Cope *et al. BMC Genomics* (2020) 21:370
https://doi.org/10.1186/s12864-020-6761-3

BMC Genomics

RESEARCH ARTICLE Open Access

Gene expression of functionally-related genes coevolves across fungal species: detecting coevolution of gene expression using phylogenetic comparative methods

Alexander L. Cope[1,2*], Brian C. O'Meara[3,4] and Michael A. Gilchrist[1,3,4]

Abstract

Background: Researchers often measure changes in gene expression across conditions to better understand the shared functional roles and regulatory mechanisms of different genes. Analogous to this is comparing gene expression across species, which can improve our understanding of the evolutionary processes shaping the evolution of both individual genes and functional pathways. One area of interest is determining genes showing signals of coevolution, which can also indicate potential functional similarity, analogous to co-expression analysis often performed across conditions for a single species. However, as with any trait, comparing gene expression across species can be confounded by the non-independence of species due to shared ancestry, making standard hypothesis testing inappropriate.

Results: We compared RNA-Seq 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 physically-interact show stronger signals of coevolution than randomly-generated pairs. Interactions with 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 in a standard hypothesis testing framework. Furthermore, our simulations indicate the use of a randomly-generated null distribution as a means of determining statistical significance for detecting coevolving genes with phylogenetically-uncorrected correlations, as has previously been done, is less accurate than PCMs, although is a significant improvement over standard hypothesis testing. These methods are further improved by using a phylogenetically-corrected correlation metric.

Conclusions: Our work highlights potential benefits of using PCMs to detect gene expression coevolution from
(Continued on next page)

*Correspondence: acope3@vols.utk.edu
[1]Genome Science and Technology, University of Tennessee, Knoxville, Tennessee, USA
[2]Chemical Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA
Full list of author information is available at the end of the article

(Continued from previous page)
high-throughput omics scale data. This framework can be built upon to investigate other evolutionary hypotheses, such as changes in transcription regulatory mechanisms across species.

Keywords: Gene expression, Coevolution, Phylogenetic comparative methods

Background

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 [2–5]. 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.

Previous work supports the hypothesis that gene expression of functionally-related genes shows stronger signals of coevolution than randomly-generated gene pairs in both unicellular yeasts and a diverse set of prokaryotes. [7–9]. Interestingly, the strength of this signal appeared to vary based on the functional groupings of the genes in question [7]. Fraser et. al. [8] proposed gene expression coevolution could be a useful method for predicting proteins which are functionally-related.

Most of the previous work examining coevolution of gene expression relied upon the Codon Adaptation Index (CAI) [10] as a proxy for gene expression. CAI and other codon-usage metrics often correlate well with gene expression in many species, but this is often not the case in species with a strong mutational bias or low effective population sizes, as is the case in many multicellular eukaryotes [11]. In fact, Lithwick and Margalit [9] were forced to eliminate organisms from their analysis which showed little adaptive codon usage. This makes detecting signals from empirical measures of gene expression, such as from RNA-Seq or mass spectrometry data, particularly useful for many species where codon usage metrics are a poor proxy for gene expression. Recent work by Martin and Fraser [12] demonstrated a method for examining coevolution of gene expression within sets of functionally-related genes using RNA-Seq data measured from the Marine Microbial Eukaryotic Transcriptome Project [13].

While it may seem appropriate to simply assess the correlation (e.g. Pearson or Spearman) between gene expression estimates across species, much like one might do in a co-expression analysis across conditions, an issue that arises is the non-independence of species due to shared ancestry [14]. This can result in biases in correlation coefficients and lead to an inflation of the degrees of freedom, making standard hypothesis testing inappropriate [14, 15]. Recent work concluded comparative analysis of gene expression data across species can be confounded by the phylogeny, leading potentially to incorrect inferences [16]. Previous work examining coevolution of gene expression did not directly account for the phylogeny when estimating correlation coefficients of gene expression across species, which is thought to reflect the strength of coevolution between gene pairs. With the exception of Clark et. al. [7], who applied a transformation to their correlation coefficients originally developed to eliminate phylogenetic signal from sequence coevolution data [17], much of the previous work used a randomly-generated null distribution created from genes not thought to coevolve as a means of determining a statistical significance cutoff. Although the use of a randomly-generated null is likely a better alternative than standard hypothesis testing, a direct assessment of these approaches' abilities to adequately control for the phylogeny have not been determined, to the best of our knowledge.

An alternative solution is to directly account for the phylogeny when assessing coevolution between pairs of genes using phylogenetic comparative methods (PCMs). Previous efforts have developed PCMs for examining coevolution of functionally-related genes based on the presence/absence of genes across species. Barker and Pagel [18] developed 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, although this work has not focused on detecting coevolution of gene expression Bedford et al. [19–27]. Much of this work relies on modeling gene expression evolution as an Ornstein-Uhlenbeck (OU) process [28, 29]. Modeling trait evolution as an OU process assumes the trait is evolving around an optimal value. A multivariate version of the OU model exists

[30], but the additional parameters used in the model often requires a greater amount of species-level data to make accurate parameter estimates. Here, we present an approach which models the coevolution of gene expression, as estimated via RNA-Seq, for pairs of proteins using the simpler multivariate Brownian Motion (BM) model [31, 32]. 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. We also find phylogenetically-uncorrected correlations tend to inflate estimates of gene expression coevolution. Unsurprisingly, simulations reveal standard hypothesis testing (i.e. $p < 0.05$) using phylogenetically-uncorrected correlations inflates the false discovery rate. We find determing statistical significance via a randomly-generated null distribution, as described in Fraser et. al. [8] is a significant improvement over standard hypothesis testing, but still performs worse than the PCM approach. The method recently described by Martin and Fraser [12] was able to obtain a low false discovery rate, but this came at the expense of statistical power to detect coevolving genes relative to the PCM, which had a comparable false discovery rate.

We expand upon previous work by looking for potential predictors reflecting the strength of coevolution between two pairs of proteins. As expected, we find protein pairs with stronger evidence of functional-relatedness show stronger coevolution at the gene expression level. We also find gene expression level and the number of protein interactions, which are considered good predictors of evolutionary rate of a gene [33], are poor predictors of the strength of coevolution between protein pairs. Consistent with previous results, we also find coevolution of gene expression is an overall weak predictor of protein sequence coevolution.

Results

The phylogenetic tree used in our analysis is shown in Fig. 1. Overall, the normalized gene expression data are moderately to strongly correlated between all species (Additional file 1, Figure S1). 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 (proteins with evidence of physically interacting, which we expect to show signals of coevolution) and control datasets (randomly-generated pairs not expected to show signals of coevolution) 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 appears weaker (Additional file 1, Figure S6 – S9). Our results are also robust to our use of the time-calibrated ultrametric tree output from treePL or the non-calibrated tree output from RAxML (Additional file 1, Figure S11).

Interacting proteins demonstrate clear coevolution of gene expression

To broadly examine coevolution of gene expression between physically-interacting proteins, a phylogenetically-corrected Covariance Ratio test (as implemented in the R package **geomorph** [34–36]) was applied to protein modules found within the protein-protein interaction network (see Methods). We found covariance between gene expression was, on average, greater within protein interaction modules compared to between modules (Covariance Ratio score = 0.8672, $p = 0.001$). This indicates gene expression within tightly-linked groups of physically-interacting proteins show greater signals of coevolution than between proteins which spuriously interact.

