Gene Expression of Functionally-Related Genes Coevolves Across Fungal Species: Detecting Coevolution of Gene Expression Using Phylogenetic Comparative Methods

Alexander L. Cope[1,*], Brian O'Meara[1,2,3], and Michael A. Gilchrist[1,2,3]

[1]Genome Science and Technology, University of Tennessee, Knoxville

[2]Department of Ecology and Evolutionary Biology, University of Tennessee,Knoxville

[3]National Institute for Mathematical and Biological Synthesis, Knoxville, TN

[*]Corresponding author: acope3@vols.utk.edu

November 15, 2019

Abstract

Functionally-related genes are often co-expressed across conditions, leading to the hypothesis that functionally-related genes also coevolve at the gene expression level across species. Previous work using codon-based proxies for gene expression supported this hypothesis;

however, this work ignored the impact of non-independence due to shared ancestry of the species. Recent work demonstrated failure to account for shared ancestry can confound the analysis of interspecies gene expression data. We analyzed empirical gene expression data across 18 fungal species using a multivariate Brownian Motion phylogenetic comparative method (PCM), which allowed us to quantify coevolution between protein pairs while directly accounting for the shared ancestry of the species. Our work indicates proteins which interact show stronger signals of coevolution than randomly-generated pairs, with interactions showing stronger empirical and computational evidence also showing stronger signals of coevolution. We examined the effects of number of protein interactions and gene expression levels on coevolution, finding both factors are overall poor predictors of the strength of coevolution between a protein pair. Simulations further demonstrate the potential issues of analyzing gene expression coevolution without accounting for shared ancestry. Overall, this work further demonstrates the potential value of PCMs for the analysis of interspecies gene expression data, although we note several key challenges still facing researchers.

Introduction

Analysis of high-throughput transcriptomics and proteomics data often focuses on how changes in environment (e.g. nutrient availability) result in changes in mRNA or protein abundances [1]. Through the concept of "guilt-by-association," genes which show similar gene expression patterns across conditions are hypothesized to be functionally-related [2, 3, 4, 5]. For example, in *S. cerevisiae*, there is significant overlap between the proteins which physically interact and the proteins which are co-expressed [6]. Such observations have naturally led researchers to ask if functionally-related genes show coordinated changes in expression across conditions, do they also show coordinated changes, or coevolve, across species.

[7] were among the first to investigate the coevolution of gene expression across four

Saccharomyces species. Using the Codon Adaptation Index (CAI) [8] as a proxy for gene expression and protein interaction data, [7] found the correlation of CAI between interacting proteins was higher on average than the correlation between random pairs of proteins. Later work by [9] obtained similar results using CAI in prokaryotes.

Neither [7] or [9] accounted for phylogeny in their analyses. Phylogenetically-uncorrected methods, such as the Pearson correlation coefficient and ordinary least squares, assume independence of species, which is incorrect due to shared ancestry. This can lead to an overestimation of the degrees of freedom and results in an inflation of statistical significance [10]. Recent work demonstrated failure to account for shared ancestry confounded two previous analyses of interspecies gene expression data [11].

This is not to say attempts at examining coevolution of gene expression have done so without attempting to control for the phylogeny. Later work by [12] came to similar conclusions as those reached by [7] and [9] using a set of 18 fungal species. Unlike [7] and [9], [12] compared CAI between interacting proteins after removing the underlying phylogenetic signal using a transformation developed to remove this signal for protein sequence coevolution data [13]. More recently, [14] presented a method which analyzed correlated changes in empirical gene expression data across hundreds of species from the Marine Microbial Eukaryotic Transcriptome Project. This approach attempted to account for the underlying phylogenetic structure via a permutation test meant to remove the underlying phylogenetic topology when assessing statistical significance.

An alternative to removing the effects of the underlying phylogeny through data transformations or randomization procedures is to deal with it directly. The work by [10] led to the development of various other phylogenetic comparative methods (PCMs), which directly account for the underlying phylogenetic structure of the species by incorporating information on the topology and branch lengths for parameter estimation. Previous effots have developed approaches for examining coevolution of functionally-related genes. [15] presented what is essentially a phylogenetically-corrected version of phylogenetic profiling, which looks

at the correlated presence/absence of genes across species. Looking across a set of fungal species and using protein-protein interaction data to determine functionally-related genes, they found incorporating the phylogeny reduced the false positive rate compared to a Fisher's exact test. Of course, this method is not applicable if the genes are present in all species under consideration, making gene expression a valuable trait for investigating coevolution of functionally-related genes.

Many PCMs have been developed for studying the evolution of gene expression [16, 17, 18, 19, 20, 21, 22, 23, 24]. Much of this work relies on modeling gene expression evolution as an Ornstein-Uhlenbeck (OU) process [25, 26]. Modeling trait evolution as an OU process assumes the trait is evolving around an optimal value. Multivariate versions of the OU model exists (see [27]), but the additional parameters used in the model often requires more species-level data to make accurate parameter estimates. Here, we present an approach which models the coevolution of gene expression for pairs of proteins using the simpler multivariate Brownian Motion (BM) model [28, 29]. This approach allows us to estimate the degree of correlation between two traits over evolutionary time while accounting for the shared ancestry of the considered species.

We find physically-interacting proteins show, on average, stronger gene expression coevolution than randomly-generated pairs of proteins using the multivariate BM approach. In addition, we find protein pairs with more evidence of functional-relatedness tend show stronger coevolution at the gene expression level. We also find gene expression level and the number of protein interactions, which are good predictors of evolutionary rate of a gene, are poor predictors of the the strength of coevolution between protein pairs. Consistent with previous results, we also find coevolution of gene expression if an overall weak predictor of protein sequence coevolution. As expected, we also find a phylogenetically-uncorrected correlation, on average, inflate estimates of gene expression coevolution.

Materials and Methods

Protein Interaction Data

18 fungal species were chosen due to availability of RNA-Seq data and for comparability to previous studies examining the evolution of functionally-related proteins [15, 12, 7]. Consistent with [7] and [15], we use physically-interacting proteins as our test case for examining functionally-related proteins. The STRING database was used to identify empirically-determined protein-protein interactions in species for which data was available [30]. We assume these protein-protein interactions are conserved across all species under consideration. This dataset will be referred to as the "binding group" for the remainder of the paper. Randomly-generated protein pairs followed by removal of any pairs which were annotated in the STRING database for the species under consideration, even if the annotation did not specify a "binding" interaction. Any proteins with overlapping Gene Ontology terms were removed to control for potential false negatives. This dataset will be referred to as the "control group" for the remainder of the manuscript.

Gene Expression Data

Gene expression levels were estimated from publicly available RNA-Seq datasets taken from SRA using the pseudo-alignment tool, Salmon [31]. Reads were mapped against protein-coding sequences obtained from NCBI Refseq, with the exception of *S. kudriavzevii*, *S. mikatae*, *S. paradoxus*, and *S. bayanus*, which were pulled from [32], and *L. kluyverii*, which was obtained from the Broad Institute (https://portals.broadinstitute.org/regev/orthogroups/) FASTQC was used to assess the quality of the RNA-Seq reads. If necessary, TrimGalore was used to remove adaptor sequences (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). Gene expression counts were obtained using Salmon's built-in ability to control for GC and position-specific biases, and these counts were converted to the transcripts per million (TPM) metric [33]. For single-end reads, mean and standard deviation for fragment lengths were

specified to be 200 and 80, respectively, except for *S. mikatae*, *S. paradoxus*, *S. paradoxus*, for which mean fragment length was specified to be 250 [34].

