Supplemental Material
RiboDiff: Detecting Changes of mRNA Translation Efficiency from Ribosome Footprints

July 30, 2016

A Library Size Normalization

Due to differences in sequencing depth, the read count of the same gene can vary in different samples (or replicates), even if no biological effect exists. Therefore, in a first step, the raw count data needs to be normalized by a library size factor in a first step. We calculate the normalization constant (a.k.a. size factor) S similar to [1] with modifications:

$$S_T^r = \underset{i:y_T^{i,r}>0}{\text{median}} \left(\frac{y_T^{i,r}+1}{\sqrt[n]{\prod_{j=1}^{n}(y_T^{i,j}+1)}} \right). \tag{1}$$

The size factors of RNA-Seq and ribosome footprinting (RF) libraries are calculated separately. Here, T denotes data type (RNA-Seq or RF); r denotes the r-th sample in data type T that includes replicates of both experimental treatments. $y_T^{i,r}$ is the observed count of type T for gene i in sample r. For all genes in all replicates, we add one to their count value to avoid the degenerate case of setting the geometric mean across all replicates (indexed by j) in the denominator to zero. We calculate the ratios of observed counts of all genes in a given sample to the geometric means and determine the median of these ratios whose count is greater than zero as the size factor.

The raw counts of gene i in sample r are normalized by the corresponding size factor.

B The Explanatory Matrix of GLM

As described in the main text, we decompose the expected count into multiple latent quantities. For RNA-Seq, it is given by $\log(\mu^i_{\text{mRNA},C}) = \beta^i_C + \beta^i_{RNA}$, whereas for RF it is given by $\log(\mu^i_{\text{RF},C}) = \beta^i_C + \beta^i_{RF} + \beta^i_{\Delta,C}$. We use a generalized linear model (GLM) to learn the latent quantities βs from the observed count data, and then calculate the means. To control the observed read counts fitting into the GLM system, an $n \times 5$ explanatory matrix X is designed, where n equals to the total number of replicates in both experimental conditions for RNA-Seq and RF. Here we show it in the context of linear predictor η of GLM (Equation 2), where $\eta = X \times \beta$.

In X matrix, the first two columns represent the baseline mRNA abundance β^i_C in the two conditions. The third and fourth columns (β^i_{RNA} and β^i_{RF}) define whether the counts are from RNA-Seq or RF, respectively.[1] The fifth column ($\beta^i_{\Delta,C}$) relates the RF count to the potential translational effect. Each

[1] In the implementation, in order to keep full rank of X, we do not include the fourth column β^i_{RF}, as it is linearly dependent with the third column.

row of X is used to control how the observed count of a specific sample should be decomposed into latent quantities in order to fit the GLM models. In this example, as indicated by the third column, the first four rows (marked in blue) model RNA-Seq counts with *two replicates* for each condition, $C0$ and $C1$, while the last six rows (marked in green) model RF counts with *three replicates* for each condition. Note the first and second columns in X are shared between mRNA and RF counts, where we couple the two different data sets. The linear predictor η then is linked with negative binomial distributed mean $\mu^i_{\text{RF},C}$ and $\mu^i_{\text{mRNA},C}$ through logarithm as the link function, namely $\log(\mu) = \eta = X \times \beta$. The βs are estimated by maximizing the likelihood of GLM [2].

In our model, the RNA and RF replicates are assumed to be independent of each other. The flexible configuration allows different numbers of replicates to be analyzed. To extend *RiboDiff* to integrate pairing information, we could modify the GLM model where the pairing term β^i_p (p indexes the mRNA and RF pair for each condition) is added to the linear equation. In this way, the pairing signal contributed from a pair of mRNA and RF is absorbed by this term in GLM fitting.

Another possible extension to *RiboDiff* is to handle time series comparison. Assuming we have data at different time points and the goal is to test whether time point n is significantly different from others in treatment in a subset of genes. The GLM model for mRNA and RF abundance can be modified as following: $\log(\mu^i_{\text{mRNA},C}) = \beta^i_C + \beta^i_{RNA} + \beta^i_n + \beta^i_{else}$ and $\log(\mu^i_{\text{RF},C}) = \beta^i_C + \beta^i_{RF} + \beta^i_n + \beta^i_{else} + \beta^i_{\Delta,C\&n}$, where β^i_n and β^i_{else} are the latent quantities of time series effects and $\beta^i_{\Delta,C\&n}$ captures the possible TE change at time point n under treatment condition.

After the β^i is estimated for each gene, the expected RNA-Seq and RF counts, $\mu^i_{\text{mRNA},C}$ and $\mu^i_{\text{RF},C}$ can be obtained. The next step is to estimate the dispersion parameter κ^i by maximizing the negative binomial likelihood function with the observed read counts and the expected counts of both RNA and RF. See details in the next section.

$$\eta = \begin{array}{c} \begin{matrix} C0 & C1 & mRNA & RF & \Delta_{Eff.} \end{matrix} \\ \begin{bmatrix} 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 \end{bmatrix} \end{array} \times \begin{bmatrix} \beta^i_{C=0} \\ \beta^i_{C=1} \\ \beta^i_{mRNA} \\ \beta^i_{RF} \\ \beta^i_{\Delta} \end{bmatrix}. \tag{2}$$

C Negative Binomial Likelihood Function

Because the count data are assumed to be sampled from a negative binomial distribution with parameter mean μ and dispersion κ, we estimate κ given observed counts and the estimated mean by maximizing the NB likelihood function (equation 4).