Gene expression evolution was modeled as a multivariate Brownian Motion (BM) process using the R package **mvMORPH** [37] in order to estimate coevolution of gene expression between pairs of proteins. This approach provides an estimate of the degree of correlation between two traits (in this case, our estimates of gene expression) across species that accounts for the phylogeny (see Methods for more details). We will refer to this correlation estimate as the phylogenetically-corrected correlation ρ_C. The phylogenetically-corrected correlation ρ_C distributions for the binding and control groups show striking differences (Fig. 2). Binding proteins have a mean phylogenetically-corrected correlation of $\bar{\rho}_C = 0.45$, which is significantly different from the expected value of 0.0 if there was no coevolution of gene expression (One-sample t-test, 95% CI: 0.436 – 0.464, $p < 10^{-200}$). In contrast, the randomly-generated control group, which is not expected to show signals of coevolution, had a much lower (but still significant) mean phylogenetically-corrected correlation of $\bar{\rho}_C = 0.03$ (One-sample t-test, 95% CI: 0.025 – 0.037, $p < 10^{-23}$). Although the mean phylogenetically-corrected correlation for the control group is significantly different from 0.0, it is important to note two things: (1) even though we did our best to eliminate possible false negatives in the control group, it

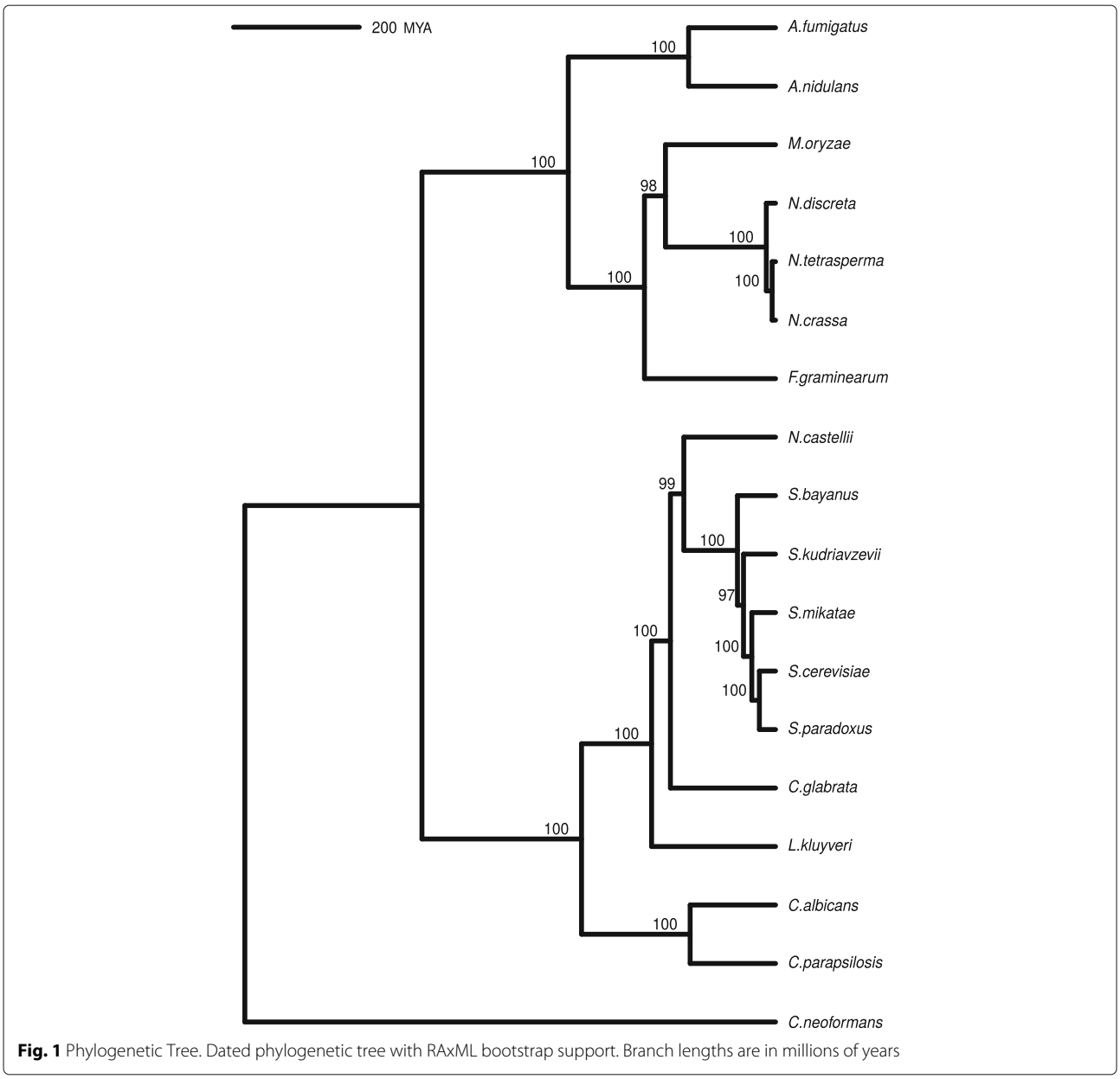

Fig. 1 Phylogenetic Tree. Dated phylogenetic tree with RAxML bootstrap support. Branch lengths are in millions of years

is unlikely all false negatives were eliminated and (2) this is consistent with previous work by Fraser et. al. [8], who also had random control groups which were not centered around 0. 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, $p < 10^{-200}$). 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, $p < 10^{-37}$, see Methods), indicating interactions which are more likely to be true and conserved show stronger coevolution of gene expression (Fig. 3). A similar result is obtained when using a metric of functional similarity between proteins based on overlapping Gene Ontology terms (Additional file 1, Figure S2).

We also compared our phylogentically-corrected approach to a phylogenetically-uncorrected approach ρ_U (the Pearson correlation coefficient, Fig. 2). Qualitatively, a similar pattern to the phylogenetically-corrected correlations ρ_C is observed: binding proteins show correlations positively skewed away from 0, consistent

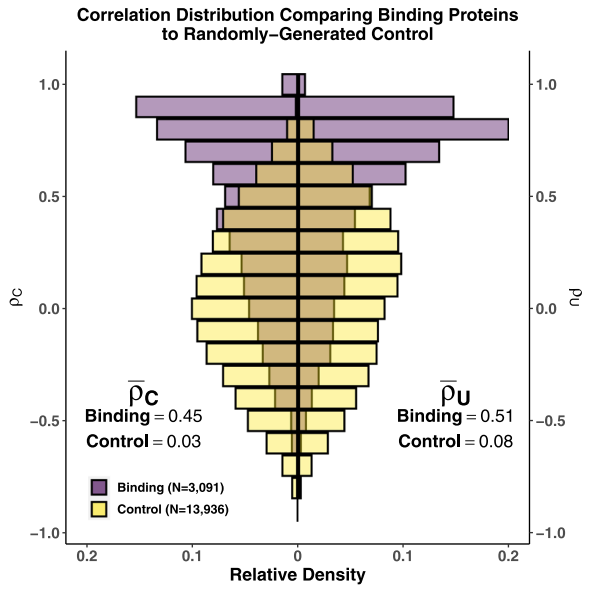

Fig. 2 Comparing phylogenetically-corrected and uncorrected correlations. Comparing the distributions of the (Left) phylogenetically-corrected correlation ρ_C and the (Right) phylogenetically-uncorrected correlation ρ_U for the binding (purple) and control (yellow) groups. (Left) Mean values for the 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

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, $p < 10^{-200}$). 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, $p < 10^{-141}$). 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 potential biasing of correlation estimates due to treatment of non-independent species data as independent [14, 15].