Given the RNA-Seq experiments are often measured different conditions, we only selected samples from the control conditions, as these are more likely to reflect natural or standard conditions for a species. For datasets which were time course experiments, we randomly selected 3 time-points which were well-correlated in gene expression estimates (Pearson correlation $\rho > 0.9$). Each RNA-Seq sample/replicate for each species was transformed to a standard lognormal distribution (i.e. $ln(X) \sim N(0, 1)$), consistent with the transformation used by [16]. A mean and standard error of normalized TPM values were calculated for each gene across all samples/replicates used. Genes with missing data, which could be because no ortholog was identified between species or no gene expression estimate was obtained, were excluded from further analysis.

We note some of the RNA-Seq datasets did not indicate replicates, making it impossible to estimate a standard error measurement for the analysis. It is generally recommend measurement error be provided for the analysis of continuous traits during phylogenetic analysis. As a proxy for the species missing replicates, we used a closely-related species to provides estimates of the standard error. This included *S. paradoxus* (proxy: *S. cerevisiae*), *S. mikatae* (proxy: *S. bayanus*), and *N. tetrasperma* and *N. discreta* (proxy: *N. crasse*).

Ortholog identification

Orthologs for fungal species were taken from FungalDB, previous publications [35, 32], or the Reciprocal Best Hits BLAST approach, which was only used for *N. castellii*. Proteins with an annotated paralog in the FungalDB or previous literature were excluded from the analysis, as introduction of a paralog could impact the gene expression of the original gene. This eliminated 4,307 possible genes.

Phylogenetic tree construction

Codon alignments of 59 complete, randomly chosen nuclear ORF were performed using TranslatorX using the MAFFT option followed by GBlocks filtering to remove poorly aligned regions [36]. These alignments were concatenated, followed by phylogenetic tree estimation using RAxML with a partitioned GTR-Γ fit allowing rate parameters for the third codon position to vary from the first and second codon position. *C. neoformans* was designated as an outgroup. Under the Brownian Motion model, it is ideal to use an ultrametric tree unless extinct species are present [37]. To convert the RAxML tree to a ultrametric tree, treePL [38] was used to date the tree, taking the divergence time of *S. cerevisiae* and *C. neoformans* (723 MYA, from TimeTree [39]) as a calibration point.

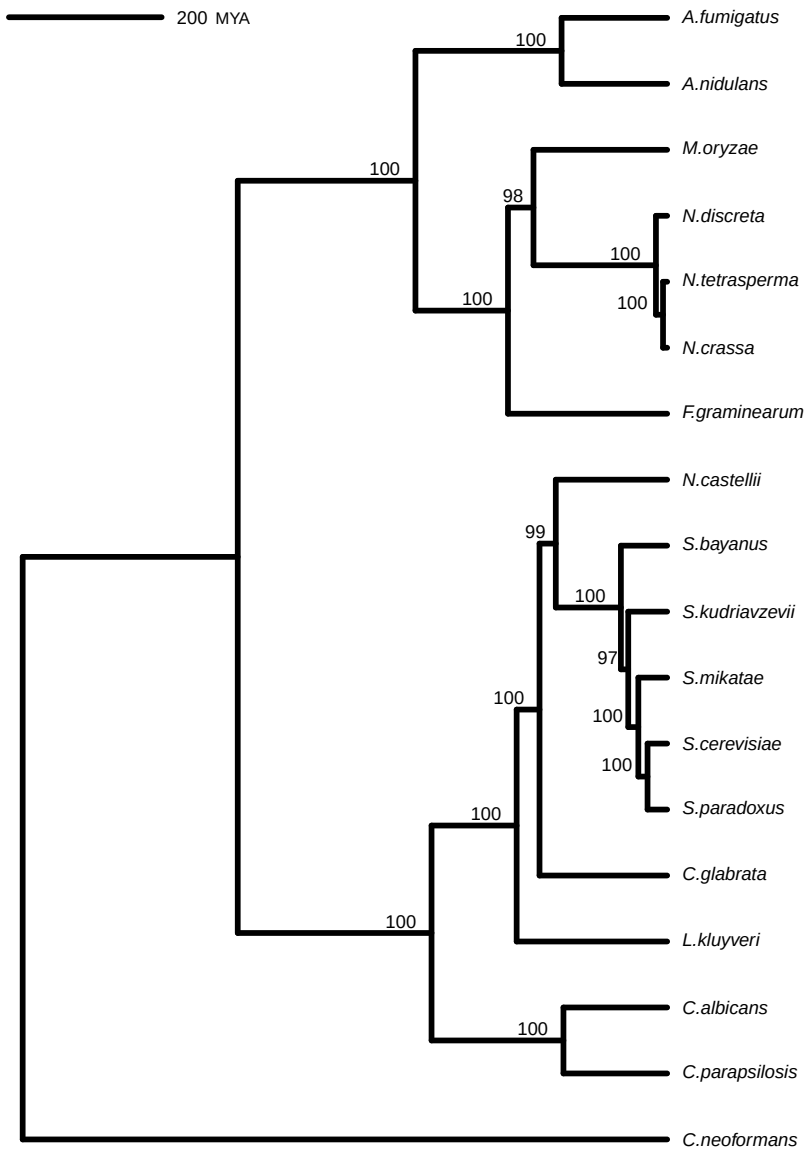


Figure 1: Dated phylogenetic tree with RAxML bootstrap support. Branch lengths are in millions of years.

Analysis of Gene Expression Data

Analyses and visualizations were performed using the R programming language. Gene expression evolution was modeled as a multivariate Brownian Motion process using the R package **mvMORPH**. Briefly, the evolutionary rate matrix for multivariate Brownian Motion represents both the trait variances on the diagonal for the individual gene expression values,

as well as the trait covariance between the gene expression estimates on the off-diagonal. The evolutionary correlation coefficient ρ_C reflects the degree to which gene expression estimates are correlated over evolutionary time and can be calculated from the evolutionary rate matrix [40, 28, 29]. The evolutionary correlation coefficient ρ_C will from here on out be referred to as the "phylogenetically-corrected correlation" to emphasize this statistic accounts for the shared ancestry of the species. Likewise, we will refer to the Pearson correlation coefficient ρ_U (estimated via the R built-in function cor.test()) as the "phylogenetically-uncorrected correlation", as this statistic ignores shared ancestry and uses variances and covariances estimated from the data at the tips of the tree.

Appropriateness of the Brownian Motion for modeling trait evolution was assessed as described in [41]. Briefly, phylogenetic independent contrasts (PICs) and standardized variances [10] were calculated from gene expression data for each ortholog set using the pic() function from the **ape** R package [42]. Pairs of genes containing a significant correlation (i.e. $p < 0.05$) between PICs and standardized variances, which indicates violation of Brownian Motion assumptions [37, 41], were excluded from further analyses.

The phylogenetically-corrected correlation ρ_C, which reflects the strength of gene expression coevolution between two genes, was compared to metrics associated with functional-relatedness of two genes. We expect stronger coevolution of gene expression between proteins which are more functionally-related. As a metric of functional-relatedness for each interaction, we used the STRING confidence score, which factors in both empirical/computational evidence supporting an interaction, as well as evidence from closely-related species. Similarly, one might expect proteins sharing a greater number of overlapping Gene Ontology (GO) terms to be more functionally-related.