The probability mass function of the negative binomial distribution is given by

$$Pr(y^{i,j}) = \binom{y^{i,j} + 1/\kappa^{i,j} - 1}{y^{i,j}} \left(\frac{1/\kappa^{i,j}}{1/\kappa^{i,j} + \mu^{i,j}}\right)^{1/\kappa^{i,j}} \left(1 - \frac{1/\kappa^{i,j}}{1/\kappa^{i,j} + \mu^{i,j}}\right)^{y^{i,j}}, \tag{3}$$

where $y^{i,j}$ is the observed RF or mRNA read count of j^{th} replicate of gene i; $\kappa^{i,j}$ is the dispersion parameter of the NB distribution where $y^{i,j}$ is drawn from; $\mu^{i,j}$ is the estimated count of j^{th} replicate. Thus the logarithmic likelihood of negative binomial of gene i is given by

$$\log \ell_{NB} = \sum_{j=1}^{n} \log(Pr(y^{i,j})) - \frac{1}{2}\log(\det(X' \cdot \mathrm{diag}(\frac{\mu^i}{1+\mu^i\kappa^i}) \cdot X)). \tag{4}$$

Note that the likelihood function is adjusted by a Cox-Reid term as suggested by Robinson *et al.* [3] to compensate bias from estimating coefficients in fitting GLM step. Again, X is the explanatory matrix with dimension $n \times 4$ or $n \times 5$, depending on H_0 or H_1, where n is the total number of RNA-Seq and RF replicates; μ^i is the vector of estimated counts; κ^i is the dispersion vector.

In the previous section, we estimate βs by starting with an arbitrary value of dispersion. After we update the dispersion from NB likelihood function, the new dispersion is plugged into the GLM again to start a new optimization cycle. This process ends when the EM-like method converges or an iteration maximum is reached.

D Empirical Bayes Shrinkage for Obtaining Final Dispersion

From previous steps, we obtain the dispersion for each gene. However, it is estimated only based on the read counts of the gene itself, thus it is less reliable due to the limited replicates. Therefore, we need a systematic method to adjust the raw dispersions. This can be accomplished by the following two steps:

1) Obtain the mean-dispersion relationship by regressing all raw dispersions κ^i given mean counts under assumption: $\kappa_F = f(\mu) = \lambda_1/\mu + \lambda_0$ [4]. Namely, for each gene with mean count μ^i, a fitted dispersion κ_F^i can be calculated.

2) To get the final dispersion κ_S^i, we follow the approach published recently [5]. This approach is based on the observation that the dispersion follows a log-normal prior distribution [6] centered at the fitted dispersion κ_F^i. The κ_S^i can be estimated by maximizing the following equation:

$$\kappa_S^i = \arg\max_{\kappa_S^i} \left(\ell_{NB}(\kappa_S^i | y^i, \mu^i) - \frac{(\log \kappa_S^i - \log \kappa_F^i)^2}{2\sigma_p^2} \right), \tag{5}$$

where σ_p^2 is the variance of the logarithmic residual between prior and the fitted dispersion κ_F^i. Moreover, the variance (σ_w^2) of the logarithmic residual between raw dispersion κ_R^i and κ_F^i is comprised of 1) the variance of sampling distribution of the logarithmic dispersion σ_x^2 and 2) σ_p^2. The σ_x^2 can be approximately obtained from a trigamma function:

$$\sigma_x^2 = \psi(\frac{m-d}{2}), \tag{6}$$

where m is the number of samples and d is the number of coefficients. Whereas, the σ_w^2 is calculated as the median absolute deviation (mad) of logarithmic residuals between pairs of κ_R^i and κ_F^i:

$$\sigma_w^2 = \underset{i}{\mathrm{mad}}(\log \kappa_R^i - \log \kappa_F^i). \tag{7}$$

Therefore, we can get the σ_p^2 by

$$\sigma_p^2 = \sigma_w^2 - \sigma_x^2, \tag{8}$$

and obtain the final dispersion κ_S^i by maximizing the posterior in equation 5.

