RiboDiff: Detecting Changes of Translation Efficiency from Ribosome Footprints

Yi Zhong,[1,*] Theofanis Karaletsos,[1,†] Philipp Drewe,[2,†] Vipin Sreedharan,[1] Kamini Singh,[3] Hans-Guido Wendel,[3] and Gunnar Rätsch[1,*]

[1] Computational Biology Program, Sloan Kettering Institute, 1275 York Avenue, New York, USA
[2] Max Delbrück Center for Molecular Medicine, Robert-Rössle-Str. 10, 13125 Berlin, Germany
[3] Cancer Biology Program, Sloan Kettering Institute, 1275 York Ave, New York, USA

Abstract

Motivation

Deep sequencing based ribosome footprint profiling can provide novel insights into the regulatory mechanisms of protein translation. However, the observed ribosome profile is fundamentally confounded by transcriptional activity. In order to decipher principles of translation regulation, tools that can reliably detect changes in translation efficiency in case-control studies are needed.

Results

We present a statistical framework and analysis tool, *RiboDiff*, to detect genes with changes in translation efficiency across experimental treatments. *RiboDiff* uses generalized linear models to estimate the over-dispersion of RNA-Seq and ribosome profiling measurements separately, and performs a statistical test for differential translation efficiency using both mRNA abundance and ribosome occupancy.

Availability

Source code and documentation are available at `http://github.com/ratschlab/ribodiff`. Supplementary Material can be found at `http://bioweb.me/ribodiff`.

Contact

zhongy@cbio.mskcc.org and raetsch@cbio.mskcc.org.

1 Introduction

The recently described ribosome profiling technology [6] allows the identification of RNA fragments that were bound by the ribosome complex. It provides valuable information on ribosome occupancy and, thereby indirectly, on protein synthesis activity. This technology can be leveraged by combining the measurements from RNA-Seq expression estimates in order to determine a gene's translation efficiency (TE): $TE = A_{\mathrm{RF}} \,/\, A_{\mathrm{mRNA}}$, where A_{mRNA} and A_{RF} are the mRNA and ribosome footprint (RF) read counts, respectively [7, 5, 13]. The normalization by mRNA abundance is designed to remove transcriptional activity as a confounder of RF abundance. The TEs in treatment/control experiments can then be compared to identify genes most affected w.r.t. translation efficiency; for instance, [13] considered a ratio (a.k.a. fold-change) of the TEs of treatment and control. However, what these initial approaches and analyses only take into account partially, is that one typically only obtains uncertain estimates of the mRNA and ribosome abundance. In particular for lowly

*to whom correspondence should be addressed
†authors contributed equally

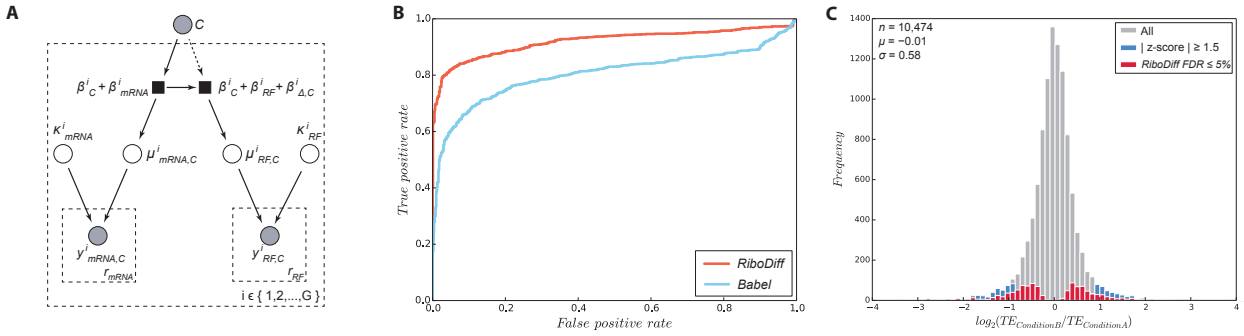

Figure 1: (**A**) Graphical model representing *RidoDiff* (Gray circle: observable variables; empty circle: unobservable variables; black square: functions; r denotes biological replicates; G is the number of genes). The dash lined denotes the relationship that we aim to test (see main text for details). (**B**) Receiver operating characteristic (ROC) curve of *RiboDiff* and *Babel* using simulated data. (**C**) Comparison of results from z-score based analysis and *RiboDiff*. Blue, $|\ z\text{-score}\ | > 1.5$; Red, FDR < 0.05 (data from GEO accession GSE56887).