It is well-established both gene expression and number of interactions in a protein-protein interaction network impact the evolutionary behavior of a protein [43, 44]; thus, we also tested if such protein-level properties also impact the strength of coevolution between two proteins. We hypothesized proteins pairs which are, on average, more highly expressed and

involved in more interactions would show stronger coevolution of gene expression. For each protein pair in the binding group, the mean degree (i.e. the average number of interactions for each protein) and the mean phylogenetically-corrected average gene expression value were calculated. The phylogenetically-corrected average gene expression value for a protein is taken as the ancestral state value estimated at the root of the tree by **mvMORPH**.

To determine if functional-relatedness, gene expression, and number of protein interactions have an impact on the strength of coevolution, a weighted rank-based (i.e. robust to non-normality in data) Spearman correlation ρ_S was used to reduce the impact of proteins found in multiple pairs. Weights for the weighted Spearman correlation ρ_S for each protein pair were calculated as

$$\text{Weight} = \frac{1}{2}\left(\frac{1}{N_1} + \frac{1}{N_2}\right)$$

where N_i is the the number of times protein i appears in the binding group. Confidence intervals and p-values for the weighted Spearman correlations were calculated using the R package **boot** [45, 46].

To assess the impact of proteins found in multiple pairs on differences observed between the binding and control groups, we generated 200 subsets of the binding and control datasets in which a protein was only allowed to appear, at maximum, in one protein pair per dataset. Each subset was restricted to a maximum size of 200 protein pairs. For each subset, the mean was calculated for ρ_C and ρ_U, creating a distribution of means. Scripts and for performing phylogenetic analysis and post-analysis of the results can be found at https://github.com/acope3/GeneExpression_coevolution.

Results

Overall, the normalized gene expression data are moderately to strongly correlated between all species (Supplemental Figure 1). Clearly, species which are more closely-related tend to show stronger correlations between normalized gene expression values, consistent with expectations. The *Candida* species appear to be exceptions, but these yeast demonstrate pathogenic traits, which could partially explain some of these differences, as well as why two of these species (*C. glabrata* and *C. parapsilosis*) appear to be better correlated with the pathogenic *Aspergillus* species.

After filtering proteins based on missing data or violation of the Brownian Motion assumption, our binding and control datasets contained 3,091 and 13,936 protein pairs respectively, consisting of 648 unique proteins. We note similar patterns are observed if not excluding genes which violate the BM assumption, although the signal overall appears weaker (see Supplemental Results).

Interacting proteins demonstrate clear coevolution of gene expression

Examination of the phylogenetically-corrected correlation ρ_C distribution for the binding and control group shows striking differences. Binding proteins have a mean phylogenetically-corrected correlation of $\bar{\rho}_C = 0.45$ (One-sample t-test, 95% CI: 0.436 – 0.464, $p < 10^{-200}$). In contrast, the randomly-generated control group had a much lower mean phylogenetically-corrected correlation of $\bar{\rho}_C = 0.03$ (One-sample t-test, 95% CI: 0.025 – 0.037, $p < 10^{-23}$). As is clear from the 95% confidence intervals, the difference between the mean phylogenetically-corrected correlations for the binding and control distributions is statistically significant (Welch's t-test, $p < 10^{-200}$). Although the mean phylogenetically-corrected correlation for the control group is significantly different from 0, it is important to note two things: (1) even though we did our best to eliminate possible false negatives in the control group, it

is unlikely all false negatives were eliminated and (2) this is consistent with previous work by [7], who also had random control groups which were not centered around 0. Despite the small, but statistically significant, deviation from 0 of the control group, the binding group shows a clear skew towards stronger coevolution between protein pairs than is observed in the control group, as expected.

We find a weak, but significant, positive correlation between the STRING confidence scores and phylogenetically-corrected correlations ρ_C (Weighted Spearman Rank Correlation $\rho_S = 0.32$, 95% CI: 0.274 – 0.371, $p < 10^{-37}$), indicating interactions which are more likely to be true and conserved show stronger coevolution of gene expression (Figure 3). A similar result is obtained when using a metric of functional similarity between proteins based on overlapping Gene Ontology terms (see Supplementary Figure 2).

We also compared how our phylogentically-corrected approach worked compared to a phylogenetically-uncorrected approach. Qualitatively, a similar pattern to the phylogenetically-corrected correlations ρ_U is observed: binding proteins show correlations positively skewed a way from 0, consistent with stronger coevolution of gene expression between the interacting pairs. Interacting proteins had a mean phylogenetically-uncorrected correlation of $\bar{\rho}_U = 0.51$ (One-sample t-test,95% CI: 0.497 – 0.523, $p < 10^{-200}$, Figure 2). In contrast, randomly-generated protein pairs had a mean phylogenetically-uncorrected correlation $\bar{\rho}_U = 0.08$ (One-sample t-test, 95% CI: 0.074 – 0.086, $p < 10^{-141}$, Figure 2). As with the phylogenetically-corrected correlations, the control group deviates significantly from the null expectation of 0.0; however, the phylogenetically-uncorrected correlation deviates further from the expectation than the phylogenetically-corrected correlations. This is consistent with inflation of correlation estimates due to treatment of non-independent species data as independent [10]. Simulations were performed to confirm potential problems with the use of non-phylogenetic methods for comparing gene expression across species (see Supplemental Methods for details on simulations). Results show failure to account for the phylogeny on data simulated under the null of no coevolution between gene expression results in an in-

crease in the false discovery rate (FDR) (Table 1), consistent with expectations. However, the distribution of ρ_U simulated under no coevolution is different from the distribution of ρ_U from the real data (Supplemental Figure 5d). In the case of simulated data in which no coevolution was allowed, the distribution of phylogenetically-uncorrected correlations ρ_U is centered around 0.0, unlike in the real data, but shows a broadening of the distribution compared to the phylogenetically-corrected correlations ρ_C (Supplemental Figure 5b).