Simulations were performed to confirm potential problems with the use of non-phylogenetic methods for comparing gene expression across species (see Additional file 1). Results show failure to account for the phylogeny on data simulated under the null hypothesis of no coevolution between gene expression results in an increase in the false discovery rate (FDR, Table 1), consistent with expectations. However, the distribution of ρ_U simulated under no coevolution differs from the distribution of ρ_U from the real data (Additional file 1, Figure S5). 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

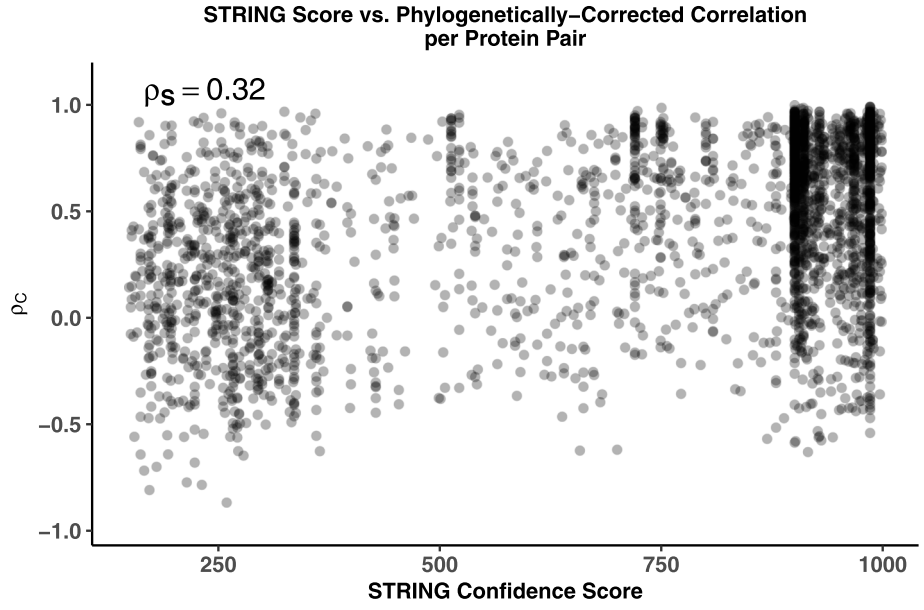

Fig. 3 Effects of functional-relatedness on phylogenetically-corrected correlation ρ_C. Positive weighted Spearman rank correlation ($\rho_S = 0.32$, $p < 10^{-37}$) between the STRING score and phylogenetically-corrected correlation ρ_C indicates more confident and/or conserved interactions tend to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

Table 1 Highlighting issues with not correcting for phylogeny

Method	Correlation	TPR (S.D)	FPR (S.D)	FDR (S.D)	Overall Accuracy (S.D)
multivariate BM PCM ($p < 0.05$)	ρ_C	0.476 (0.0004)	0.026 (0.0030)	0.053 (0.0056)	0.725 (0.0013)
cor.test() ($p < 0.05$)	ρ_U	0.574 (0.0006)	0.209 (0.0075)	0.267 (0.0068)	0.682 (0.0035)
Fraser et. al. [8]	ρ_C	0.567 (0.0363)	0.053 (0.0152)	0.084 (0.0212)	0.757 (0.0148)
	ρ_U	0.511 (0.0432)	0.097 (0.0285)	0.156 (0.0316)	0.708 (0.0144)
Martin and Fraser [12]	ρ_C	0.476 (0.0108)	0.025 (0.0008)	0.050 (0.0010)	0.726 (0.0051)
	ρ_U	0.305 (0.0155)	0.016 (0.0010)	0.050 (0.0015)	0.644 (0.0073)

Comparison of 4 methods for detecting coevolution of gene expression using data simulated under Brownian Motion. The 4 methods represent the multivariate Brownian Motion (BM) PCM described in this manuscript, hypothesis testing with the phylogenetically-uncorrected correlation, the method described in Fraser et. al. [8], and the method described in Martin and Fraser [12]. Mean and standard deviations for true positive rates (TPR), false positive rates (FPR), false discovery rate (FDR), and overall accuracy are reported

shows a broadening of the distribution compared to the phylogenetically-corrected correlations ρ_C.

Instead of determining statistical significance for the phylogenetically-uncorrected correlations ρ_U using $p < 0.05$, we used approaches similar to those described by Fraser et. al [8] and Martin and Fraser [12]. We found the method described in Fraser et. al. to have a greater true positive rate (TPR) compared to the PCM (0.511 compared to 0.476), but still had an inflated false discovery rate (FDR) of 0.156, although this was a significant improvement over standard hypothesis testing (Table 1). An approach similar to Martin and Fraser [12] was actually underpowered compared to the PCM, with a true positive rate (TPR) of 0.305, when controlling the FDR to be 0.05. This method had the overall worst accuracy of 0.644. Unsurprisingly, both methods described by Fraser et. al. and Martin and Fraser are improved when using the phylogenetically-corrected correlation ρ_C. When the data is consistent with a Brownian Motion process, methods based on ρ_C are superior to the methods based on ρ_U.

We note these methods all have fairly low true positive rates (TPR). We hypothesized part of this could be due to the presence of false positives in the binding group, which are unlikely to show much coevolution of gene expression, resulting in protein pairs in the simulated data with potentially small effects unlikely to be detected with only 18 species. After excluding potential false positives in the binding group (i.e. protein pairs with a STRING Score <400), the TPR and overall accuracy of all methods increased (Additional file 1, Table S2). However, the general pattern remained the same: when data is consistent with a phylogenetic model of trait evolution (which is the case for our simulations), the methods based on correcting for the phylogeny are superior.

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 [38–42]. 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 $\rho_S = 0.26$ (Fig. 4a, 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 $\rho_S = -0.09$ (Fig. 4b, 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 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. 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 (Additional file 1, Figure S3). This could be due to the representation of certain proteins in the