expressed genes, the error bars for the ratio of two TE values can be very large. As in proper RNA-Seq analyses, one should consider the uncertainty in these abundance measurements when making statements about differentiality. For RNA-Seq, this has been described in various ways often based on generalized linear models taking advantage of dispersion information from biological replicates (for instance, [11, 2, 3]). In [14, 16]i, a way to adopt an approach for RNA-Seq analysis for this problem was described, which had several conceptual and practical limitations. Here, we describe a novel statistical framework that also uses generalized linear model to detect effects of a treatment on RNA translation. Additionally, our approach accounts for the fact that two different sequencing protocols with distinct statistical characteristics are used. We compare it to a recently published tool *Babel* [10].

2 Methods

In sequencing-based ribosome footprint profiling, the RF read count is naturally confounded by mRNA abundance (Fig. 1A). We seek a strategy to compare RF measurements taking mRNA abundance into account, in order to accurately discern the translation effect in case-control experiments. We model the vector of mRNA and RF read counts y^i_{mRNA} and y^i_{RF}, respectively, and for gene i with Negative Binomial (NB) distributions, as described before (for instance, [11, 8, 3]): $y^i \sim NB(\mu^i, \kappa^i)$, where μ^i is the expected count and κ^i is the estimated dispersion across (biological) replicates. Formulating the problem as a generalized linear model (GLM) with the logarithm as link function, we can express expectations on read counts as a function of latent quantities related to *mRNA abundance* β_C in the two conditions ($C = \{0, 1\}$), a quantity β_{mRNA} that relates mRNA abundance to RNA-Seq read counts, a quantity β_{RF} that relates mRNA abundance to RF read counts and a quantity $\beta_{\Delta,C}$ that captures the effect of the treatment on translation. In particular, the expected mRNA read count $\mu^i_{\text{mRNA},C}$ is given by the equation $\mu^i_{\text{mRNA},C} = \log(\beta^i_C + \beta^i_{RNA})$.

We assume that transcription and translation are successive cellular processing steps and that abundances are linearly related. The expected RF read count, $\mu^i_{\text{RF},C}$, is given by $\mu^i_{\text{RF},C} = \log(\beta^i_C +$

$\beta^i_{RF} + \beta^i_{\Delta,C}$). A key point to note is that β^i_C is revealed to be a shared parameter between the expressions governing the expected mRNA and RF counts. It can be considered to be a proxy for shared transcriptional/translation activity under condition C in this context. Then, $\beta^i_{\Delta,C}$ indicates the deviation from that activity under condition C, with $\beta^i_{\Delta,C} = 0$ for $C = 0$ and free otherwise.[‡]

Fitting the GLM consists of learning the parameters β^i and dispersions κ^i given mRNA and RF counts for the two conditions $C = \{0, 1\}$. We perform alternating optimization of the parameters β^i given dispersions κ^i and the dispersion parameters κ^i given β^i, similar to the EM algorithm:

$$\beta^i = \arg\max_{\beta^i} \ell_{glm}(\beta^i | y^i, \kappa^i) \quad \text{and} \quad \kappa^i = \arg\max_{\kappa^i} \ell_{NB}(\kappa^i | y^i, \mu^i).$$

As experimental procedures for measuring mRNA counts and RF counts differ, we enable the estimating of separate dispersion parameters for the data sources of RNA-Seq and RF profiling to account for different characteristics. As in [2], we use the mean-dispersion relationship $\kappa = f(\mu) = \lambda_1/\mu + \lambda_0$ and a Gamma distribution to obtain the function $f(\mu)$. We perform empirical Bayes shrinkage [8] to shrink κ^i towards $f(\mu)$ to stabilize estimates. See Section D in Supplementary Material for details.

In a treatment/control setting, we can then evaluate whether a treatment ($C = 1$) has a significant differential effect on translation efficiency compared to control ($C = 0$), which is equivalent to determining whether the inferred parameter $\beta_{\Delta,1}$ differs significantly from 0. This is whether the relationship denoted by the dashed line in Fig. 1A is needed or not. We can compute significance levels based on the χ^2 distribution by analyzing log-likelihood ratios of the Null model ($\beta^i_{\Delta,1} = 0$) and the alternative model ($\beta^i_{\Delta,1} \neq 0$).

