 2008) after applying soft-threshold exponent of 6 to the PCC co-expression matrix.

References

Assis R, Bachtrog D (2015) Rapid divergence and diversification of mammalian duplicate gene functions. *BMC Evolutionary Biology*, **15**, 138.

Blanc WG, Wolfe KH Functional Divergence of Duplicated Genes Formed by Polyploidy during Arabidopsis Evolution.

Chan YF, Marks ME, Jones FC *et al.* (2010) Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. *Science (New York, N.Y.)*, **327**, 302–5.

Daub CO, Steuer R, Selbig J, Kloska S (2004) Estimating mutual information using B-spline functions--an improved similarity measure for analysing gene expression data. *BMC bioinformatics*, **5**, 118.

Dutilh B, Huynen M, Snel B (2006) A global definition of expression context is conserved between orthologs, but does not correlate with sequence conservation. *BMC Genomics*, **7**, 10.

Faith JJ, Hayete B, Thaden JT *et al.* (2007) Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS biology*, **5**, e8.

Gerstein MB, Rozowsky J, Yan K-K *et al.* (2014) Comparative analysis of the transcriptome across distant species. *Nature*, **512**, 445–448.

Gu Z, Rifkin SA, White KP, Li W-H (2004) Duplicate genes increase gene expression diversity within and between species. *Nature Genetics*, **36**, 577–579.

Huminiecki L, Wolfe KH (2004) Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. *Genome research*, **14**, 1870–9.

Kinoshita K, Obayashi T (2009) Multi-dimensional correlations for gene coexpression and application to the large-scale data of Arabidopsis. *Bioinformatics (Oxford, England)*, **25**, 2677–84.

Langfelder P, Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics*, **9**, 559.

Madar A, Greenfield A, Vanden-Eijnden E, Bonneau R (2010) DREAM3: network inference using dynamic context likelihood of relatedness and the inferelator. *PloS one*, **5**, e9803.

Maere S, De Bodt S, Raes J *et al.* (2005) Modeling gene and genome duplications in eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 5454–9.

Meysman P, Sánchez-Rodríguez A, Fu Q, Marchal K, Engelen K (2013) Expression Divergence between Escherichia coli and Salmonella enterica serovar Typhimurium Reflects Their Lifestyles. *Molecular Biology and Evolution*, **30**, 1302–1314.

Monaco G, van Dam S, Casal Novo Ribeiro JL, Larbi A, de Magalhães JP (2015) A comparison of human and mouse gene co-expression networks reveals conservation and divergence at the tissue, pathway and disease levels. *BMC evolutionary biology*, **15**, 259.

Mutwil M, Klie S, Tohge T *et al.* (2011) PlaNet: combined sequence and expression comparisons across plant networks derived from seven species. *The Plant cell*, **23**, 895–910.

Netotea S, Sundell D, Street NR, Hvidsten TR ComPlEx : Conservation and divergence of co-expression networks in A. thaliana, Populus and O. sativa.

Netotea S, Sundell D, Street NR, Hvidsten TR (2014) ComPlEx: conservation and divergence of co-expression networks in A. thaliana, Populus and O. sativa. *BMC Genomics*, **15**, 106.

Obayashi T, Hayashi S, Saeki M, Ohta H, Kinoshita K (2009) ATTED-II provides coexpressed gene networks for Arabidopsis. *Nucleic acids research*, **37**, D987-91.

Obayashi T, Kinoshita K (2009) Rank of correlation coefficient as a comparable measure for biological significance of gene coexpression. *DNA research : an international journal for rapid publication of reports on genes and genomes*, **16**, 249–60.

Ohno S (1970) *Evolution by Gene Duplication*. Springer Berlin Heidelberg, Berlin, Heidelberg.

Ohyanagi H, Takano T, Terashima S *et al.* (2015) Plant Omics Data Center: An Integrated Web Repository for Interspecies Gene Expression Networks with NLP-Based Curation. *Plant and Cell Physiology*, **56**, e9–e9.

Oti M, van Reeuwijk J, Huynen MA, Brunner HG (2008) Conserved co-expression for candidate disease gene prioritization. *BMC bioinformatics*, **9**, 208.

Papp B, Pál C, Hurst LD (2003) Dosage sensitivity and the evolution of gene families in yeast. *Nature*, **424**, 194–197.

Schmutz J, Cannon SB, Schlueter J *et al.* (2010) Genome sequence of the palaeopolyploid soybean. *Nature*, **463**, 178–183.

Stuart JM, Segal E, Koller D, Kim SK (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science (New York, N.Y.)*, **302**, 249–55.

Tirosh I, Barkai N (2007) Comparative analysis indicates regulatory neofunctionalization of yeast duplicates. *Genome Biology*, **8**, R50.

Tirosh I, Bilu Y, Barkai N (2007) Comparative biology: beyond sequence analysis. *Current Opinion in Biotechnology*, **18**, 371–377.

Vanneste K, Maere S, Van de Peer Y (2014) Tangled up in two: a burst of genome duplications at the end of the Cretaceous and the consequences for plant evolution. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **369**, 20130353.

Vilella AJ, Severin J, Ureta-Vidal A *et al.* (2009) EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome research*, **19**, 327–35.

Wang Y, Wang X, Tang H *et al.* (2011) Modes of Gene Duplication Contribute Differently to Genetic Novelty and Redundancy, but Show Parallels across Divergent Angiosperms (SR Proulx, Ed,). *PLoS ONE*, **6**, e28150.

Zarrineh P, Fierro AC, Sánchez-Rodríguez A *et al.* (2011) COMODO: an adaptive coclustering strategy to identify conserved coexpression modules between organisms. *Nucleic Acids Research*, **39**, e41–e41.