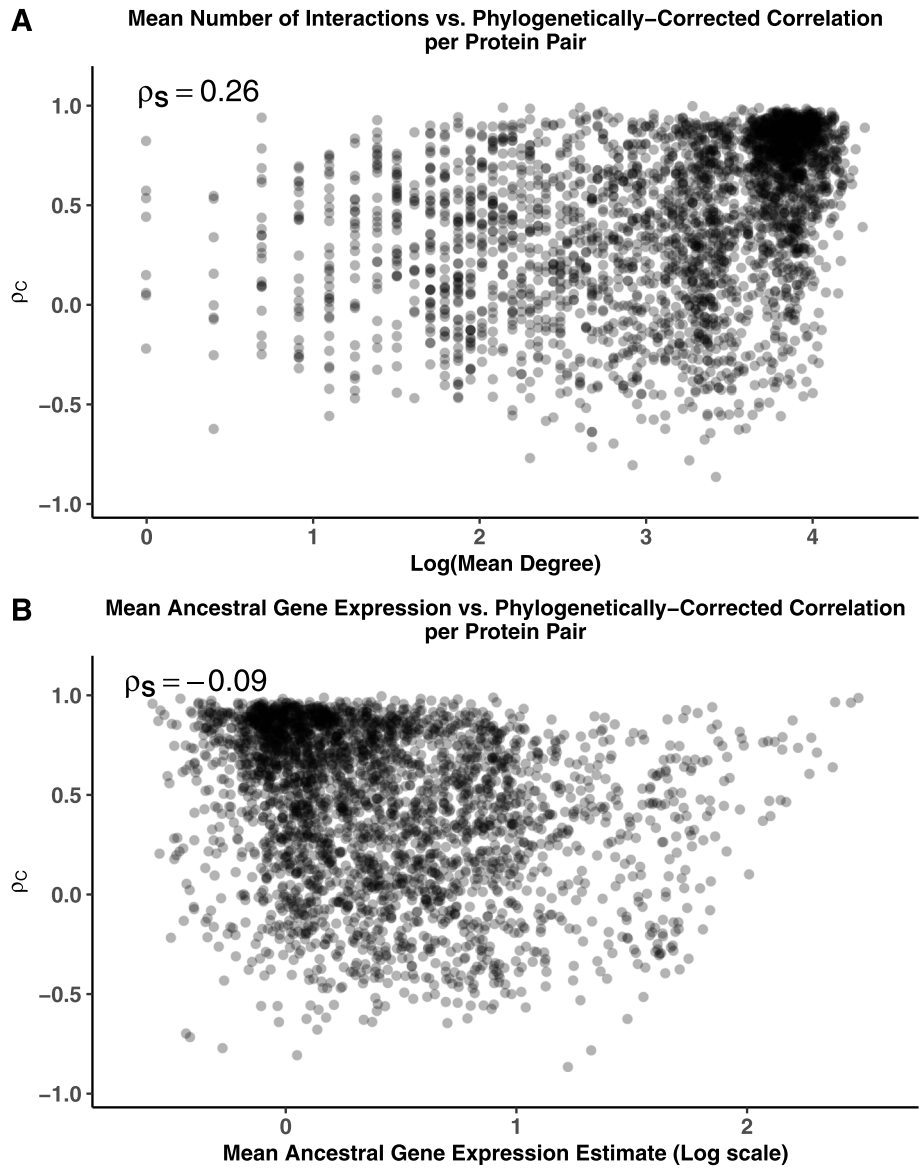

Fig. 4 Effects of number of interactions and gene expression on strength of coevolution. The relationship of **a** the mean degree (average number of interactions between a protein pair) and **b** mean ancestral gene expression estimate with the phylogenetically-corrected correlation ρ_C for the binding 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

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.

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 [7, 8]. Using estimates of protein sequence coevolution across a yeast phylogeny taken from Clark et. al. [7], we found protein sequence coevolution and the phylogenetically-corrected correlations ρ_C were weakly, but significantly correlated (Weighted Spearman Rank correlation $\rho_S = 0.10$, 95% CI: 0.037 – 0.155, $p = 0.0015$, Fig. 5a). We also found a significant correlation between our phylogenetically-corrected correlation ρ_C and the measure of gene expression coevolution from Clark et. al. [7] (Weighted Spearman Rank correlation $\rho_S = 0.22$, 95% CI:

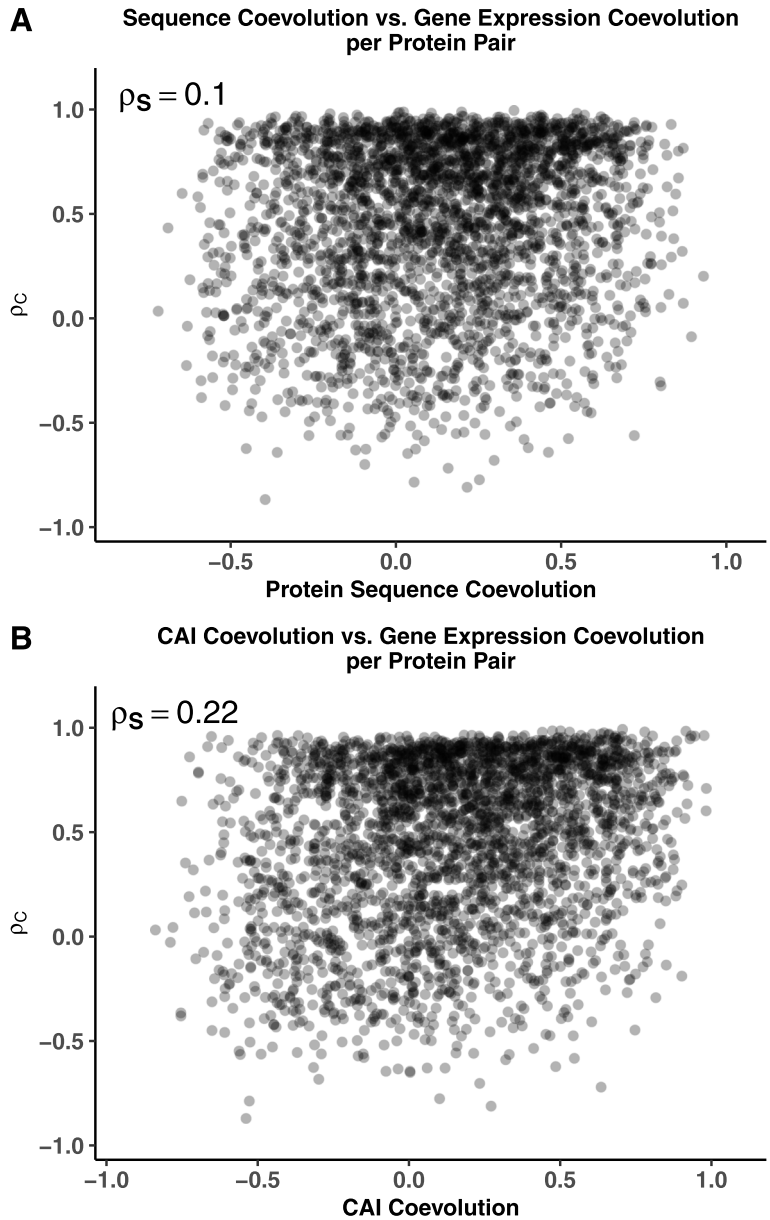

Fig. 5 Comparison to other coevolution metrics. **a** Comparing coevolution of gene expression, represented by the phylogenetically-corrected correlation ρ_C, and protein sequences, taken from Clark et al [7]. There is a weak but significant correlation (Weighted Spearman Rank Correlation $\rho_S = 0.10, p = 0.0015$) between the measures of gene expressions and protein sequence coevolution. **b** A similar comparison using the measures of CAI coevolution from [7]. Again, there is a weak, but significant correlation (Weighted Spearman Rank correlation $\rho_S = 0.22, p < 10^{-16}$)

0.171 – 0.275, $p < 10^{-16}$, Fig. 5b). 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, $p < 10^{-4}$, Additional file 1, Figure S4). 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.

Discussion

A broad-scale analysis based on the Covariance Ratio test [34, 35] found coevolution of gene expression was stronger within groups of tightly-linked protein interactions compared to coevolution between proteins with weaker or no interactions (Covariance Score =

0.8672, $p = 0.001$). Consistent with this, 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 show 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 [7, 8]. We speculate this is 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 Clark et. al. [7] and our measure 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 [7, 8, 42–44]. 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 [7] do not come from the same 18 species. Clark et. al. [7] also used 18 fungal species, 11 of which are from the *Saccharomyces* or *Candida* genera, of which 7 overlap with the species used in this study. This undoubtedly introduced noise into these comparisons, but there are additional reasons to expect discrepancies between coevolution estimates based on CAI and empirical gene expression measurements. CAI, as well as other proxies for gene expression based on codon usage, reflect the evolutionary average expression level for a given gene (assuming strength of selection on codon usage scales with gene expression), but this may not reflect expression of a gene for a given experimental treatment [7, 8, 42, 43, 45, 46]. Additionally, empirical gene expression is subject to measurement error, which will also increase the discrepancy between CAI and gene expression, particularly for low to moderate expression genes [42, 47]. Fortunately, many PCMs allow for the incorporation of measurement error of a trait, which can be estimated via experimental replicates. Furthermore, using multivariate PCMs allows for the treatment of gene expression measured under various conditions as separate traits [1].

Unlike previous approaches, our results are based on both a multivariate PCM and empirical gene expression data. This offers two clear advantages. One advantage is our approach directly accounts for the phylogeny, recognizing the non-independence of species, allowing for standard hypothesis testing. Although previous efforts attempted to control for the phylogeny by using randomly-generated null distributions to determine statistical significance for phylogenetically-uncorrected correlations, our simulations indicate these approaches are generally worse than phylogenetic-based approaches if the underlying model of gene expression evolution is consistent with the BM model (Table 1). The second advantage is while CAI often correlates well with gene expression in organisms with a high effective population size [11], low effective population size species often 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].