3 Results and Discussion

We simulated data to illustrate the performance of *RiboDiff* and to compare it with a recently published tool *Babel*. For details on data simulation see Section F in Supplementary Material. Fig. 1B shows the receiver operating characteristic curve of *RiboDiff* and *Babel*, indicating superior quantitative performance of *RiboDiff*. We also re-analyzed previously released ribosome footprint data (GEO accession GSE56887). After multiple testing correction, *RiboDiff* detected 601 TE down-regulated genes and 541 up-regulated ones with FDR < 0.05, which is about twice as many as reported in [14]. The new TE down set includes 92.4% genes identified in the previous study, whereas the TE up set contains 94.7% previously identified ones. The result of *RiboDiff* is also compared to TE fold change analysis, which classifies genes with the most extreme Δ_{TE} as candidates (Fig. 1C). We run *RiboDiff* on a machine with 1.7 GHz CPU and 4GB RAM, it took 23 mins of computing time to finish (10,474 genes having both RF and mRNA counts).

In summary, we propose a new statistical model and analysis tool to analyze the effect of a of a treatment on RNA translation. It assumes a rich model of data generation and can be used accurate differential testing. A major advantage of this method is facilitating comparisons of RF abundance by taking mRNA abundance variability as a confounding factor. Moreover, *RiboDiff* is specifically tailored to produce robust dispersion estimates for different sequencing protocols measuring gene expression and ribosome occupancy that have different statistical properties. The

[‡]We described the model in an easy to follow way. It turns out that one variable is linearly dependent and in the implementation we omit β^i_{RF}.

described approach is statistically sound and identifies a similar set of genes from a less developed method that was used in [14]. The release of this tool is expected to enable proper analyses of data from many future RF profiling experiments.

Acknowledgements This work was funded by the Marie Curie ITN framework (Grant # PITN-GA-2012-316861), MSKCC, the National Cancer Institute (R01-CA142798-01 to H.-G.W.) and the Experimental Therapeutics Center (H.-G.W.).

A Sequencing library size and normalization

The sequencing library sizes of RNA-Seq and Ribosome footprinting (RF) counts are normalized separately. We calculate the library size S similar to [8] with modifications:

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

where T denotes data type (mRNA or RF); j indexes the replicates (or samples); $y_T^{i,j}$ is the observed count of type T for gene i in replicate j. For all genes in all replicates, we add one to the count value to avoid the geometric mean across all replicates in the denominator equals to zero. The ratios of gene read counts in a given replicate to the geometric means are calculated, and we take the median of these ratios whose count is greater than one as the library size. The read counts are normalized by the library size before being used in the next step.

B The explanatory matrix of GLM

To control the observed read counts fitting into the GLM system as we described in the main text, an explanatory matrix X is designed. Here we show it in the context of linear predictor η of GLM:

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

In this X matrix example, the first four rows absorb mRNA count with two replicates for each condition. The last six rows absorb RF count with three replicates for each condition. Please note the first and second columns in X are shared between mRNA and RF counts, where we couple the two different data set. The linear predictor then are linked with negative binomial distributed mean $\mu^i_{\text{RF},C}$ and $\mu^i_{\text{mRNA},C}$ through logarithm as the link function.

C Negative binomial likelihood function

The probability mass function of 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 Cox-Reid term as suggested by Robinson *et al.* [12] to compensate bias from estimating coefficients in fitting GLM step. Here, X is the explanatory matrix with dimension of $n \times 4$ or $n \times 5$, depending on H_0 or H_1, where n is the total number of replicates of RF and mRNA data; μ^i is the vector of estimated counts; κ^i is the dispersion vector.

D Empirical Bayes shrinkage for obtaining final dispersion

We follow the approach published recently [8] to get the final dispersion κ_S^i. Assumption is based on the observation that the dispersion follows a log-normal prior distribution [15] centered at the fitted dispersion κ_F^i which is obtained from the dispersion-mean relationship $\kappa = f(\mu) = \lambda_1/\mu + \lambda_0$ (see in the main text). 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.

E Estimating dispersion for different sequencing protocols separately

As experimental procedures for representing mRNA and RF abundances can vary, such as the samples are sequenced in different platforms, we enable *RiboDiff* uses separate dispersion parameters κ for different data sources. Here we show an example that estimating κ separately is needed. The example data are from a recent publication [4].

The empirical dispersion for RNA-Seq and RF counts are calculated from the following equation [8, 11, 1, 9]:

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

Fig. 2 shows the mean-dispersion relationship. It demonstrates the deviation of empirical dispersion of RNA-Seq and ribosome footprint data in this experimental setting. The deviation between these two data sets becomes small while the count increases.

F Data simulation

We simulated the RF and mRNA read count for 2,000 genes with 500 genes showing translational efficiency down regulated and 500 genes showing up regulated. There are three replicates for each of the two treatments in both "ribosome profiling" and "RNA-Seq" counts. Therefore, the dimension of count matrix is 2,000 $\times$ 12.