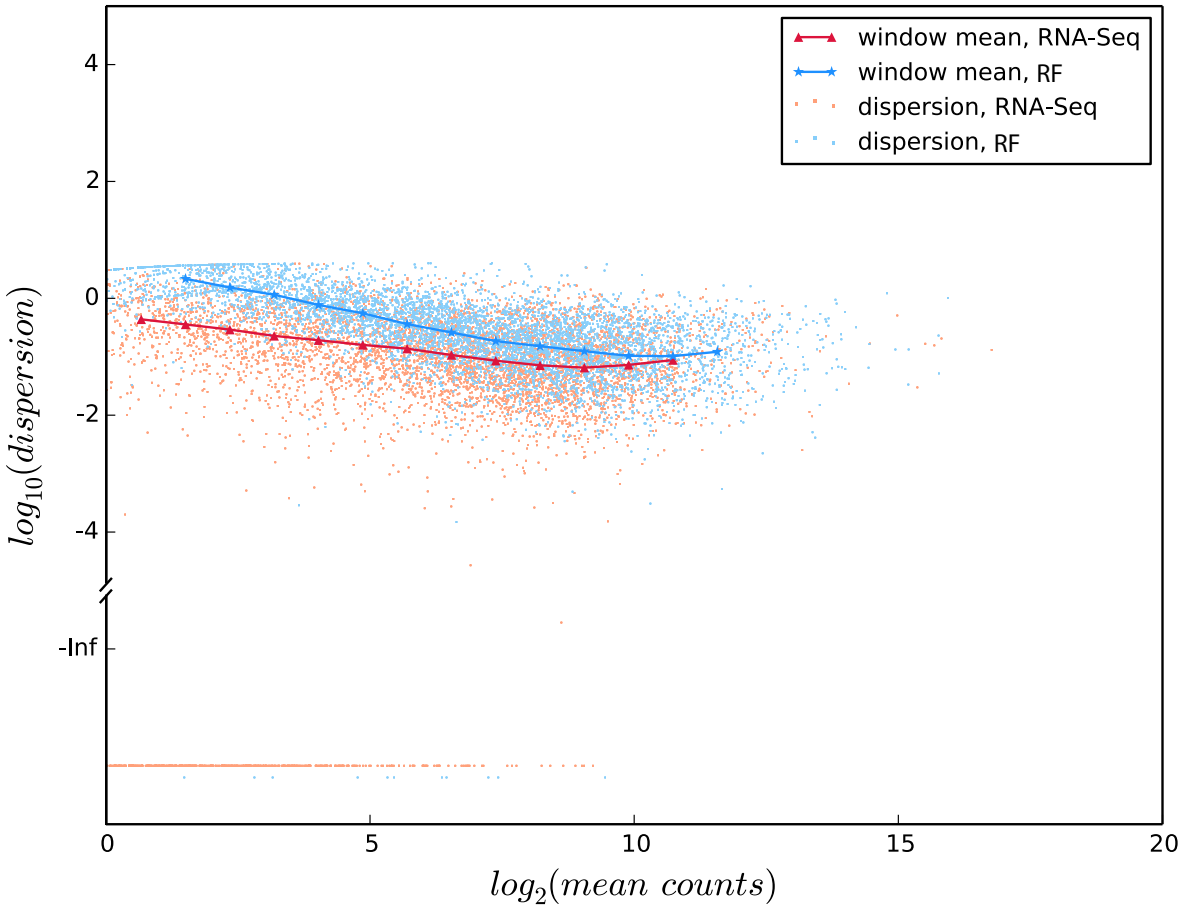

Figure S-1: Scatter plot of empirical dispersions. The X-axis is split into several bins and the median κ in each bin is highlighted and connected. The empirical κ smaller than zero are plotted at the bottom of the figure.

E Estimating Dispersion for Different Sequencing Protocols Separately

Because RNA-Seq and ribosome footprinting are different sequencing protocols, the properties of the read counts from these two protocols can vary. Therefore, we enable *RiboDiff* to infer dispersion parameters κ for different data sources. Here we show an example where estimating κ separately may be needed. The example data are from a recent publication [7].

The empirical dispersion estimates for RNA-Seq and RF counts are calculated from the following equation [5, 8, 1, 9]:

$$\sigma^2 = \mu + \kappa\mu^2. \tag{9}$$

Fig. S-1 shows the mean-dispersion relationship. It demonstrates the deviation of the empirical dispersion of RNA-Seq and ribosome footprint data in this experimental setting. The deviation between these two data sets becomes small when read count increases.

F Data Simulation

To test the performance of *RiboDiff* and compare it to other methods, we simulated the RF and RNA-Seq read count for 2,000 genes with 500 genes showing down regulated translation efficiency (TE) and 500 genes showing up regulated translation efficiency. There are three replicates for each of the two conditions (i.e., treatment and control) for RNA-Seq and RF. Therefore, count matrix dimensions are $2{,}000 \times 12$.

We first generated the mean counts for two treatments of both RF and RNA-Seq across all 2,000 genes assuming their mean counts are randomly drawn from a negative binomial distribution with parameter n and p, where $n = 1/\kappa$ and $p = n/(n+\mu)$. Then, for each mean count μ^i, we generated three count values as three replicates from a negative binomial distribution with parameter μ^i and κ^i, where κ^i is calculated as $\kappa^i = f(\mu^i) = \lambda_1/\mu^i + \lambda_0$. To simulate the genes with TE changes in two treatments, we multiply the fold difference to the mean count of the target genes, assuming the fold changes follow a gamma distribution that is observed from real data (GEO accession GSE56887). The gamma distribution has a shape parameter α and a scale parameter s, and its mean $\mu_G = \alpha \cdot s$. In the following simulation, we fix s and only change α to obtain different means for the two treatments and simulate genes having different fold changes using these two means. The fold increase F_I is obtained by

$$F_I = X_G(\alpha, s) + 1, \tag{10}$$

where X_G is a random vector containing 500 elements generated from a gamma density function. And the fold decrease F_D is obtained by

$$F_D = \frac{1}{F_I}. \tag{11}$$

Here, we simulated five groups of count data. In each group, 1,000 out of 2,000 genes showing TE changes:

- mean count has a fold change only for RF count, with $\alpha = 0.8$;
- mean count has a fold change only for mRNA count, with $\alpha = 0.6$;
- mean count has a fold change only for RF count, with $\alpha = 1.5$;
- mean count has a fold change only for mRNA count, with $\alpha = 1.5$;
- mean count has a fold change for RF with $\alpha = 0.8$ AND for mRNA with $\alpha = 0.6$, referred as "combined" in Fig. S-2.

Note that in the last group, if the gene has fold increase in RF, it must have a fold decrease in RNA-Seq. By doing this, the effect at the mRNA level is added to the TE change outcome instead of offsetting the effect caused by RF. Other simulation parameters are as follow: for all RF and RNA-Seq, $n = 1$, $\lambda_1 = 0.1$, $\lambda_0 = 0.0001$, $s = 0.5$. The parameter p controls the scale of the count. We use 0.008 for RF and 0.0002 for mRNA. We run *RiboDiff* with the five dataset to estimate its sensitivity and specificity (Fig. S-2).