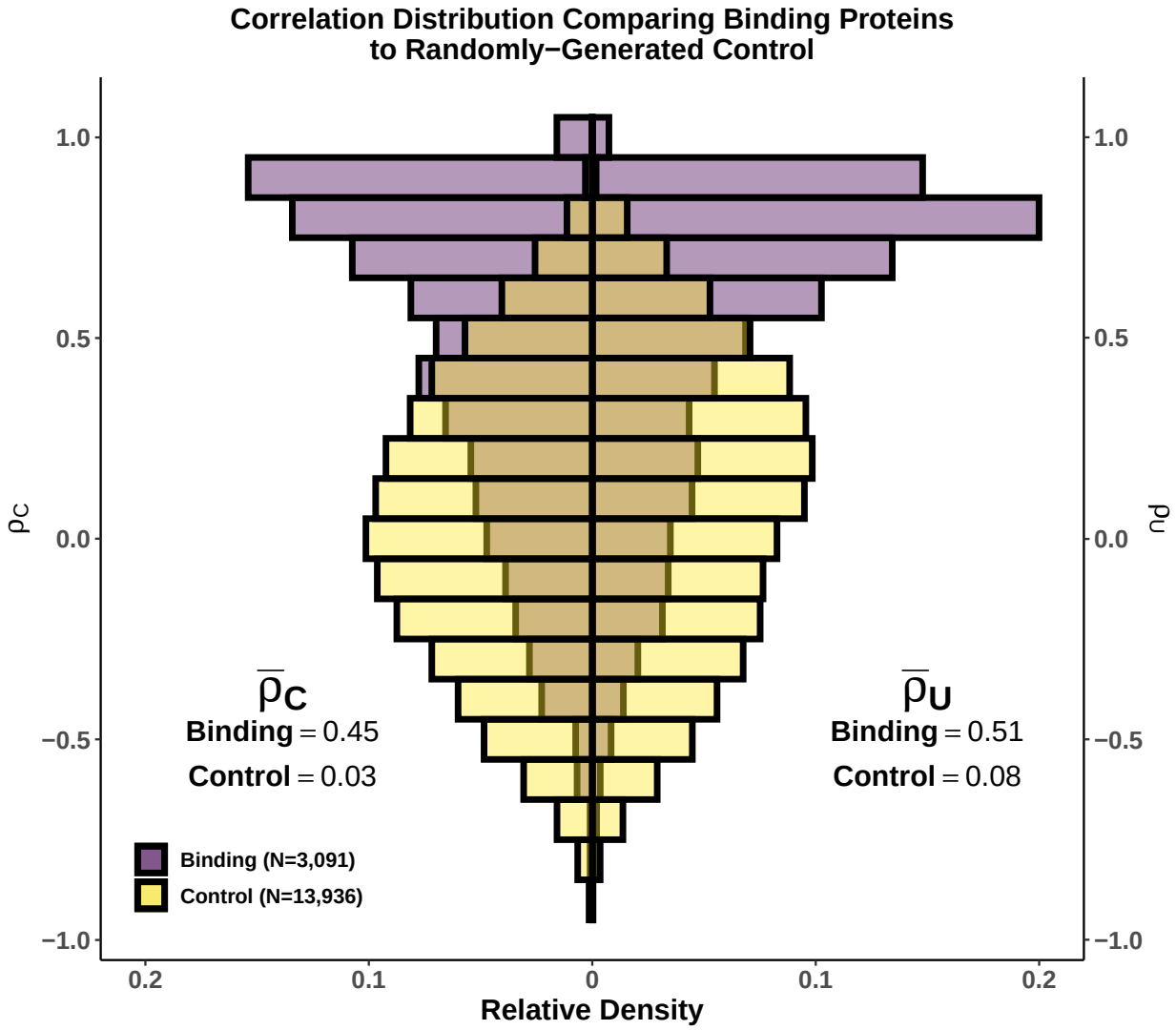


Figure 2: Comparing the distributions of the (Left) phylogenetically-corrected correlation ρ_C and the (Right) phylogenetically-uncorrected correlation ρ_U for the binding (purple) and control (yellow) groups. (Left) Mean values for the binding and control group phylogenetically-corrected correlation ρ_C distributions are 0.45 (95% CI: 0.436 – 0.464) and 0.03 (95% CI: 0.025 – 0.037), respectively. (Right) Mean values for the binding and control group phylogenetically-uncorrected correlation ρ_U distributions are 0.51 (95% CI: 0.497 – 0.523) and 0.08 (95% CI: 0.074 – 0.086), respectively

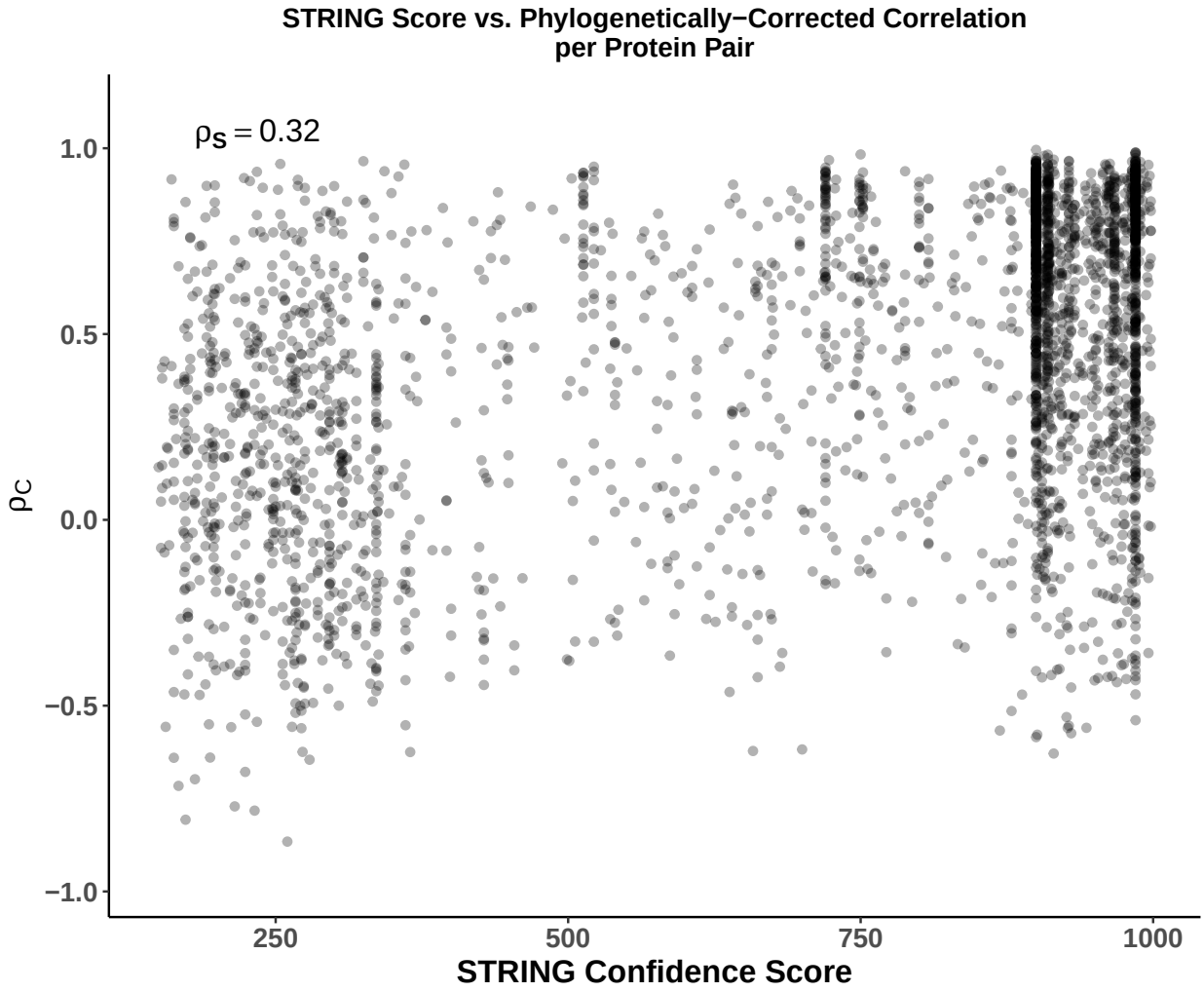


Figure 3: Effects of metric representing functional-relatedness on phylogenetically-corrected correlation ρ_C. Positive weighted Spearman rank correlation ($\rho_S = 0.32$, $p < 10^{-37}$) between the STRING score and phylogenetically-corrected correlation ρ_C indicates more confident and/or conserved interactions tend to have higher ρ_C, indicating stronger coevolution at the gene expression level.