Our results indicate 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), although this could vary depending on the tree and data in question. In theory, it is possible to expand this approach to test for gene expression coevolution in larger gene sets or correlate changes in gene expression with changes in other phenotypes, such as body size (see [37] 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 [48]. For now, it appears best to restrict the analysis to as few traits as possible when using approaches like **mvMORPH**. Alternative approaches to examine coevolution of gene expression with more than 2 genes include the Covariance Ratio test [34, 35] and the approach described by Adams and Felice using partial least squares [49]. Unlike the Covariance Ratio test, which reflects the degree of coevolution

within modules of traits (in this case, gene expression), the approach described by Adams and Felice tests for coevolution between modules. Another alternative is the method developed by Martin and Fraser [12].

We note very few traits in biology likely evolve in a true Brownian Motion manner [14]. Consistent with this, most of the genes in our dataset violated the BM assumption based on the test proposed by Garland et. al. [50]. Although the Ornstein-Uhlenbeck (OU) model may be a more appropriate model, and is 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 BM model combined with filtering of genes which significantly deviated from the assumptions of BM [50]. 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 patterns (Additional file 1 Figure S6 – S9). These analyses are exactly the same as described above, but includes genes for which gene expression evolution is better described by other models of trait evolution, such as the OU process. Given these models often incorporate additional parameters to describe trait evolution across species, incorrectly using the BM model likely results in inaccurate estimates of ρ_C and a weakening of the some of the patterns we observe when filtering out genes violating the BM assumption. Future work should focus on the examination of coevolution of gene expression using the OU model. 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 [1, 51, 52]. For our analysis, we transformed species-level data to a standard lognormal distribution, consistent with previous work using microarray data [19]. While other methods for normalizing RNA-Seq measurements for across species exist, our results indicate transformation to the standard lognormal was suitable for the purpose of determining if functionally-related genes show stronger coevolution of gene expression than randomly-generated pairs. To the best of our knowledge, there is no current consensus on the best approach for comparing RNA-Seq measurements across species. Brawand et. al. [20] 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. Dunn et. al. [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. Muesser and Wagner [52] 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.

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, we were still able to pick up evolutionary signals indicating coevolution of gene expression. Additionally, the normalized gene expression data used here were moderately to strongly correlated across species (Additional file 1, Figure S1) and species which were more closely related tended to show higher correlations, consistent with 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. Future efforts in this area may consider using proteomics data instead of transcriptomics data. 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 [53].

Finally, our analysis does not directly account for possible discordance between the species tree and the gene trees of the protein pairs used. This was done out of practicality, as **mvMORPH** only takes into account one phylogenetic tree. Although we eliminate one possible source of discordance by removing genes with evidence of gene duplications, other possible sources include introgression, incomplete lineage sorting (ILS), and horizontal gene transfer (HGT) [54]. Removal of protein pairs with genes marked as possible introgression or HGT events from a population genomics study on 1,011 *S. cerevisiae* isolates [55] had little impact on the phylogenetically-corrected correlation ρ_C distributions for the binding and control sets (Additional file 1, Figure S10). Although this does not exclude ILS as a source of discordance, previous work found ILS reduced phylogenetic signal as estimated

by Pagel's λ, which reflects similarity to a BM process [56, 57]. Based on this, we speculate many genes subject to ILS may have been eliminated by filtering out genes inconsistent with the BM process. Further work is needed to understand the effects of ILS and other sources of gene tree discordance on multivariate trait evolution.

Conclusion

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 [58]. 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** [37]. 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 [59] 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. Rohlfs et. al. 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 [26]. Researchers should assess the impact of phylogeny of their data and make the appropriate decisions on what tools best answer the questions at hand.

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 [7, 8, 18]. Consistent with [8] and [18], 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 [60]. We assume these protein-protein interactions are conserved across all species under consideration. This dataset will be referred to as the "binding group". 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".

Gene expression data

Gene expression levels were estimated from publicly available RNA-Seq datasets taken from SRA using the pseudo-alignment tool, Salmon [61]. Reads for each species were mapped against their respective protein-coding sequences taken from NCBI Refseq/Genbank [62, 63], ENSEMBL [64], the Joint Genome Institute [65], the Broad Institute (https://portals.broadinstitute.org/regev/orthogroups/), the Aspergillus Genome Database [66], or http://www.saccharomycessensustricto.org/cgi-bin/s3.cgi?data=Annotations&version=current [67]. 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 [51]. 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 [68].

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 $\rho > 0.98$). Each RNA-Seq sample/replicate for each species was transformed to a standard lognormal distribution (i.e. $ln(X) \sim N(0,1)$, where X is the gene expression vector for a species), consistent with the transformation used by [19]. Notably, the log-transformation removes the 0 boundary from the data, which better reflects the assumptions of Brownian Motion [50]. 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. crassa*).

Ortholog identification

Orthologs for fungal species were taken from FungiDB [69], previous publications [67, 70], or the Reciprocal Best Hits BLAST approach, which was only used for *N. castellii*. Proteins with an annotated paralog in the FungiDB or previous literature were excluded from the analysis, as introduction of a paralog could impact the gene expression of the original gene. This eliminated 3669 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 [71]. 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. The Brownian Motion model assumes branch lengths of the phylogenetic tree are proportional to time [50, 72]. To convert the RAxML phylogenetic tree to an ultrametric tree with branch lengths in millions of years, treePL [73] was used to date the tree, taking the divergence time of *S. cerevisiae* and *C. neoformans* (723 millions of years ago (MYA), from TimeTree [74]) as a calibration point. The final phylogenetic tree used for all analyses can be observed in Fig. 1. A summary of the species used, the RNA-Seq data used, and the availability of protein-protein interaction data from STRING can be found in Additional file 1, Table S1.

Analysis of gene expression data

Analyses and visualizations were performed using the R programming language.

Coevolution of gene expression was broadly examined using the Covariance Ratio test implemented in **geomorph** [34–36]. Briefly, this test compares the degree of covariation between traits within predefined modules to covariation between modules. In this case, modules were defined as groups of tightly-linked proteins within a protein-protein interaction network. Modules were determined by applying the Markov Clustering algorithm (as implemented in the clusterMaker2 Cytoscape plug-in [75]) to the protein-protein interaction data using the STRING confidence scores as edge weights. The Covariance Ratio test was applied to all modules with at least 15 proteins. A covariance ratio score of 1 indicates covariance of a trait between modules is equal to the covariance within modules. The closer the covariance ratio is to 0, the more modular the data (i.e. the greater the covariance of a trait within modules is relative to between modules).

Gene expression evolution was modeled as a multivariate Brownian Motion process using the R package **mvMORPH** [37] in order to examine the strength of coevolution between pairs of proteins (as opposed to coevolution within modules). 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 [31, 32, 37]. 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 [59]. Briefly, phylogenetic independent contrasts (PICs) and standardized variances [14] were calculated from gene expression data for each ortholog set using the pic() function from the **ape** R package [76]. Pairs of genes containing a significant correlation (i.e. $p < 0.05$) between PICs and standardized variances, which indicates violation of Brownian Motion assumptions [50, 59], were excluded from further analyses.

Under no coevolution of gene expression, the expected value for the phylogenetically-corrected correlation ρ_C is 0.0. A one-sample t-test was performed to assess if the mean value of ρ_C for the binding and control groups were significantly different from 0.0. Under the hypothesis that gene expression coevolves between proteins which physically-interact, we expect the mean value of ρ_C for the binding group to be significantly different from 0. In contrast, we do not expect the mean value of ρ_C for the control group to be significantly different from 0. A Welch's t-test was also used to assess if the mean values of ρ_C were significantly different from each other. Similar tests were performed for the phylogenetically-uncorrected correlations ρ_U.

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 [38, 45]; 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**.