To evaluate how the number of replicates influences the dispersion estimation, RF and RNA-Seq counts for $5,000$ genes with two to ten replicates for each condition were simulated using the same way as described above. For instance, two replicates for condition A and two replicates for condition B in RF, and the same number of replicates for condition A and B in RNA-Seq. In total, we have 9 data sets, and each of them has a certain number of replicates ranging from two to ten. Next, we run *RiboDiff* on these 9 data sets:

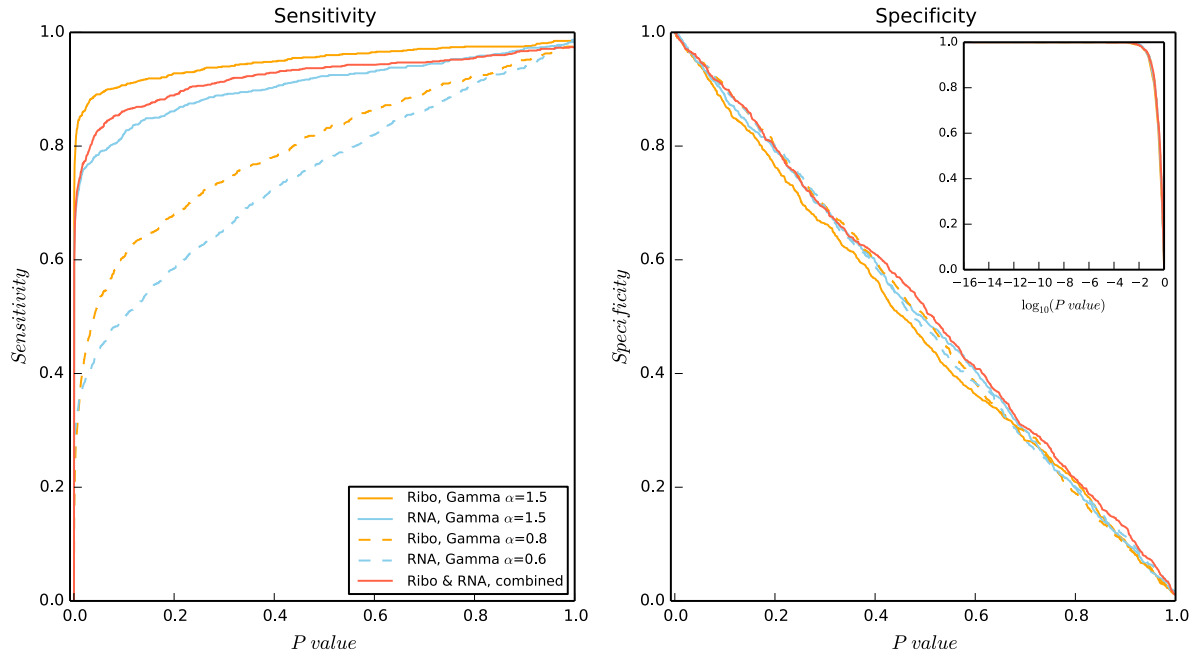

Figure S-2: Sensitivity and specificity of *RiboDiff* on simulated data.