Type	TPR	TNR	FDR	Overall Accuracy
Phylogenetically-corrected Correlation (ρ_C)	0.52	0.96	0.05	0.74
Phylogenetically-uncorrected Correlation (ρ_U)	0.61	0.72	0.25	0.65

Table 1: Performance metrics for 18 species fungal tree using simulated data for detecting correlated evolution. While the phylogenetically-uncorrected correlation ρ_U has greater statistical power, it also has a much lower true negative rate when the true model is Brownian Motion.

Gene expression and number of interactions are poor predictors of coevolution of gene expression

It is well-established both gene expression and location in a protein-protein interaction network significantly impact the evolutionary behavior of a protein [47, 48, 44, 49, 50]. One might expect an imbalance in the number of proteins involved in a greater number of interactions or more highly expressed interactions to have a more negative impact on fitness, leading to greater constraints on the evolution of gene expression. However, we find both the number of interactions and the gene expression to be weak predictors of the strength of coevolution of gene expression. Based on the number of interactions for each protein in our binding dataset, the weighted Spearman rank correlation between the number of interactions and the phylogenetically-corrected correlations ρ_C is $\rho_S = 0.26$ (95% CI: 0.196 – 0.315, $p < 10^{-16}$), indicating protein pairs involved in more interactions tend to show stronger constraint on the evolution of gene expression. Surprisingly, the mean ancestral gene expression estimates are negatively correlated with the phylogenetically-corrected correlations ρ_C, with $\rho_S = -0.09$ (95% CI: - 0.143 – -0.035, $p = 0.00131$).

Given phylogenetically-corrected correlations ρ_C correlate with the number of interactions and mean ancestral gene expression, differences between the binding and control groups in terms of number of interactions and gene expression could introduce small biases when comparing the ρ_C distributions. The average mean ancestral gene expression estimate distributions for the binding and control group are extremely similar (0.414 vs. 0.416, respectively, Welch's t-test, $p = 0.8316$). This makes differences in the gene expression distributions an unlikely source of bias when comparing the binding and control groups. To determine if protein membership biases causes biases in the results, 200 subsets of the binding and control groups were sampled, restricting a protein appearing in each group a maximum of 1 time. The 200 subsets resulted in distributions of the mean phylogenetically-corrected correlations $\bar{\rho}_C$, which were qualitatively consistent with the full datasets (Supplemental Figure 2). We do note there appears to be less of a difference between the binding and control

group $\bar{\rho_C}$ distributions compared to $\bar{\rho}_C$ estimated from the full dataset (Supplemental Figure 3). This could be due to the representation of certain proteins in the binding group inflating the correlation, or could be due to decreased power to detect differences due to the significantly reduced dataset. Despite this, the overall interpretation is the same: interacting proteins show greater coevolution at the gene expression level than randomly generated pairs of proteins.

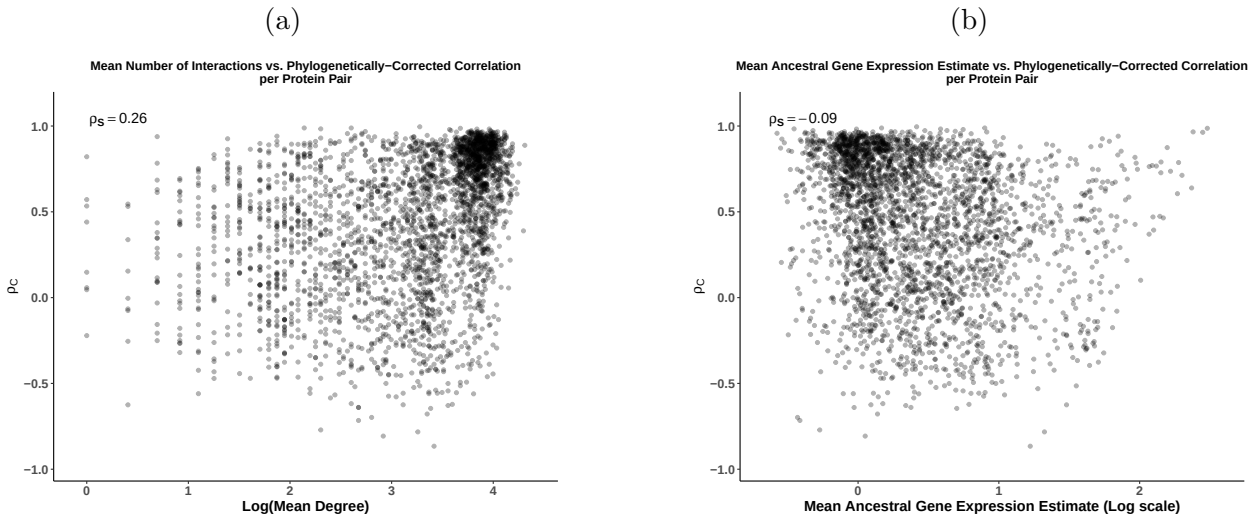
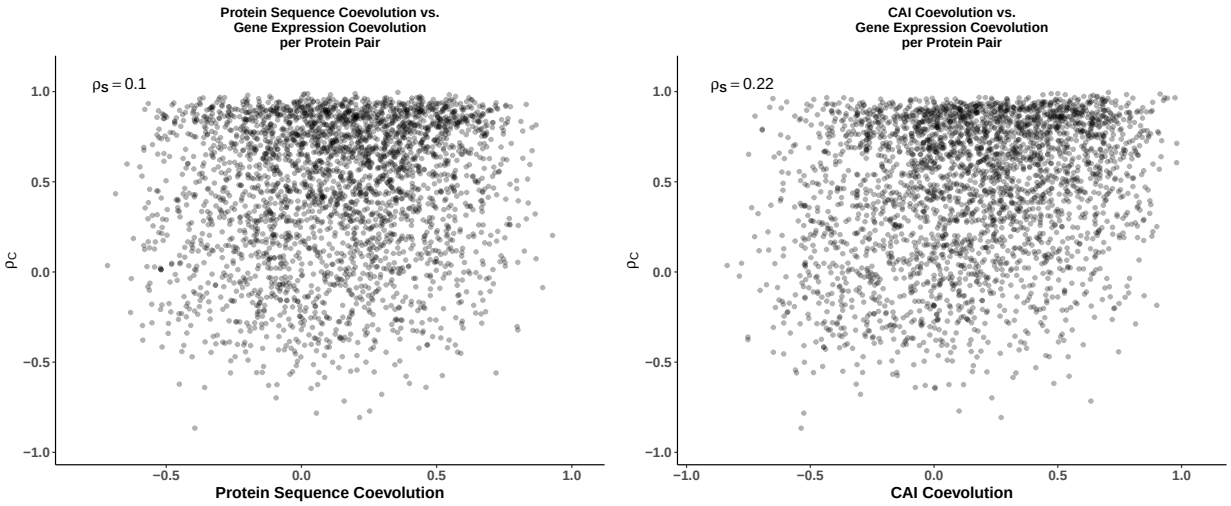


Figure 4: The relationship of (a) the mean degree (average number of interactions between a protein pair) and (b) mean ancestral gene expression estimate with the phylogenetically-corrected correlation ρ_C for the binding group. Both protein pair metrics are weakly, but significantly correlated with the phylogenetically-corrected correlation ρ_C: weighted Spearman rank correlation $\rho_S = 0.26$ ($p < 10^{-16}$) for mean degree and $\rho_S = -0.09$ ($p = 0.00131$) for mean ancestral gene expression. This suggests both metrics are poor predictors of the strength of coevolution of gene expression between protein pairs.