We first generated the mean counts for two treatments of both RF and mRNA across all 2K genes assuming they 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 add 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 the s, only specify different α to make genes having different fold changes on their 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 as

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

Here, we simulated five groups of count data, in every group 1,000 out of 2,000 genes showing TE changes:

- mean count has fold change only for RF count, with $\alpha = 0.8$;
- mean count has fold change only for mRNA count, with $\alpha = 0.6$;

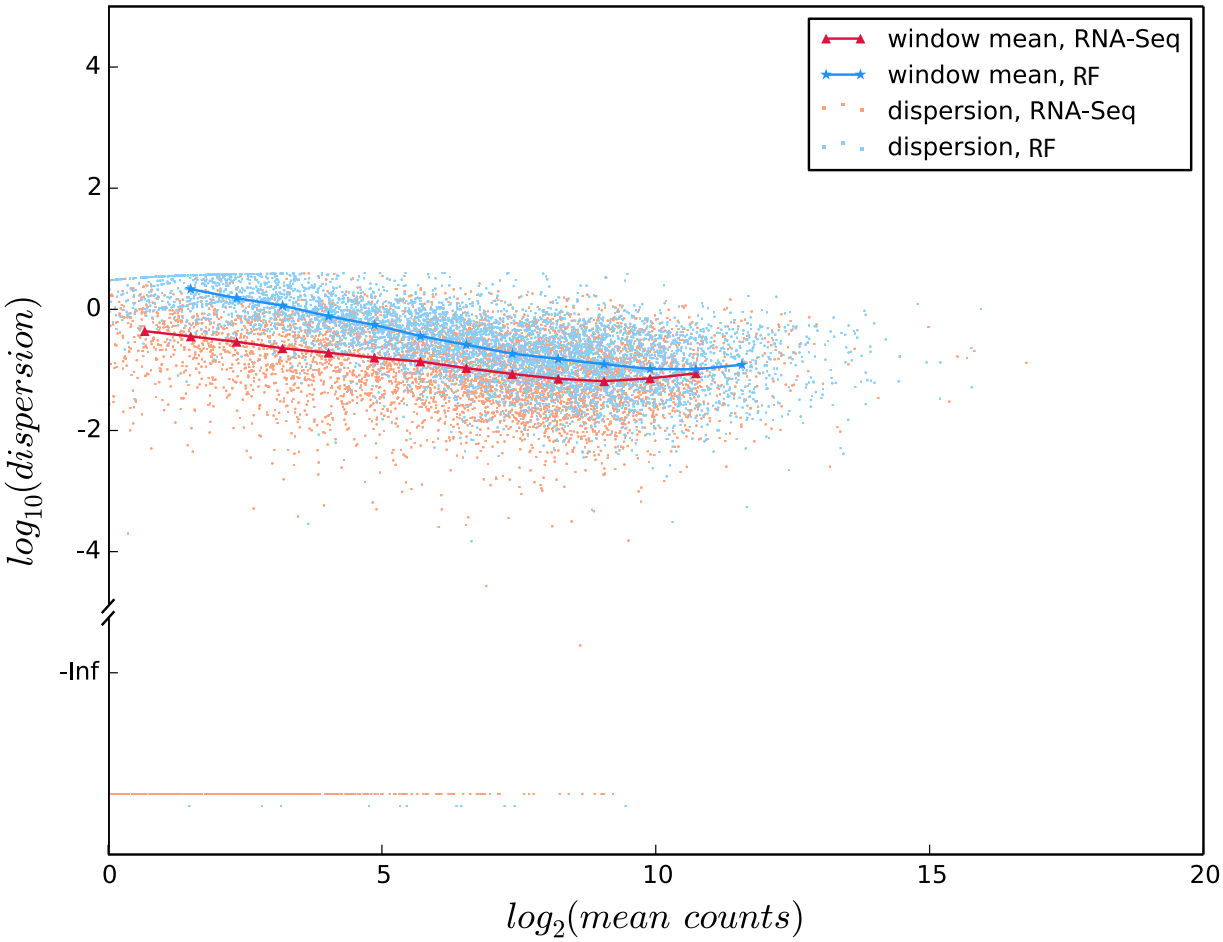

Figure 2: Scatter plot of empirical dispersions. The X-axis is split into several bins and the median κ in each bin is highlighted and connected. The empirical κ smaller than zero are plotted at the bottom of the figure.

