Supplemental Information

De novo identification of actively translated open reading frames with ribosome profiling data

Yanan Zhu, Fajin Li, Xuerui Yang, Zhengtao Xiao

Evaluating the dependence between two p-values

RiboCode combines two p-values (i.e., pval_frame0_vs_frame1 and pval_frame0_vs_frame2) for assessing whether the number of ribosome-protected fragment (RPF) reads in the open reading frame (ORF) 0 (i.e., in-frame RPF reads, represented by F0) are consistently higher than those of the RPF reads in frame 1 and frame 2 (represented by F1 and F2, respectively). If both p-values are significant, the combined p-value would be smaller than the individual p-value, as expected. However, considering the usually high noise level of the ribosome profiling data, it would be arbitrary to accept/reject the null hypothesis at a significant level α if only one of the p-values is smaller than α. Thus, in such cases, combining the strategies could provide a more informative p-value by summarizing information from two tests. That is why we prefer combining the p-values.

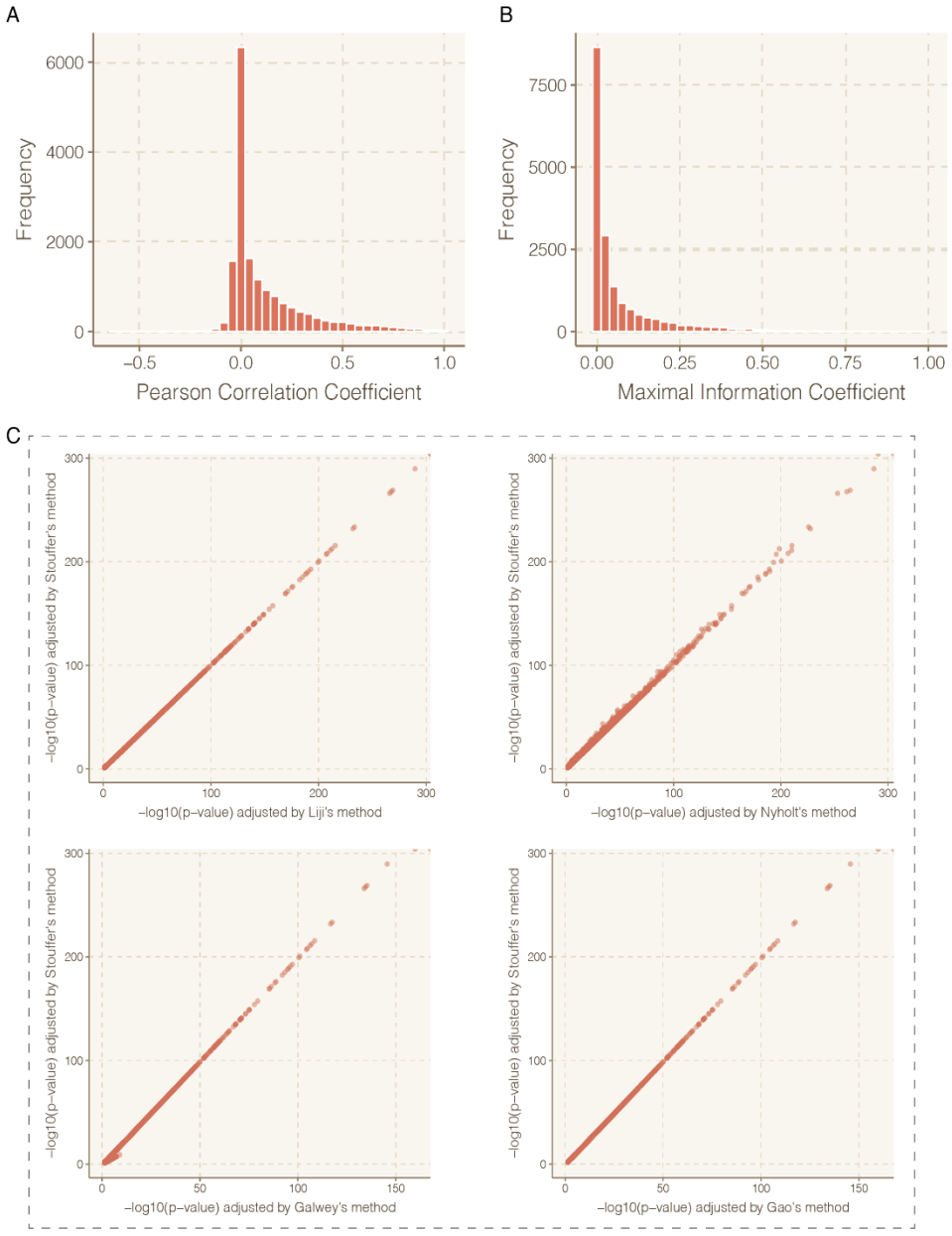
Stouffer's method is a commonly used method for integrating multiple statistical tests, which has been proved to perform better than other popular methods[1], e.g., Fisher's and Tippett's methods. This method is robust in many applications when the number of combining tests is smaller than 5. Therefore, it is suitable for this context as only two tests are combined for each ORF in this system.