Coevolution of gene expression weakly reflects coevolution of protein sequences

Previous work found an overall weak correlation between coevolution at the protein sequence level and coevolution at the gene expression level based on CAI [12, 7]. Using estimates of protein sequence coevolution across a yeast phylogeny taken from [12], we found protein

sequence coevolution and the phylogenetically-corrected correlations ρ_C were weakly, but significantly correlated ($\rho_S = 0.10$, 95% CI: 0.037 – 0.155, $p = 0.0015$). We also found a significant correlation between our phylogenetically-corrected correlation ρ_C and the measure of gene expression coevolution from [12] ($\rho_S = 0.22$, 95% CI: 0.171 – 0.275, $p < 10^{-16}$). Notably, we find overall better agreement between CAI and empirical-based measures of coevolution for protein pairs which are, on average, more highly expressed (Weighted Spearman Rank correlation $\rho_S = -0.12$, 95% CI: -0.176 – -0.065, $p < 10^{-4}$, Supplemental Figure 4). This is unsurprising, given that many highly expressed genes are likely to be housekeeping genes, such as ribosomal proteins, and thus highly expressed across most conditions and evolutionary time, making CAI a reliable proxy for gene expression in these cases.



(a) (b)

Figure 5: (a) Comparing coevolution of gene expression, represented by the phylogenetically-corrected correlation ρ_C, and protein sequences, taken from Clark et al [12]. There is a weak but significant correlation (Weighted Spearman Rank Correlation $\rho_S = 0.10$, $p = 0.0015$) between the measures of gene expressions and protein sequence coevolution. (b) A similar comparison using the measures of CAI coevolution from [12]. Again, there is a weak, but significant correlation (Weighted Spearman Rank correlation $\rho_S = 0.22$, $p < 10^{-16}$).

Discussion

Consistent with previous results, we find physically- interacting proteins show a clear signal of gene expression coevolution compared to randomly-generated pairs of proteins, with mean phylogenetically-corrected correlations $\bar{\rho}_C$ of 0.45 vs. 0.03, respectively. We find interacting proteins are correlated with the STRING confidence score (weighted Spearman Rank correlation $\rho_S = 0.32$), indicating protein-protein interactions with stronger evidence of being true and conserved have stronger coevolution of gene expression, on average. We also find the number of protein-protein interactions a protein is involved in and its gene expression level – two common metrics known to affect the evolution of protein sequence – are overall weak predictors of gene expression coevolution. Protein pairs involved in more interactions do tend to show stronger gene expression coevolution (weighted Spearman rank correlations $\rho_S = 0.26$), consistent with the idea that proteins involved in more interactions in a protein-protein interaction network have more constraints on the evolution of their gene expression. Surprisingly, highly expressed protein pairs actually tended to show weaker coevolution of gene expression (weighted Spearman rank correlation $\rho_S = -0.09$). We also find an overall weak correlation between gene expression coevolution and protein sequence coevolution (weighted Spearman rank correlation $\rho_S = 0.10$), consistent with previous work [12, 7]. This is likely because relatively small regions of two protein sequences may be important for the proteins to be able to bind, forcing strong sequence coevolution at the binding sites, but weaker coevolution for the remainder of the protein sequences. Surprisingly, there was overall poor agreement between CAI coevolution from [12] and our measure of of gene expression coevolution based on empirical RNA-Seq data (weighted Spearman Rank correlation $\rho_S = 0.22$). The stronger correlation between ρ_C and CAI coevolution compared to protein sequence coevolution is unsurprising. CAI and similar codon usage metrics often show moderate to strong correlations with empirical gene expression estimates. However, the correlation between ρ_C and CAI coevolution is still very weak, indicating these measures of gene expression coevolution can give radically different interpretations about the degree of

gene expression coevolution at the individual protein-pair level. It is worth noting that our estimates of gene expression coevolution and the estimates from [12] do not come from the same 18 species. [12] also used 18 fungal species, 11 of which are from the *Saccharmoyces* or *Candida* genera, of which 7 overlap with the species used in this study. This undoubtedly introduced noise into these comparisons. Additionally, empirical gene expression in inherently noisy and can vary between conditions, which also will also increase the discrepancy between CAI and gene expression, particularly for low to moderate expression genes. Fortunately, many PCMs allow for the incorporation of measurement error, and through the use of multivariate methods, gene expression estimates across conditions can also be compared by treating each condition as an additional trait.

Unlike previous approaches, our results are based on a multivariate PCM and empirical gene expression data. This offers two clear advantages. The first is our approach accounts for the phylogeny, recognizing the non-independence of species. The second is although CAI is often well-correlated with gene expression in organisms with a high effective population size [51], many low effective population size species show little adaptive codon usage bias, making CAI a poor proxy for gene expression. As a result, the use of empirical gene expression measurements are highly valuable for studying the evolution of gene expression, as others have noted [1].

As with previous work, our results suggest this multivariate PCM could be used to identify functionally-related proteins. However, simulations indicate more species might be needed to have sufficient statistical power (see Table 1). Additionally, it is possible to use this multivariate approach to correlate changes in gene expression with changes in other phenotypes, such as body size (see [40] for more details on using **mvMORPH**). With that in mind, recent work finds multivariate PCMs are in need of improvement, as parameter estimation accuracy decreases quickly as the number of traits (i.e. parameters) increases [52]. For now, it appears best to restrict the analysis to as few traits as possible.

We note very few traits in biology likely evolve in a true Brownian Motion manner [10].

Consistent with this, most of the genes in our dataset violated the BM assumption based on the test proposed by [37]. Although the Ornstein-Uhlenbeck (OU) model may be a more appropriate model, and is, in fact, used in many other PCMs for examining gene expression evolution, it often requires more species to make accurate parameter estimates. As we only used 18 fungal species, we opted to use the simpler Brownian Motion model combined with filtering of genes which significantly deviated from assumptions of the Brownian Motion model [37]. Future work should focus on the examination of coevolution of gene expression using the OU model. Based on our results, inclusion of genes which violate the BM assumption does not change overall conclusions of this work, but it does appear to weaken some of the observed signals (see Supplemental Results). Other phylogenetic models, such as the OU model, may be more appropriate for the analysis of gene expression for other gene pairs. A major advantage of PCMs is other models can easily be incorporated into the analysis of the trait, with the best model being determined via a hypothesis testing (e.g. Likelihood ratio test) or model comparison (e.g. AIC) framework.