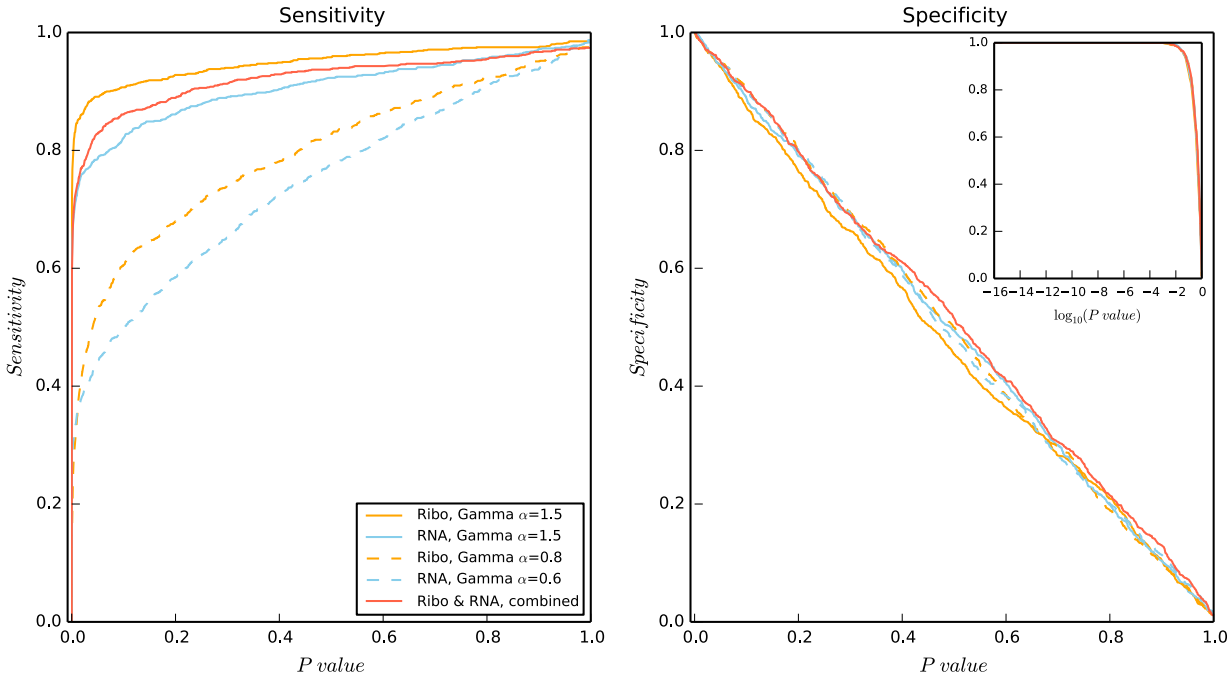

Figure 3: Sensitivity and specificity of *RiboDiff* on simulated data.

- mean count has fold change only for RF count, with $\alpha = 1.5$;
- mean count has fold change only for mRNA count, with $\alpha = 1.5$;
- mean count has fold change for RF with $\alpha = 0.8$ AND for mRNA with $\alpha = 0.6$, referred as "combined" in Fig. 3.

Note that in the last group, if the gene has fold increase in RF, it must has fold decrease in mRNA. By doing this, the effect at mRNA level is added to the TE change outcome instead of offsetting the effect caused by RF. Other parameters for simulating are as follow: for all RF and mRNA, $n = 1$, $\beta_1 = 0.1$, $\beta_2 = 0.0001$, $s = 0.5$. 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 groups of data set to estimate the sensitivity and specificity (Fig. 3). We also compared the performances of *RiboDiff* with *Babel* [10] using the simulated data of the combined setting.

References

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

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

[3] Philipp Drewe, Oliver Stegle, Lisa Hartmann, André Kahles, Regina Bohnert, Andreas Wachter, Karsten Borgwardt, and Gunnar Rätsch. Accurate detection of differential rna processing. *Nucleic Acids Res*, 41(10):5189–98, May 2013.

[4] 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.

[5] Andrew C Hsieh, Yi Liu, Merritt P Edlind, Nicholas T Ingolia, Matthew R Janes, Annie Sher, Evan Y Shi, Craig R Stumpf, Carly Christensen, Michael J Bonham, Shunyou Wang, Pingda Ren, Michael Martin, Katti Jessen, Morris E Feldman, Jonathan S Weissman, Kevan M Shokat, Christian Rommel, and Davide Ruggero. The translational landscape of mtor signalling steers cancer initiation and metastasis. *Nature*, 485(7396):55–61, May 2012.

[6] Nicholas T Ingolia, