I am adding this comment because I believe that it may have been better for me to wait a little longer to prepare / submit this paper in order to **1)** gain more understanding of processed data and different analysis methods before paper submission (which could be useful in many situations, if it emphasizes the need for labs to gradually expand their skillsets and focus on a limited number of projects that they can study in-depth) and **2)** prepare public data / code for the paper (since readers may have questions substantially after the paper is published).

In this paper, I had the idea to test predicting fRNAs in yeast (starting with EvoFold predictions), but I had the least bioinformatics experience among the 3 authors. So, instead of a Bioinformatics Specialist supporting labs from a shared resource, I think this may be helpful in terms of an experience that may share some similarities to those that I am currently supporting (although I still genuinely believe there are interesting results within the paper that are worth consideration among the broader community).

To be clear, I do not want this to reflect poorly on the other authors. Much to the contrary, I have continued to stay in contact with Soojin after I graduated, and I have contacted the other authors while preparing this comment. ***However, this comment represents my own personal opinions / observations (so, it should not be assumed that there is complete agreement between the authors about the conclusions in this comment).***

<u>High-Level Comments</u>:

1) **In the process of making additional data / code publicly available, I discovered that I have backed up some files but they aren't quite enough to reproduce all the results.** Namely, I don't have the EvoFold output, or the RNAfold *mfe* / DicodonShuffle EFP values for the EvoFold folds (and I don't think the EvoFold program is available anymore).
    a. **I think this can be a useful example of the need to plan ahead for having well-organized, publicly available code and data in future projects, since there can be questions about the paper considerably after the paper has been published.**
    b. In terms of what I currently can do for this paper, I have added additional analysis with data / code for 169 genes (without ribosomal genes) as well as the set of 195 genes (<u>with</u> ribosomal genes) in this GitHub repository
    c. Beyond the Coding fRNAs, I created an **all_fRNA_track_sacCer1.bed** file (among the GitHub files) with all the EvoFold prediction locations (to include non-coding folds, such as intronic folds).
        i. I also checked the location of the HAC1 fold (fold chr6:75804-75850 in YFL031W). This fold overlaps the donor site on the 5` splice junction; so, the fold overlaps the intron sequence, although the prediction is not for the full intron (but it also isn't in an unrelated coding position).
2) It is a bit of a side note in terms of data sharing, but the 169 gene set includes removal of ribosomal genes (from a starting set of 195 genes). The ribosomal genes were removed to match previous precedents for gene filtering for ***evolutionary divergence*** (such as the earlier Drummond et al. 2005 PNAS article); for example, we had a

response to reviewers saying "we removed the ribosomal genes because they are a known source of bias in functional genomic studies and our initial dataset was significantly enriched with ribosomal genes." However, there is an interesting cluster that can be seen for fRNA coverage versus ***CAI***, which wasn't emphasized in the original paper and is shown in **Figure C1** in the new additional analysis (on GitHub). I think this is worth noting in the additional analysis since it is a clear pattern that relates to the GO enrichment (and indicates value in the overall topic for the paper).

   a. Also, even with the 169 pre-calculated non-ribosomal genes, the association between fRNA coverage and adjusted synonymous site divergence is interesting, but the abundance/expression and CAI were more highly correlated with divergence. I find visualizations are helpful (and these results were only presented as tables in the paper), so I have added some scatter plots that I believe should make these trends more clear (**Figure C4**, GitHub re-analysis).
   b. Also, while different than the Principal Component Regression analysis in the paper (and the separate correlation with dS`), the R-base scaled PCA analysis shows a similar trend on PC1 as well as a relatively greater similarity between fRNA Coverage and dN/dNdS/dNdS` on both evolutionary scales on PC2 (**Figure C9**, with or without ribosomal gene removal).

3) The methods say *"An R code for our method of partial correlation analysis that controls for the influence of multiple variables (which was used to produce the data in Table 1) is available at Yi lab website (www.yilab.gatech.edu)"*

   a. That code is indeed on the Yi Lab website, but the multi-variate partial correlation **pcor()** function was for the Kim and Yi 2007 paper (which was on the website before this fRNA paper was published). So, please don't look for the code under the Warden et al. 2008 paper.
   b. Part of the reason I feel confident that I did in fact use multi-variate partial correlation analysis is that the paper also says "We modified R scripts available from the supplemental material for Drummond et al. 2006 for partial correlation (factoring out only expression)." In other words, the use of the word "only" refers to the more commonly used partial correlation analysis (adjusting for one-variable, only gene expression), and this indicates that additional variables (expression / CAI / dispensability) were presented in this paper (confirming the 1st sentence referring to the Yi Lab website is accurate, although partial correlation analysis is presented in both Table 1 and 3).
   c. However, the new analysis on GitHub uses the original **partial.cor.test()** function from the Drummond et al. 2006 paper. The reason for this is explained in the 4th section below.