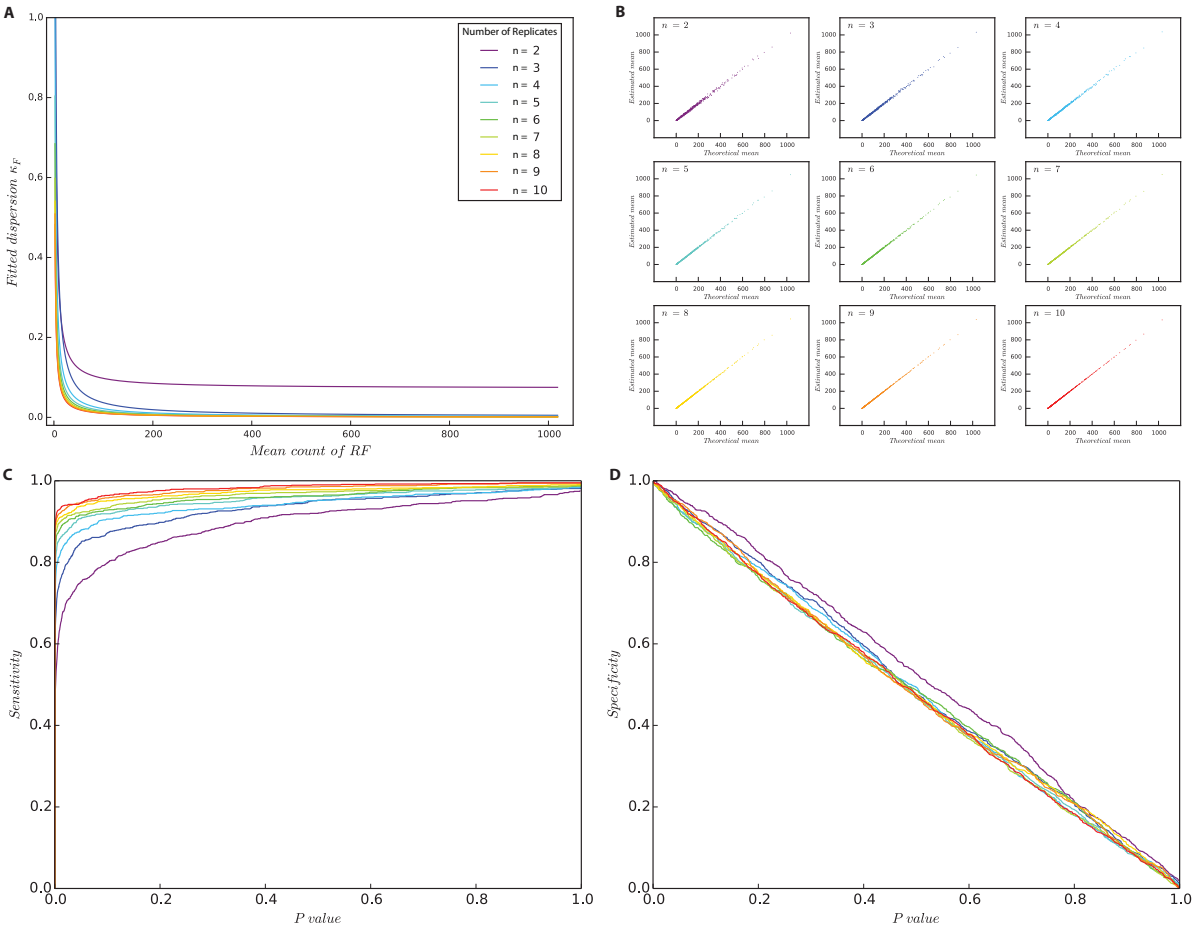

Figure S-3: Evaluation of *RiboDiff* by using different number of replicates. (A) Mean-dispersion relationship. (B) Comparison between the theoretical mean and estimated mean. (C) and (D) Sensitivity and specificity of *RiboDiff* calculated under different number of replicates.

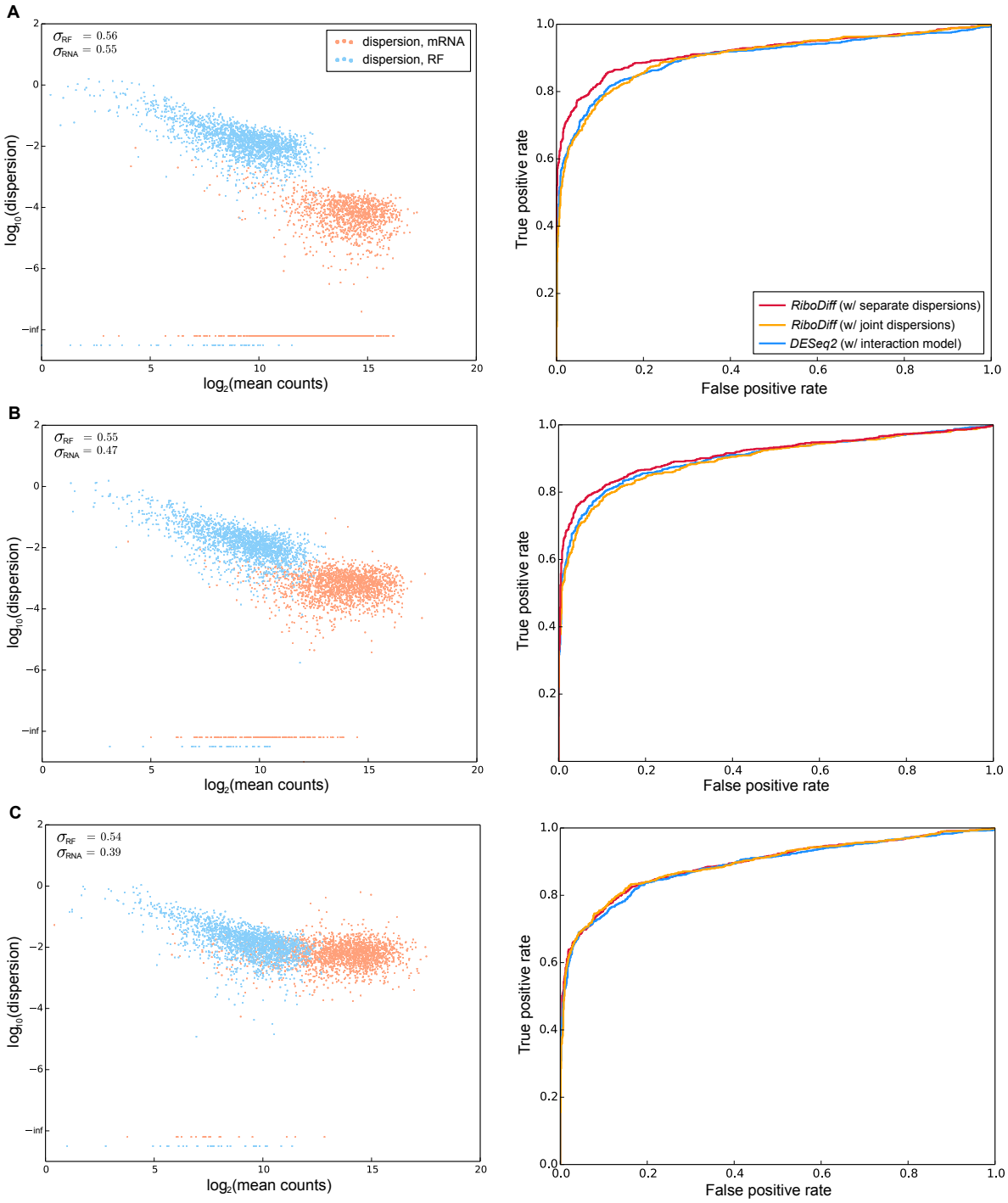

Figure S-4: Comparison of ROC curves of *RiboDiff* and *DESeq2* using simulated data. (A-C) The left panel are the dispersions of mRNA and RF; the right panel are the corresponding ROC curves. From the top to the bottom, the differences of dispersion are large, moderate and small, respectively.