We also note comparison of RNA-Seq data across species presents its own challenges [33, 1, 53]. Other methods for normalizing RNA-Seq data in order to make them comparable across species have been proposed, but to the best of our knowledge, there is no current consensus on the best approach. [17] developed a method for normalizing gene expression by identifying the genes with the most conserved ranks across samples, calculating species-specific scaling factors to make the median expression of these conserved rank genes equal across all species, and using those scaling factors to re-scale all gene expression estimates. [1] proposed a method based on comparing fold-changes (differential expression) across species-specific samples, which assumes a clear control and experimental condition and these measurements exists for all species under consideration. [53] proposed a method for re-scaling the TPM metric based on the largest genome in the dataset, but this assumes the genes represented in the smaller genomes are subsets of the genes in the larger genome, which was not the case for our data based on the orthologs we identified. We transformed

species-level data to a standard lognormal distribution, consistent with previous work using microarray data [16]. We make no claims that this is the best normalization approach for across species comparisons of RNA-Seq data, but state it was suitable for our overall purpose of determining if functionally-related genes show stronger coevolution of gene expression than randomly-generated pairs. Additionally, the log-transformed data likely better reflects the assumptions of Brownian Motion, as this eliminates the lower bound of 0.0 [37]. Furthermore, the RNA-Seq data used in this study were pulled from various non-related experiments which differed in terms of protocols, sequencers, sequencing depth, read type (single vs. paired), experimental conditions, and other factors which could impact the quantifications. It cannot be understated that this also introduces large amounts of variability to the quantified RNA-Seq data, making comparisons across species even more difficult. We attempted to control for this by using Salmon's abilities to automatically adjust quantifications based on biases its detects within the RNA-Seq reads, as well as using the control conditions for each species for our analysis. Undoubtedly, this did not control for all of the variability introduced by pulling data from different experiments. Despite this, the normalized gene expression data were moderately to strongly correlated across species (Supplemental Figure 1) and species which were more closely related tended to show higher correlations, consistent with expectations. Our results regarding coevolution of gene expression were also mostly consistent with previous work and expectations. However, analyses attempting to make more precise conclusions about the evolution or coevolution of gene expression should ideally use measurements produced under better controlled conditions. Additionally, future efforts in this area may consider using proteomics data as opposed to transcriptomics. Previous work finds protein abundances appear to be more conserved between species compared to mRNA abundances, which could indicate stronger selection on maintaining the former [54].

Given our results and the ease of use of many tools implementing PCMs, we strongly recommend the use of PCM approaches when performing interspecies analysis. The phylogenetic research community has databases where phylogenetic trees can be easily accessed,

such as TreeBase [55]. If a phylogenetic tree is not available for the species of interest, multiple sequence alignment tools and phylogenetic tree estimation tools have made building a reasonable phylogenetic tree efficient and easy, even for non-computational researchers. The phylogenetics community has made access to complex phylogenetic parameter estimation accessible via open-source, easy-to-use R packages, such as **mvMORPH** [40]. Although we strongly recommend the use of PCMs for interspecies data analysis, we emphasize that such approaches come with their own challenges and, in some cases, the PCM may not perform better than standard statistical approaches (see [41] for more details). Even so, approaches for assessing the impact of shared ancestry on the data still requires the generation of a phylogenetic tree and analysis of the trait in a phylogenetic context. Some researchers have also suggested PCMs likely will not provide different results from non-PCMs if analyzing gene expression for a small number of species, with a larger number of species resulting in more complex phylogenetic patterns and complicating the downstream data analyses [23]. Researchers should assess the impact of phylogeny of their data and make the appropriate decisions on what tools best answer the questions at hand.

Acknowledgments

Project was originally developed by A.L Cope and B.C. O'Meara as part of the NSF-funded Phylogenetic Methods course (NSF BIO 1453424) taught at the University of Tennessee, Knoxville. Financial support provided by by NSF grant MCB-1546402 (A. VonArnim and M.A. Gilchrist), NSF grant MCB-1120370 (M.A. Gilchrist), DEB-1355033 (B.C. O'Meara, M.A. Gilchrist, and R. Zaretzki), the Graduate School of Genome Science and Technology (University of Tennessee), and the U.S. Department of Energy, Biological and Environmental program through funding of the Center for Bioenergy Innovation at the Oak Ridge National Laboratory. ORNL is managed by the UT – Battelle, LLC for the U.S. Department of Energy (DOE). Additional support was provided by the National Institute for Mathematical and

Biological Synthesis (NSF:DBI-1300426) and the Dept. of Eco/Evol. Biology (University of Tennessee).

References

[1] Dunn CW, Luo X, Wu Z. Phylogenetic Analysis of Gene Expression. Integr Comp Biol. 2013;53(5):847–856. doi:10.1093/icb/ict068.

[2] Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci. 1998;95(25).

[3] Grigoriev A. A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast Saccharomyces cerevisiae. Nucleic Acids Res. 2001;29(17):3513–3519. doi:10.1093/nar/29.17.3513.

[4] Gillis J, Pavlidis P. The Impact of Multifunctional Genes on "Guilt by Association" Analysis. PLoS One. 2011;6(2):e17258. doi:10.1371/journal.pone.0017258.

[5] Michalak P. Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. Genomics. 2008;91(3):243–248. doi:10.1016/J.YGENO.2007.11.002.

[6] Ge H, Liu Z, Church GM, Vidal M. Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. Nat Genet. 2001;29(4):482–486. doi:10.1038/ng776.

[7] Fraser HB, Hirsh AE, Wall DP, Eisen MB. Coevolution of gene expression among interacting proteins. Proc Natl Acad Sci U S A. 2004;101(24):9033–8. doi:10.1073/pnas.0402591101.

[8] Sharp PM, Li W. The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential applications. Nucl Acids Res. 1987;15(3):1281–1295.

[9] Lithwick G, Margalit H. Relative predicted protein levels of functionally associated proteins are conserved across organisms. Nucleic Acids Res. 2005;33(3):1051–1057. doi:10.1093/nar/gki261.

[10] J Felsenstein. Phylogenies and the Comparative Method on JSTOR. Am Nat. 1985;125(1):1–15.

[11] Dunn CW, Zapata F, Munro C, Siebert S, Hejnol A. Pairwise comparisons across species are problematic when analyzing functional genomic data. Proc Natl Acad Sci U S A. 2018;115(3):E409–E417. doi:10.1073/pnas.1707515115.

[12] Clark NL, Alani E, Aquadro CF. Evolutionary rate covariation reveals shared functionality and coexpression of genes. Genome Res. 2012;22(4):714–20. doi:10.1101/gr.132647.111.

[13] Sato T, Yamanishi Y, Kanehisa M, Toh H. The inference of protein-protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. Bioinformatics. 2005;21(17):3482–3489. doi:10.1093/bioinformatics/bti564.

[14] Martin T, Fraser HB. Comparative expression profiling reveals widespread coordinated evolution of gene expression across eukaryotes. Nat Commun. 2018;9(1):4963. doi:10.1038/s41467-018-07436-y.

[15] Barker D, Pagel M. Predicting Functional Gene Links from Phylogenetic-Statistical Analyses of Whole Genomes. PLoS Comput Biol. 2005;1(1):e3. doi:10.1371/journal.pcbi.0010003.

[16] Bedford T, Hartl DL. Optimization of gene expression by natural selection. Proc Natl Acad Sci U S A. 2009;106(4):1133–8. doi:10.1073/pnas.0812009106.

[17] Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, et al. The evolution of gene expression levels in mammalian organs. Nature. 2011;478(7369):343–348. doi:10.1038/nature10532.

[18] Eng KH, Bravo HC, Keleş S. A Phylogenetic Mixture Model for the Evolution of Gene Expression. Mol Biol Evol. 2009;26(10):2363–2372. doi:10.1093/molbev/msp149.

[19] Gu X, Zou Y, Huang W, Shen L, Arendsee Z, Su Z. Phylogenomic Distance Method for Analyzing Transcriptome Evolution Based on RNA-seq Data. Genome Biol Evol. 2013;5(9):1746–1753. doi:10.1093/gbe/evt121.

[20] Liang C, Musser JM, Cloutier A, Prum RO, Wagner GP. Pervasive Correlated Evolution in Gene Expression Shapes Cell and Tissue Type Transcriptomes. Genome Biol Evol. 2018;10(2):538–552. doi:10.1093/gbe/evy016.