4) In terms of the original code, gene expression (as well as CAI and dispensability) is more significantly associated with nucleotide divergence, but I believe factoring out gene length is important because it is more significantly correlated with fRNA coverage (see **Figure C2** in the additional GitHub figures). So, the important difference between these strategies is that the main tables with the multi-variate analysis focus on factors most

relevant for nucleotide divergence, but gene length is most closely correlated with fRNA coverage.

a. There is an adjustment for gene length in *Table S2* in the original paper (in addition to a non-significant difference between short and long genes, using either a 15 nt or 20 nt cutoff for binary length status, in Table S7B), although the relative length-adjusted trend for 4-species divergence is different for dN versus dS`.
b. In the re-analysis (which includes primary variables other fRNA coverage), the overall conclusions with adjusting for only gene length are similar (see "*fRNAcov_versus_dSadj_adj_length*" in **Table C1**; partial correlation with 4-species dS` adjusting for gene length is -0.203, with glm p-value of 0.0495). However, to be fair, the FDR for the 2-variable comparison is 0.078 instead of 0.00021 for the 1-variable comparison (still significant if using FDR < 0.10 or FDR < 0.25, but noticeably lower than the 1-variable association).
    i. If gene length was included in the multi-variate partial correlation analysis, this would have almost certainly caused the association with fRNA coverage to not be significant (since the 1-variable unadjusted p-value is already almost not significant with p-value < 0.05). However, *Tables 1-4* only use 3 other variables (Gene Expression, CAI, and Dispensability), indicating some appreciation for the need to focus on the variables most relevant to nucleotide divergence. In other words, over-fitting could be more of a problem with multi-variate analysis with 4-18 variables for comparisons that use as few as 20 genes (in the paper), and this is one reason why gene length was not included in the multi-variate analysis.
c. With the extra fRNA Coverage versus CAI plots, it may be worth noting that the "*fRNAcov_versus_dSadj_adj_length"* partial correlation without ribosomal genes in **Table C1** is more significant than the partial correlation with ribosomal genes in **Table C3.**
d. While the very important comparison of gene length *instead* of fRNA coverage was provided (in *Table S2* of the original paper; as recommended by a helpful reviewer comment) and it was very helpful to have R code for the earlier Drummond et al. 2006 paper, **I think it is important to understand why comparisons were performed a certain way when using a template for code (otherwise, it will be more difficult to determine when they should/need to be modified).** In other words, I think I was more likely to focus on gene expression than gene length because that is what was done in previous publications, but gene length was something important for this particular study.

5) With the set of 169 genes, I can reproduce that fRNA coverage is more significant than gene length for the 1-variable analysis (and most significant for the adjusted dS values; as shown in Table S2). **However, the original paper contains some discrepancies for dS` (between 4-species divergence in Tables 1/3/S2 and 2-species divergence in Tables S4; or, between Tables 2/4 and Tables S5-S6), and this discrepancy still holds true with the full set of genes (or the set of genes with 4-species divergence values).**

a. For analysis as a Bioinformatics Specialist (in a different context, but particularly for differential expression testing, even for 1-variable analysis), I have found that you should have at least some benchmarks for each project (and a way to evaluate your results, such as an independently calculated expression value). **So, the presentation of analysis needs to be a fair representation of results (to avoid cherry picking results that are not representative of the overall findings), but also with some acknowledgement that the methods for analysis aren't perfect (which can be due to either biological or technical reasons).**
b. If possible, taking time to think of creative experiments to try and understand discrepancies is the sort of critical thinking that should help improve the quality and impact of a paper. However, realistically, there are going to be situations where all possible explanations for all observations are not possible. So, I think the strategy of presenting both results (such as 4-species versus 2-species dS` divergence in this paper) is a solution that should be encouraged (instead of excluding discordant results that are not clearly understood), although we shouldn't have used the word "similar" to describe the smaller (2-species) divergence analysis (since significance of the dS` trends noticeably varies).

6) There were several missing dN values in the 169 gene table because they were pre-calculated in the Drummond et al. 2006 supplement file (*Wall-dn-ds.txt*), although I did re-calculate divergence for more genes among the smaller time-scale (for 2-species instead of 4-species).
    a. In general, I would recommend expecting extra time to investigate observations like this, prior to publication. The Project Lead / 1st Author needs to be comfortable describing the entire project, and he/she should at least be capable of performing analysis similar to creation of the processed datasets (**otherwise, processing of similar data should likely necessary as a training exercise to prove understanding of everything described in the paper**).
    b. To be clear, the need to spend more time and understand the subject/project more deeply was not immediately obvious to me; however, I know that there are some things in the bioinformatics field that I didn't really notice could benefit from improvement until after gaining 5-10 years of experience.

<u>Specific Corrections</u>:

- The table legend for Tables 1 and 4 say "Pearson Correlations are shown in parenthesis below partial correlation in the above table." This does match Supplemental Tables S3 and S4. However, in Tables 1 and 4, the Pearson Correlations are shown in parenthesis, but to the right of the partial correlation values (not below them).
- *Incorrect Citations within Methods*: We used data from Wall et al. [8] (available from the supplementary material for Drummond et al. [7]) → Drummond et al. is citation #9, Wall et al is citation #10
- *There is a minor typo in the discussion*: "RNAz and EvolFold" → "RNAz and EvoFold"

While it can be difficult to parse through an incomplete set of data ~10 years after a paper has been published, I am very grateful for several rounds of discussion with Soojin; her help considerably improved the quality and precision of this comment.

In terms of possible solutions (to have more time to work on a publication), Georgia Tech now has a BS/MS degree in Bioinformatics (which has overlap with the extra math and computer science courses that I took as free electives as an undergrad). It is possible that it may still be better to wait for 1st author publication after earning the dual BS/MS degree, but I would recommend that option for current undergraduates.