RiboDiff firstly estimates the raw dispersion κ^i for each gene based on their RF and RNA-Seq counts. Then, a mean-dispersion relationship $\kappa_F = f(\mu) = \lambda_1/\mu + \lambda_0$ is obtained by regressing the raw dispersion κ^i given the mean count μ^i using GLM to learn λ_1 and λ_0. Fig. S-3A shows the mean-dispersion relationship function for different number of replicates. From this plot we can see that the estimated mean-dispersion relationships, using three to ten replicates, are rather similar to each other, whereas the result using only two replicates deviates from the rest. This indicates that the raw dispersion κ^i estimated using two replicates is less reliable. We observed that the dispersion estimates of high read count genes are larger if only two replicates are used, which can decreases true positive rate.

We use the same simulated data set to show how the number of replicates affects the latent quantity β. For each gene, there are multiple β's that represent different latent quantities, and these βs are summed up to obtain the estimated counts of RNA-Seq or RF. Hence, we compare the estimated RF count (μ^i_{RF}) of every gene i against their mean counts (theoretical means) that are used to generate the negative binomial counts in the data simulation. In Fig. S-3B, each subplot is the comparison of estimated counts (Y axis) from n replicates against the theoretical means (X axis). As we can see, the theoretical means and the estimated means correlate well in all 9 experiments (all $r > 0.99$).

Fig. S-3C and D show how sensitivity and specificity depend on a chosen p-value threshold. For the sensitivity, the area under curves for 2 to 10 replicates increases when the number of replicates increase, whereas the specificities from the same data set do not have large difference among them. This illustrates that the test is well-calibrated and that one can recommend using three replicates to achieve a close-to-best sensitivity.

As *RiboDiff* uses similar technical concept as *DESeq2* [5], we compare the performances of the two methods. Here, *DESeq2* uses a specific design formula: condition + protocol + condition:protocol. The interaction term between sequencing protocol and experimental condition represents the possible condition differences controlling for protocol type.

We simulated three data sets of RNA-Seq and RF counts where gradient differences of dispersion between mRNA and RF were applied to the two data types. The same simulation strategy was used as we described before with modifications. Briefly, $1,000$ out of $2,000$ genes were chosen to show ΔTE fold change by altering their mean counts of mRNA and RF. The following parameters were used to generate the mRNA count: $n = 1$, $p = 0.5 \times 10^{-4}$, $\lambda_1 = 0.1$, $\lambda_0 = 0.1 \times 10^{-3}$, $\alpha = 0.8$, $s = 0.5$. And for RF count, we used $n = 1$, $p = 0.1 \times 10^{-2}$, $\lambda_1 = 10.0$, $\lambda_0 = 0.01$, $\alpha = 0.8$, $s = 0.5$. Next, we multiplied the mRNA dispersion of every gene in the first data set by a factor of 10, and used the new dispersion to generate mRNA counts for the second data set. Similarly, to obtain the third data set, the original mRNA dispersions were multiplied by 100. We used the same parameters to generate the RF counts for all three data sets. In Fig. S-4, from the top to the bottom, the three dispersion plots on the left side show the three simulated data sets where mRNA dispersions are approaching to merge with RF dispersions. The ROC curves on the right side are the corresponding performances of *RiboDiff* with joint and separate dispersion estimates and *DESeq2*. Although *RiboDiff* with joint dispersion estimate performs similar to DESeq2, estimating dispersion separately yields better results under the condition of different dispersions of the two protocols.

G Results from Real Biological Data

We use previously published ribosome footprint and RNA-Seq data (GEO accession GSE56887) to compare *RiboDiff* with a Z-score based method [10]. The sequencing data were processed in a similar way as before [11], which includes trimming the adapter tail in the reads, aligning the reads, filtering the ribo-

somal RNA contamination, and counting the reads for genes, *etc.* For gene i, the change of translation efficiency ΔTE^i is calculated by

$$\Delta TE^i = \frac{K^i_{RF,A}/K^i_{RNA,A}}{K^i_{RF,B}/K^i_{RNA,B}}, \quad \text{with} \quad K^i_{t,c} = \underset{j}{\text{mean}}(y^{i,j}_{t,c}), \tag{12}$$

where t denotes the data type, as $t = \{\text{RF, RNA-Seq}\}$; c denotes the treatment condition A or B, as $c = \{A, B\}$; j indexes the replicates. $y^{i,j}_{t,c}$ means the t type of read count y of gene i in its j^{th} replicate under condition c. A Z-score was then calculated for each gene as following:

$$z^i = \frac{\Delta TE^i - \mu_{\Delta TE}}{\sigma_{\Delta TE}}, \tag{13}$$

where $\mu_{\Delta TE}$ is the mean of ΔTE of all genes; $\sigma_{\Delta TE}$ is the standard deviation. The genes with $\mid z^i \mid \geq 1.5$ are selected as significant. Fig. S-5A and B show the overlap of significant genes between *RiboDiff* and Z-score based method are limited in both TE down and up regulated gene sets. Further analysis indicates most of the significant genes detected by the Z-score based method having their mean RF counts smaller than 100 with only a few exceptional cases. In contrast, the significant genes detected by *RiboDiff* scatter over a wide range of mean RF count (Fig. S-5C and D). It is rational that for highly translated genes, it is more confident to identify significant TE change between two treatments due to enough supported read counts. This is the reason that *RiboDiff* can detect highly translated genes as significant ones even though their absolute value of Z-score are less than 1.5 ($\mid \Delta TE \mid$ below the dashed lines in Fig. S-5C and D). This comparison indicates *RiboDiff* identifies more sensible hits and is not biased towards genes with low mean count that inherently have more uncertainty rather than statistically significant differences.