However, Stouffer's method assumes that the tests to be combined are independent, which urges us to check the dependence across two tests in these cases. As the two tests to be combined share the common F0, the testing of the dependence between F1 and F2 can determine the relationship between these two tests. We added an argument (--dependence_test) to the updated RiboCode for assessing the significance of dependence between F1 and F2. By setting the "--dependence_test" to "pcc" or "mic", RiboCode will calculate the Pearson correlation coefficient (PCC) and maximal information coefficient (MIC) between F1 and F2.

We found that for the overwhelming majority of ORFs, the PCCs between F1 and F2 are close to zero, suggesting that there is no linear relationship between the two tests (**Supplemental Figure 1A**). However, PCC alone cannot guarantee that there is no nonlinear relationship between F1 and F2. The MIC is a metric for measuring the correlation between paired variables regardless of linear or nonlinear relationship[2]. Thus, we calculated the MIC between F1 and F2 to further evaluate their nonlinear correlations. The result showed that MIC values of most ORFs are close to zero (**Supplemental Figure 1A**). These analyses together suggested that the relationship between F1 and F2 is very weak or likely unimportant.

To further demonstrate the rationality of Stouffer's method in this system, we also calculated the p-values using other methods designed for handling the dependence among the tests (see more

details at https://search.r-project.org/CRAN/refmans/poolr/html/stouffer.html). The results showed that the adjusted p-values generated by these methods are highly consistent with the original p-values produced by Stouffer's method, suggesting that Stouffer's method is valid for this system.