[21] Oakley TH, Gu Z, Abouheif E, Patel NH, Li WH. Comparative Methods for the Analysis of Gene-Expression Evolution: An Example Using Yeast Functional Genomic Data. Mol Biol Evol. 2005;22(1):40–50. doi:10.1093/molbev/msh257.

[22] Rohlfs RV, Harrigan P, Nielsen R. Modeling Gene Expression Evolution with an Extended Ornstein–Uhlenbeck Process Accounting for Within-Species Variation. Mol Biol Evol. 2014;31(1):201–211. doi:10.1093/molbev/mst190.

[23] Rohlfs RV, Nielsen R. Phylogenetic ANOVA: The Expression Variance and Evolution Model for Quantitative Trait Evolution. Syst Biol. 2015;64(5):695–708. doi:10.1093/sysbio/syv042.

[24] Schraiber JG, Mostovoy Y, Hsu TY, Brem RB. Inferring Evolutionary Histories of Pathway Regulation from Transcriptional Profiling Data. PLoS Comput Biol. 2013;9(10):e1003255. doi:10.1371/journal.pcbi.1003255.

[25] Butler MA, King AA. Phylogenetic Comparative Analysis: A Modeling Approach for Adaptive Evolution. Am Nat. 2004;164(6):683–695. doi:10.1086/426002.

[26] Hansen TF. STABILIZING SELECTION AND THE COMPARATIVE ANALYSIS OF ADAPTATION. Evolution (N Y). 1997;51(5):1341–1351. doi:10.1111/j.1558-5646.1997.tb01457.x.

[27] Bartoszek K, Pienaar J, Mostad P, Andersson S, Hansen TF. A phylogenetic comparative method for studying multivariate adaptation. J Theor Biol. 2012;314:204–215. doi:10.1016/J.JTBI.2012.08.005.

[28] Revell LJ, Collar DC. PHYLOGENETIC ANALYSIS OF THE EVOLUTIONARY CORRELATION USING LIKELIHOOD. Evolution (N Y). 2009;63(4):1090–1100. doi:10.1111/j.1558-5646.2009.00616.x.

[29] Revell LJ, Harmon LJ. Testing quantitative genetic hypotheses about the evolutionary rate matrix for continuous characters; 2008. Available from: `http://www.evolutionary-ecology.com/issues/v10n03/ccar2235.pdf`.

[30] Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019;47(D1):D607–D613. doi:10.1093/nar/gky1131.

[31] Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14(4):417–419. doi:10.1038/nmeth.4197.

[32] Scannell DR, Zill OA, Rokas A, Payen C, Dunham MJ, Eisen MB, et al. The Awesome Power of Yeast Evolutionary Genetics: New Genome Sequences and Strain Resources for the Saccharomyces sensu stricto Genus. G3 (Bethesda). 2011;1(1):11–25. doi:10.1534/g3.111.000273.

[33] Wagner GP, Kin K, Lynch VJ. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory Biosci. 2012;131(4):281–285. doi:10.1007/s12064-012-0162-3.

[34] Yang JR, Maclean CJ, Park C, Zhao H, Zhang J. Intra and Interspecific Variations of Gene Expression Levels in Yeast Are Largely Neutral: (Nei Lecture, SMBE 2016, Gold Coast). Mol Biol Evol. 2017;34(9):2125–2139. doi:10.1093/molbev/msx171.

[35] Brion C, Pflieger D, Souali-Crespo S, Friedrich A, Schacherer J. Differences in environmental stress response among yeasts is consistent with species-specific lifestyles. Mol Biol Cell. 2016;27(10):1694–1705. doi:10.1091/mbc.E15-12-0816.

[36] Abascal F, Zardoya R, Telford MJ. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. Nucleic Acids Res. 2010;38(suppl_2):W7–W13. doi:10.1093/nar/gkq291.

[37] Garland T, Harvey PH, Ives AR. Procedures for the Analysis of Comparative Data Using Phylogenetically Independent Contrasts. Syst Biol. 1992;41(1):18–32. doi:10.1093/sysbio/41.1.18.

[38] Smith SA, O'Meara BC. treePL: divergence time estimation using penalized likelihood for large phylogenies. Bioinformatics. 2012;28(20):2689–2690. doi:10.1093/bioinformatics/bts492.

[39] Kumar S, Stecher G, Suleski M, Hedges SB. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Mol Biol Evol. 2017;34(7):1812–1819. doi:10.1093/molbev/msx116.

[40] Clavel J, Escarguel G, Merceron G. mv morph : an r package for fitting multivariate evolutionary models to morphometric data. Methods Ecol Evol. 2015;6(11):1311–1319. doi:10.1111/2041-210X.12420.

[41] Revell LJ. Phylogenetic signal and linear regression on species data. Methods Ecol Evol. 2010;1(4):319–329. doi:10.1111/j.2041-210X.2010.00044.x.

[42] Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 2019;35(3):526–528. doi:10.1093/bioinformatics/bty633.

[43] Drummond DA, Wilke CO. Mistranslation-Induced Protein Misfolding as a Dominant Constraint on Coding-Sequence Evolution. Cell. 2008;134:341–352.

[44] Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW. Evolutionary rate in the protein interaction network. Science. 2002;296(5568):750–2. doi:10.1126/science.1068696.

[45] Canty A, Ripley BD. boot; 2019.

[46] Davison AC, Hinkley DV. Bootstrap Methods and Their Applications. Cambridge: Cambridge University Press; 1997. Available from: `http://statwww.epfl.ch/davison/BMA/`.

[47] Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. Why highly expressed proteins evolve slowly. Proc Natl Acad Sci. 2005;102(40):14338–14343.

[48] Drummond DA, Raval A, Wilke CO. A Single Determinant Dominates the Rate of Yeast Protein Evolution. Mol Biol Evol. 2005;23(2):327–337.

[49] Feyertag F, Berninsone PM, Alvarez-Ponce D. Secreted Proteins Defy the Expression Level-Evolutionary Rate Anticorrelation. Mol Biol Evol. 2017;34(3):692–706. doi:10.1093/molbev/msw268.

[50] Gilchrist MA, Chen WC, Shah P, Landerer CL, Zaretzki R. Estimating Gene Expression and Codon-Specific Translational Efficiencies, Mutation Biases, and Selection Coefficients from Genomic Data Alone. Genome Biol Evol. 2015;7:1559–1579.

[51] Charlesworth B. Effective population size and patterns of molecular evolution and variation. Nat Rev Genet. 2009;10:195–205. doi:10.1038/nrg2526.

[52] Adams DC, Collyer ML. Multivariate Phylogenetic Comparative Methods: Evaluations, Comparisons, and Recommendations. Syst Biol. 2018;67(1):14–31. doi:10.1093/sysbio/syx055.

[53] Musser JM, Wagner GP. Character Trees From Transcriptome Data: Origin and Individuation of Morphological Characters and the So-Called "Species Signal". J Exp Zool (Mol Dev Evol). 2015;324:588–604. doi:10.1002/jez.b.22636.

[54] Laurent JM, Vogel C, Kwon T, Craig SA, Boutz DR, Huse HK, et al. Protein abundances are more conserved than mRNA abundances across diverse taxa. Proteomics. 2010;10(23):4209–4212. doi:10.1002/pmic.201000327.

[55] Piel WH, Donoghue M, Sanderson M. TreeBASE: A database of phylogenetic information. In: e-Biosphere; 2009.Available from: `http://phylogeny.harvard.edu/treebase`.