Here, we also compare the new TE change gene sets detected by *RiboDiff* and the previous corresponding gene sets published in [11] (Fig. S-6). *RiboDiff* detected twice as many as before, and more than 90% genes from the old study are included in the new gene sets.

H Pipeline of ribosome footprinting and RNA-Seq data processing

The deep sequencing based ribosome footprinting has many unique features compared to regular RNA-Seq data. The flowchart of computational procedure for preprocessing the footprint data is shown in Figure S-7. Here we discuss its distinct features at each step of data processing before running *RiboDiff*'s differential test on the count input. The relevant scripts that specifically aim for doing each data processing tasks can be found in the *RiboDiff* package.

- Sequencing procedure can introduce bias to the library of replicates. It is always helpful to check the quality of the FASTQ file. Statistics on sequence GC content, length distribution, duplication level, adaptor content and kmer enrichment can provide information from different angles to identify outliers from usable libraries. Publically available tools for doing these tasks are well established and can easily be obtained online.

- In the raw FASTQ file of footprinting, rRNA contamination can take up 25 to 70% of the entire sequences of a library. We construct rRNA databases for specific organisms by collecting their rRNA sequences from SILVA [12]. Both footprint and RNA-Seq reads are aligned to the rRNA database to identify rRNA reads that need to be filtered.

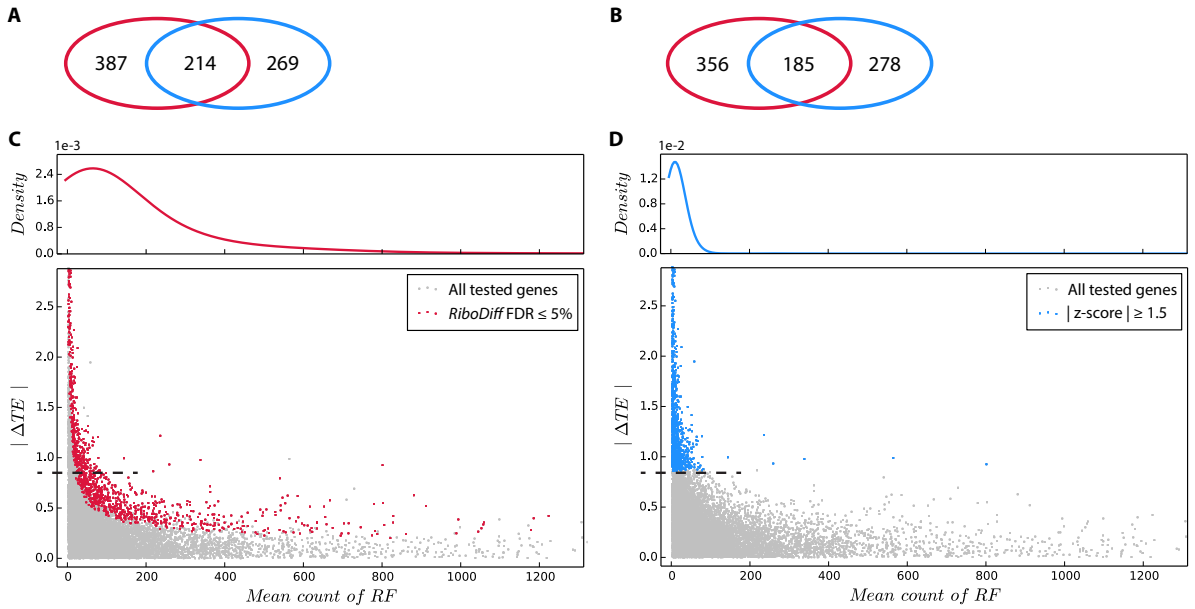

Figure S-5: Comparison between *RiboDiff* and a Z-score based method on real biological data. (**A** and **B**) Venn diagrams showing the number of overlapping and self specific genes detected by *RiboDiff* and Z-score based method. (**A**) TE down regulated genes. (**B**) TE up regulated genes. Red ellipse: results from *RiboDiff*; blue ellipse: results from Z-score based method. (**C** and **D**) Scatter plot of mean RF count against the $\mid \Delta TE \mid$. (**C**) Result of *RiboDiff*. Significant genes are labeled as red. (**D**) Result of Z-score based method. Significant genes are labeled as blue. The narrow panels above the scatter plots are the estimated density functions of significant genes on x-axes by using non-parametric kernel density estimation.

Figure S-6: Comparison of results from *RiboDiff* and Wolfe *et al.* (**A**) TE down regulated genes. (**B**) TE up regulated genes. Red circle: results from *RiboDiff*; purple circle: results reported in previous study [11].