Furthermore, previous studies have examined the relationship between sequence evolution and gene expression evolution [7, 8]. We compared our estimates of gene expression coevolution to measures of sequence evolution taken from Clark et al. [7]. Clark et. al. also examined gene expression coevolution using the Codon Adaptation Index (CAI), which allowed us to compare our results based on empirical estimates of gene expression with a commonly-used proxy based on codon usage [10].

To determine if functional-relatedness, gene expression, number of protein interactions, and sequence coevolution have an impact on the strength of gene expression 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 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** [77, 78].

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.

Assessing accuracy of methods for detecting coevolution of gene expression

Data simulated under Brownian Motion were used to assess the ability to detect coevolution of gene expression (see Additional file 1 for details). Briefly, protein pairs from the binding set were simulated allowing for coevolution (i.e. the covariance term for the simulations was allowed to be non-zero), forming the simulated binding set. On the other hand, protein pairs from the control set were simulated forcing independent evolution of gene expression (i.e. the covariance term between them was set to 0 in the simulations), forming the simulated control set. The number of true positives (significant result from simulated binding set), true negatives (non-significant result from simulated control set), false positives (significant result from simulated control set), and false negatives (non-significant result from simulated binding set) were determined using the statistical tests described below. From these, a true positive rate (TPR, proportion of significant results from the simulated binding set) and a false positive rate (FPR, proportion of significant results from the simulated control set) were calculated to assess statistical power and specificity of each method. Similarly, a false discovery rate (FDR, proportion of false positives out of all significant results from both the simulated binding and simulated control sets) to determine potential trade-offs between statistical power and specificity for each method. Finally, an overall accuracy score (proportion of true positives and true negatives out of all simulated protein pairs) was calculated for each method.

For the PCM approach, protein pairs were considered coevolving if a Likelihood Ratio test (as implemented in **mvMORPH**) comparing the model allowing coevolution of gene expression to a null model forcing independent evolution of gene expression had a Benjamini-Hochberg corrected p-value <0.05. Similarly, for the non-PCM approach (cor.test() function in R), protein pairs were considered significantly coevolving if the phylogenetically-uncorrected correlation ρ_U had a Benjamini-Hochberg corrected p-value <0.05

Previous work proposed using randomly-generated null distributions (i.e. the control group) as a means of determining statistically significant gene expression coevolution using phylogenetically-uncorrected correlations. This approach is thought to be an adequate approach to

control for the phylogeny when the phylogeny is unknown [12]. We implement approaches similar to those described in Fraser et. al. [8] and Martin and Fraser [12] using both the phylogenetically-uncorrected and phylogenetically-corrected correlations.

Fraser et. al. [8] compared the relative histograms of correlations from a binding and a control group to determine the bin at which the relative frequencies of the binding group were greater than the control group for all subsequent bins. Pairs of proteins were considered significantly coevolving if they had a correlation greater than this point. To assess the accuracy of this method, we split both the binding and control groups into training and test sets (80% and 20% of the data, respectively). The binding and control training sets were used to determine the significance cutoff, while the test sets were then used to assess the accuracy of this approach.

Martin and Fraser [12] presented an approach to determine if gene sets (i.e. more than 2 genes) showed significant coevolution of gene expression by comparing the median phylogenetically-uncorrected correlation to the median correlations from 10,000 randomly-generated gene sets. As we only deal with protein pairs, we compared the number of times (out of 1000) a randomly-generated protein pair had a correlation greater than the correlation of the target protein pair. This procedure was repeated for each protein pair in the binding and control groups. A p-value for each pair was calculated as described in Martin and Fraser [12], and a p-value cutoff was empirically-determined such that the false discovery rate was approximately 5%.

We note accuracy scores can be skewed by large differences in the size of the binding and control groups. For example, if a method is underpowered and the size of the control group is much larger than the binding group, then failure to detect significant differences in the binding group is heavily outweighed by successfully not detecting significant differences in the control group. This results in a higher, and potentially misleading, accuracy score for the method. To account for this, each method was assessed using a subsample of the control group which is the same size as the binding group. Model assessments were made 100 times to obtain mean TPR, FPR, FDR, and overall accuracy scores.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12864-020-6761-3.

Additional file 1: Supplemental Material and Methods. Files contains supplemental materials and methods, including supplementarytables and figures referenced in this manuscript. **Table S1.** Species and data sources used in this study. Basic information on species used in analysis, including the citation corresponding to the RNA-Seq data used and whether or not STRING data was available at the time of analysis. **Table S2.** Comparing performance of different methods for assessing coevolution of gene expression after removing protein pairs with a STRING Score less than 400. **Figure S1.** Heatmap showing overall similarity between species gene expression estimates. **Figure S2.** Strength of coevolution as function of overlapping Gene Ontology terms between protein pairs. **Figure S3.** Effects of controlling for number of interactions of proteins on phylogenetically-corrected and phylogenetically-uncorrected correlations. **Figure S4.** Distance between CAI and empirical-based measures of coevolution as function of gene expression. **Figure S5.** Results from simulated data. **Figure S6.** Phylogenetically-corrected correlation distributions without filtering genes violating BM. **Figure S7.** Effects of functional-relatedness on phylogenetically corrected correlation without filtering genes violating BM. **Figure S8.** Effects of number of interactions and gene expression on strength of coevolution without filtering genes violating BM. **Figure S9.** Comparison to other coevolution metrics based on protein sequence and CAI without filtering genes violating BM. **Figure S10.** Assessing the effects of potential discordance between the species tree and gene trees (e.g. introgression, ILS, etc.). **Figure S11.** Assessing the effects of using branch lengths based on mean nucleotides substitutions per site (as output by RAxML) instead of a time-calibrated tree.

Abbreviations

PCM: Phylogenetic comparative methods; BM: Brownian motion; OU: Ornstein-Uhlenbeck; TPM: Transcripts per million; TPR: True positive rate; FPR: False positive rate; FDR: False discovery rate; CAI: Codon adaptation index; ILS: Incomplete lineage sorting; MYA: Millions of years ago

Acknowledgements

We would like to thank the two anonymous reviewers whose comments and suggestions greatly improved the quality of this article.

Authors' contributions

ALC performed all analyses described in this work. BCO and MAG assisted in interpretation of data and writing of the manuscript. All the authors read and approved the final manuscript.

Funding

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 to A.L. Cope 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). The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

Scripts, protein-coding sequence files, and processed data for running the analysis described in this paper can be found on GitHub (https://github.com/acope3/GeneExpression_coevolution). Protein-coding sequences from *S. kudriavzevii*, *S. mikatae*, *S. paradoxus*, and *S. bayanus* were pulled from http://www.saccharomycessensustricto.org/cgi-bin/s3.cgi?data=Annotations&version=current. Protein-coding sequences for *L. kluyverii* were pulled from https://portals.broadinstitute.org/regev/orthogroups/nt/Sklu.fasta. Protein-coding sequences for *A. nidulans* were pulled from http://www.aspergillusgenome.org/download/sequence/A_nidulans_FGSC_A4/archive/A_nidulans_FGSC_A4_version_s10-m04-r11_orf_coding.fasta.gz Protein-coding sequences for were pulled from ENSEMBL for *F. graminearum* (ftp://ftp.ensemblgenomes.org/pub/fungi/release-46/fasta/fusarium_graminearum/cds/Fusarium_graminearum.RR1.cds.all.fa.gz) and *A. fumigatus* (ftp://ftp.ensemblgenomes.org/pub/fungi/release-46/fasta/aspergillus_fumiga