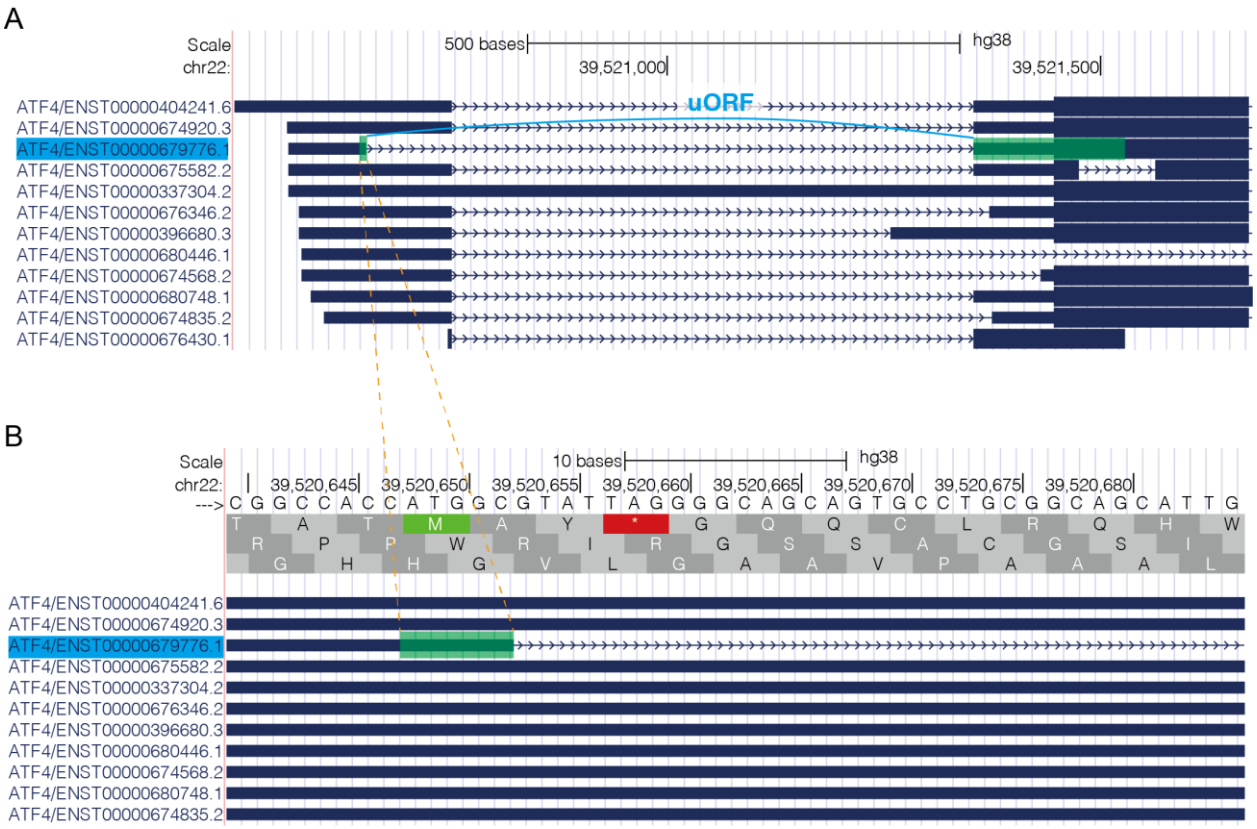


Supplemental Figure 1: Pearson correlation coefficients and Maximal information coefficients. Distribution of (**A**) Pearson correlation coefficients and (**B**) Maximal information coefficients between frame 1 and frame 2. (**C**) Comparisons between the adjusted p-values calculated by various adjustment methods and the original p-values generated by Stouffer's method.

Briefly, we explained why the two p-values should be combined. We also proved that: (1) for the overwhelming majority of ORFs, the two tests conducted by RiboCode are not (or at least very weakly) dependent; (2) Stouffer's method is suitable for p-value combination in this context.

Explanation of RiboCode results using the upstream ORF (uORF) of ATF4 as an example:
Considering that alternative splicing of precursor mRNA creates multiple distinct transcripts, RiboCode searches for candidate ORFs from each transcript. Introns are not included in the candidate ORFs. Each candidate ORF has only one in-frame start codon and one in-frame stop codon. To help users visualize the positions of the predicted ORFs on the genome, RiboCode reports the coordinates of each ORF on the genome (i.e., ORF_gstart and ORF_gstop) and on the mRNA transcript (i.e., ORF_tstart to ORF_tstop).

One of the predicted ORFs "ENSG00000128272_39520648_39521528_59", has 59 codons and is located in transcript "ENST00000679776" of ATF4. On the genome, this ORF is interrupted by an intron and its in-frame stop codon is located in the next exon (**Supplemental Figure 2A**). Be aware that a stop codon at the right of this ORF's start codon is located in the intron region of the transcript ENST00000679776 and therefore is **NOT** included in this ORF (**Supplemental Figure 2B**).



Supplemental Figure 2: Predicted ORF "ENSG00000128272_39520648_39521528_59". The screenshots from the UCSC website show the constitute (**A**) and the start codon (**B**) of the predicted ORF " ENSG00000128272_39520648_39521528_59" on the human genome.

This predicted uORF has been also identified by other studies[3,4]. The amino acid sequence of this ORF is also collected by UniProt (https://www.uniprot.org/uniprot/A0A6Q8PF56). More details about this ORF and other transcript isoforms harboring this ORF are available in the outputs of RiboCode.

References:

1 Kim, S. C. et al. Stouffer's test in a large scale simultaneous hypothesis testing. *PLoS One.* **8** (5), e63290 (2013).

2 Reshef, D. N. et al. Detecting novel associations in large data sets. *Science.* **334** (6062), 1518–1524 (2011).

3 Vattem, K. M., Wek, R. C. Reinitiation involving upstream ORFs regulates ATF4 mRNA translation in mammalian cells. *Proceedings of the Nationall Academy of Sciences of the United States of America.* **101** (31), 11269–11274 (2004).

4 Lewerenz, J. et al. Mutation of ATF4 mediates resistance of neuronal cell lines against oxidative stress by inducing xCT expression. *Cell Death and Differentiation.* **19** (5), 847–858 (2012).