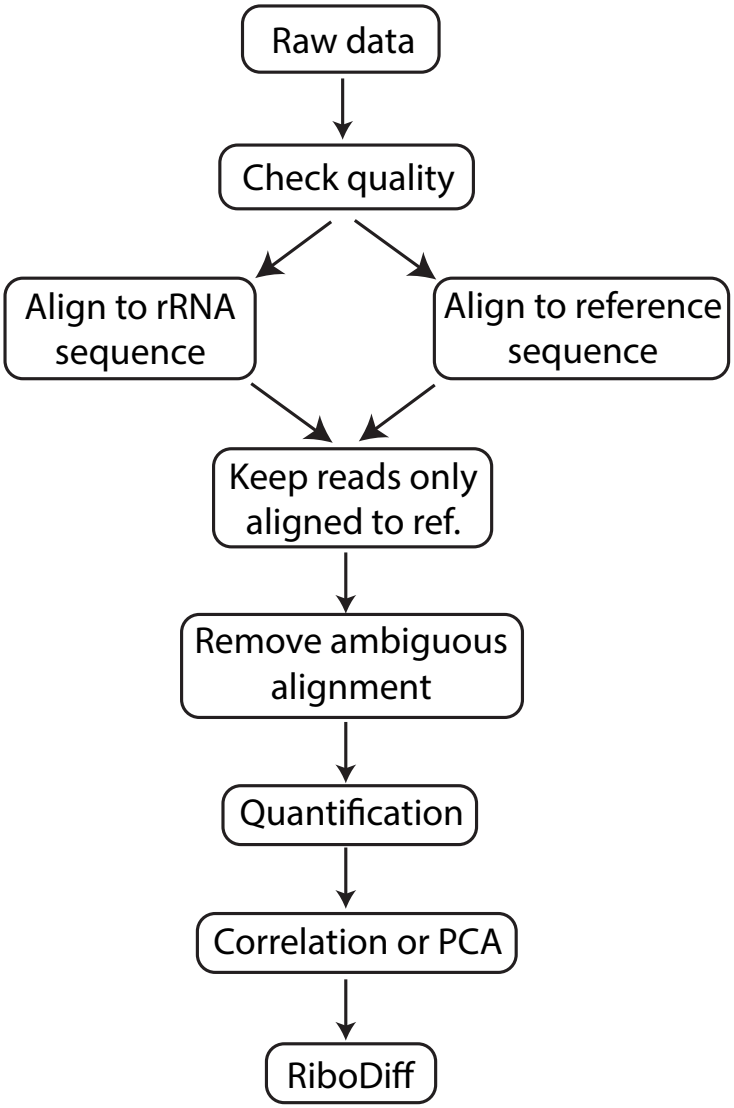

Figure S-7: The flowchart of computational pipeline for analyzing ribosome footprint data.

- Next, we use STAR [13] to align both footprint and RNA-Seq reads to the reference genome to get mapping information. STAR is an ultrafast and accurate aligner that also supports junction reads crossing exon splicing site. It also trims the linker sequence (CTGTAGGCACCATCAAT) on the 3' of footprint reads while aligning them to the reference. Ribosome protected mRNA sequences are short (normally from 20 to 40 nt), therefore, to minimize the effect of multiple mapping, only uniquely aligned reads are used.

- The identified rRNA reads from step 2 are removed from the alignment of footprint and RNA-Seq. It is also recommended to check whether footprint reads are clipped even after trimming the linker sequence. Over-clipped short reads are prone to produce ambiguous alignments.

- The last step of preprocessing the data is counting reads for each gene. We add a counting script that only takes the reads mapping the exonic regions guided by an annotation GTF file.

References

[1] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biol*, 11(10):R106, 2010.

[2] P. McCullagh and John A. Nelder. *Generalized Linear Models, Second Edition*. Chapman and Hall, 1989.

[3] Mark D Robinson and Gordon K Smyth. Small-sample estimation of negative binomial dispersion, with applications to sage data. *Biostatistics*, 9(2):321–32, Apr 2008.

[4] Simon Anders, Alejandro Reyes, and Wolfgang Huber. Detecting differential usage of exons from rna-seq data. *Genome Res*, 22(10):2008–17, Oct 2012.

[5] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biol*, 15(12):550, 2014.

[6] Hao Wu, Chi Wang, and Zhijin Wu. A new shrinkage estimator for dispersion improves differential expression detection in rna-seq data. *Biostatistics*, 14(2):232–43, Apr 2013.

[7] Christian Gonzalez, Jennifer S Sims, Nicholas Hornstein, Angeliki Mela, Franklin Garcia, Liang Lei, David A Gass, Benjamin Amendolara, Jeffrey N Bruce, Peter Canoll, and Peter A Sims. Ribosome profiling reveals a cell-type-specific translational landscape in brain tumors. *J Neurosci*, 34(33):10924–36, Aug 2014.

[8] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–40, Jan 2010.

[9] Davis J McCarthy, Yunshun Chen, and Gordon K Smyth. Differential expression analysis of multifactor rna-seq experiments with respect to biological variation. *Nucleic Acids Res*, 40(10):4288–97, May 2012.

[10] Carson C Thoreen, Lynne Chantranupong, Heather R Keys, Tim Wang, Nathanael S Gray, and David M Sabatini. A unifying model for mtorc1-mediated regulation of mrna translation. *Nature*, 485(7396):109–13, May 2012.

[11] Andrew L Wolfe, Kamini Singh, Yi Zhong, Philipp Drewe, Vinagolu K Rajasekhar, Viraj R Sanghvi, Konstantinos J Mavrakis, Man Jiang, Justine E Roderick, Joni Van der Meulen, Jonathan H Schatz, Christina M Rodrigo, Chunying Zhao, Pieter Rondou, Elisa de Stanchina, Julie Teruya-Feldstein, Michelle A Kelliher, Frank Speleman, John A Porco, Jr, Jerry Pelletier, Gunnar Rätsch, and Hans-Guido Wendel. Rna g-quadruplexes cause eif4a-dependent oncogene translation in cancer. *Nature*, 513(7516):65–70, Sep 2014.

[12] Christian Quast, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. The silva ribosomal rna gene database project: improved data processing and web-based tools. *Nucleic Acids Res*, 41(Database issue):D590–6, Jan 2013.

[13] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21, Jan 2013.