tus/cds/Aspergillus_fumigatus.ASM265v1.cds.all.fa.gz). Protein-coding sequences for *N. discreta* were pulled from the Joint Genome Institute https://genome.jgi.doe.gov/portal/Neudi1/download/Ndiscreta.FilteredModels2.CDS.fasta.gz. The *N. discreta* sequence data were produced by the US Department of Energy Joint Genome Institute http://www.jgi.doe.gov/ in collaboration with the user community. All other protein-coding sequences were downloaded from the NCBI RefSeq/GenBank database with the following genome assembly accessions: GCF_000146045.2 (*S. cerevisiae*), GCF_000149245.1 (*C. neoformans*), GCF_000237345.1 (*N. castellii*), GCF_000002545.3 (*C. glabrata*), GCF_000182965.3 (*C. albicans*), GCA_000182765.2 (*C. parapsilosis*), GCF_000002495.2 (*M. oryzae*), GCF_000182925.2 (*N. crassa*), GCF_000213175.1 (*N. tetrasperma*). Protein interaction data was downloaded from the STRING database version 11.0 (https://string-db.org/) with the following species identification numbers: 5476 (*C. albicans*), 5480 (*C. parapsilosis*), 318829 (*M. oryzae*), 5141 (*N. crassa*), 4932 (*S. cerevisiae*), 162425 (*A. nidulans*), 5478 (*C. glabrata*), 5518 (*F. graminearum*), 27288 (*N. castellii*). Protein-protein interaction data from the STRING database used in this study can be found in the file <species identification number>.protein.actions.v11.0.txt.gz under the Downloads page of the STRING Database website (https://string-db.org/) after filtering by species name or species identification number. All raw RNA sequencing reads can be found in the NCBI Sequence Read Archive (SRA) or the EMBL Nucleotide Sequence Database (ENA) with the following accession numbers: SRA PRJNA319029 (*S. cerevisiae*, *C. neoformans*), SRA PRJNA320926 (*S. paradoxus*, *S. mikatae*), SRA PRJNA278671 (*S. bayanus*, *S. kudriavzevii*, *N. castellii*), ENA PRJEB10946 (*L. kluyverii*), SRA PRJNA261678 (*C. glabrata*), SRA PRJNA485524 (*C. albicans*), SRA PRJNA429457 (*C. parapsilosis*), SRA PRJNA326901 (*F. graminearum*), SRA PRJNA223667 (*M. oryzae*), SRA PRJNA381768 (*A. fumigatus*), SRA PRJNA481968 (*A. nidulans*), SRA PRJNA177178 (*N. crassa*), SRA PRJNA257829 (*N. discreta*), SRA PRJNA257828 (*N. tetrasperma*). Other references for raw sequencing data (including publication citations) can be found in Additional file 1, Table S1.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Genome Science and Technology, University of Tennessee, Knoxville, Tennessee, USA. [2]Chemical Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA. [3]Department of Ecology and Evolutionary Biology, University of Tennessee, Knoxville, Tennessee, USA. [4]National Institute of Mathematical and Biological Synthesis, University of Tennessee, Knoxville, Tennessee, USA.

Received: 9 January 2020 Accepted: 29 April 2020
Published online: 20 May 2020

References

1. Dunn CW, Luo X, Wu Z. Phylogenetic Analysis of Gene Expression. Integr Comp Biol. 2013;53(5):847–56. Available from: https://academic.oup.com/icb/article-lookup/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):. https://doi.org/10.1073/pnas.95.25.14863.
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):3513–9. Available from: https://academic.oup.com/nar/article-lookup/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. Available from: http://dx.plos.org/10.1371/journal.pone.0017258.
5. Michalak P. Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. Genomics. 2008;91(3):243–8. Available from: https://www.sciencedirect.com/science/article/pii/S0888754307002807#bib30.
6. Ge H, Liu Z, Church GM, Vidal M. Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. Nat Genet. 2001;29(4):482–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11694880http://www.nature.com/articles/ng776z.
7. Clark NL, Alani E, Aquadro CF. Evolutionary rate covariation reveals shared functionality and coexpression of genes. Genome Res. 2012;22(4):714–20. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22287101. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3317153.
8. 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. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15175431. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC439012.
9. Lithwick G, Margalit H. Relative predicted protein levels of functionally associated proteins are conserved across organisms. Nucleic Acids Res. 2005;33(3):1051–7. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gki261.
10. 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):1281–95.
11. Charlesworth B. Effective population size and patterns of molecular evolution and variation. Nat Rev Genet. 2009;10:195–205.
12. Martin T, Fraser HB. Comparative expression profiling reveals widespread coordinated evolution of gene expression across eukaryotes. Nat Commun. 2018;9(1):4963. Available from: http://www.nature.com/articles/s41467-018-07436-y.
13. Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, Amaral-Zettler LA, et al. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. PLoS Biol. 2014;12(6):e1001889. Available from: https://dx.plos.org/10.1371/journal.pbio.1001889.
14. Felsenstein J. Phylogenies and the Comparative Method on JSTOR. Am Nat. 1985;125(1):1–15. Available from:https://www.jstor.org/stable/2461605?seq=1#metadata_info_tab%_contents.
15. Rohlf FJ. A Comment on Phylogenetic Correction. Soc Study Evol. 2006. Available from: https://www.jstor.org/stable/4095344.
16. 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. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29301966. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5776959.
17. 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):3482–9. Available from:https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bti564.
18. Barker D, Pagel M. Predicting Functional Gene Links from Phylogenetic-Statistical Analyses of Whole Genomes. PLoS Comput Biol. 2005;1(1):e3. Available from: http://dx.plos.org/10.1371/journal.pcbi.0010003.
19. Bedford T, Hartl DL. Optimization of gene expression by natural selection. Proc Natl Acad Sci U S A. 2009;106(4):1133–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19139403. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2633540.
20. 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):343–8. Available from: http://www.nature.com/articles/nature10532.
21. Eng KH, Bravo HC, Keleş S. A Phylogenetic Mixture Model for the Evolution of Gene Expression. Mol Biol Evol. 2009;26(10):2363–2372. Available from: https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msp149.
22. 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):1746–53. Available from: https://academic.oup.com/gbe/article-lookup/doi/10.1093/gbe/evt121.
23. 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. Available from: https://academic.oup.com/gbe/article/10/2/538/4823540.
24. 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. Available from: https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msh257.
25. 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):201–11. Available from: https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/mst190.
26. Rohlfs RV, Nielsen R. Phylogenetic ANOVA: The Expression Variance and Evolution Model for Quantitative Trait Evolution. Syst Biol. 2015;64(5):695–708. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26169525. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4635652.
27. 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. Available from: https://dx.plos.org/10.1371/journal.pcbi.1003255.
28. Butler MA, King AA. Phylogenetic Comparative Analysis: A Modeling Approach for Adaptive Evolution. Am Nat. 2004;164(6):683–95. Available from: http://www.journals.uchicago.edu/doi/10.1086/426002. http://www.ncbi.nlm.nih.gov/pubmed/29641928.
29. Hansen TF. Stabilizing Selection and the Comparative Analysis of Adaptation. Evolution (N Y). 1997;51(5):1341–51. Available from: http://doi.wiley.com/10.1111/j.1558-5646.1997.tb01457.x.
30. Bartoszek K, Pienaar J, Mostad P, Andersson S, Hansen TF. A phylogenetic comparative method for studying multivariate adaptation. J Theor Biol. 2012;314:204–15. Available from: https://www.sciencedirect.com/science/article/pii/S0022519312003918.
31. Revell LJ, Collar DC. Phylogenetic Analysis of the Evolutionary Correlation Using Likelihood. Evolution (N Y). 2009;63(4):1090–1100. Available from: http://doi.wiley.com/10.1111/j.1558-5646.2009.00616.x.
32. 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.
33. Zhang J, Yang JR. Determinants of the rate of protein sequence evolution. Nat Rev Genet. 2015;16(7):409–20. Available from: http://www.nature.com/articles/nrg3950.
34. Adams DC. Evaluating modularity in morphometric data: Challenges with the RV coefficient and a new test measure. Methods Ecol Evol. 2016;7(5):565–72. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.12511.
35. Adams DC, Collyer ML. Comparing the strength of modular signal, and evaluating alternative modular hypotheses, using covariance ratio effect sizes with morphometric data. Evolution (N Y). 2019;73(12):2352–67. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/evo.13867.
36. Adams DC, Collyer ML. Geomorph: Geometric Morphometric Analyses of 2D/3D Landmark Data [R package geomorph version 3.2.1]. Compr R Arch Netw (CRAN). 2020. Available from: https://cran.r-project.org/package=geomorph.
37. 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):1311–9. Available from: http://doi.wiley.com/10.1111/2041-210X.12420.
38. Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW. Evolutionary rate in the protein interaction network. Science. 2002;296(5568):750–2. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11976460.
39. Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. Why highly expressed proteins evolve slowly. Proc Natl Acad Sci. 2005;102(40):14338–43.
40. Drummond DA, Raval A, Wilke CO. A Single Determinant Dominates the Rate of Yeast Protein Evolution. Mol Biol Evol. 2005;23(2):327–37.
41. Feyertag F, Berninsone PM, Alvarez-Ponce D. Secreted Proteins Defy the Expression Level-Evolutionary Rate Anticorrelation. Mol Biol Evol. 2017;34(3):692–706.
42. 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;7:1559–79.
43. Gilchrist MA. Combining Models of Protein Translation and Population Genetics to Predict Protein Production Rates from Codon Usage Patterns. Mol Biol Evol. 2007;24(11):2362–72.
44. dos Reis M, Wernisch L, Savva R. Unexpected correlations between gene expression and codon usage bias from microarray data for the whole \textit{Escherichia coli} K-12 genome. Nucleic Acids Res. 2003;31(23):6976–85.
45. Drummond DA, Wilke CO. Mistranslation-Induced Protein Misfolding as a Dominant Constraint on Coding-Sequence Evolution. Cell. 2008;134:341–52.
46. Shah P, Gilchrist M. Explaining complex codon usage patterns with selection for translational efficiency, mutation bias, and genetic drift. PNAS. 2011;108(25):10231–6.
47. Wallace EWJ, Airoldi EM, Drummond DA. Estimating Selection on Synonymus Codon Usage from Noisy Experimental Data. Mol Biol Evol. 2013;30(6):1438–53.
48. Adams DC, Collyer ML. Multivariate Phylogenetic Comparative Methods: Evaluations, Comparisons, and Recommendations. Syst Biol. 2018;67(1):14–31. Available from: http://academic.oup.com/sysbio/article/67/1/14/3867043.
49. Adams DC, Felice RN. Assessing Trait Covariation and Morphological Integration on Phylogenies Using Evolutionary Covariance Matrices. PLoS ONE. 2014;9(4):e94335. Available from: https://dx.plos.org/10.1371/journal.pone.0094335.
50. Garland T, Harvey PH, Ives AR. Procedures for the Analysis of Comparative Data Using Phylogenetically Independent Contrasts. Syst Biol. 1992;41(1):18–32. Available from: https://academic.oup.com/sysbio/article/41/1/18/1617342.
51. 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):281–5. Available from: http://link.springer.com/10.1007/s12064-012-0162-3.
52. 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. Available from: https://onlinelibrary-wiley-com.proxy.lib.utk.edu/doi/pdf/10.1002/jez.b.22636.
53. 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):4209–12. Available from: http://doi.wiley.com/10.1002/pmic.201000327.
54. Maddison WP. Gene Trees in Species Trees. Syst Biol. 1997;46(3):523–36. Available from: https://academic.oup.com/sysbio/article-abstract/46/3/523/1651369.
55. Peter J, De Chiara M, Friedrich A, Yue JX, Pflieger D, Bergström A, et al. Genome evolution across 1,011 Saccharomyces cerevisiae isolates. Nature. 2018;556(7701):339–44.
56. Mendes FK, Fuentes-González JA, Schraiber JG, Hahn MW. A multispecies coalescent model for quantitative traits. Elife. 2018;7:. https://doi.org/10.7554/elife.36482.
57. Pagel M. Inferring the historical patterns of biological evolution: Nature Publishing Group; 1999. https://doi.org/10.1038/44766.
58. Piel WH, Donoghue M, Sanderson M. TreeBASE: A database of phylogenetic information. In: e-Biosphere; 2009. Available from: http://phylogeny.harvard.edu/treebase. Accessed 9 Dec 2019.
59. Revell LJ. Phylogenetic signal and linear regression on species data. Methods Ecol Evol. 2010;1(4):319–29. Available from: http://doi.wiley.com/10.1111/j.2041-210X.2010.00044.x.
60. 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):D607–13. Available from: https://academic.oup.com/nar/article/47/D1/D607/5198476.
61. 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):417–9. Available from: http://www.nature.com/articles/nmeth.4197.
62. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44(D1):D733–45.
63. Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res. 2016;44(Database):67–72. Available from: www.ncbi.

64. Cunningham F, Achuthan P, Akanni W, Allen J, Amode MR, Armean IM, et al. Ensembl 2019. Nucleic Acids Res. 2019;47(Database):D745–51. Available from: http://test-metadata.ensembl.org/.
65. Nordberg H, Cantor M, Dusheyko S, Hua S, Poliakov A, Shabalov I, et al. The genome portal of the Department of Energy Joint Genome Institute 2014 updates. Nucleic Acids Res. 2014;42(Database):D26–D31. Available from: http://genome.jgi.doe.
66. Cerqueira GC, Arnaud MB, Inglis DO, Skrzypek MS, Binkley G, Simison M, et al. The Aspergillus Genome Database: multispecies curation and incorporation of RNA-Seq data to improve structural gene annotations. Nucleic Acids Res. 2014;42(Database):D705–D710. Available from: http://www.aspergillusgenome.org/.
67. 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. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22384314. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3276118.
68. 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):2125–39. Available from: https://academic.oup.com/mbe/article/34/9/2125/3858070.
69. Basenko E, Pulman J, Shanmugasundram A, Harb O, Crouch K, Starns D, et al. FungiDB: An Integrated Bioinformatic Resource for Fungi and Oomycetes. J Fungi. 2018;4(1):39. Available from: http://www.mdpi.com/2309-608X/4/1/39.
70. 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):1694–705.
71. 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. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkq291.
72. O'Meara BC, Ané C, Sanderson MJ, Wainwright PC. Testing for Different Rates of Continuous Trait Evolution Using Likelihood. Evolution (N Y). 2006;60(5):922–33. Available from: http://doi.wiley.com/10.1111/j.0014-3820.2006.tb01171.x.
73. Smith SA, O'Meara BC. treePL: divergence time estimation using penalized likelihood for large phylogenies. Bioinformatics. 2012;28(20):2689–90. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts492.
74. Kumar S, Stecher G, Suleski M, Hedges SB. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Mol Biol Evol. 2017;34(7):1812–9. Available from: https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msx116.
75. Morris JH, Apeltsin L, Newman AM, Baumbach J, Wittkop T, Su G, et al. ClusterMaker: A multi-algorithm clustering plugin for Cytoscape. BMC Bioinformatics. 2011;12(1):436. Available from: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-436.
76. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 2019;35(3):526–8. Available from: https://academic.oup.com/bioinformatics/article/35/3/526/5055127.
77. Canty A, Ripley BD. boot: Bootstrap R (S-Plus) Functions. R package version 1.3–20. 2017.
78. Davison AC, Hinkley DV. Bootstrap Methods and Their Applications. 1997. Available from: http://statwww.epfl.ch/davison/BMA/.

Publisher's Note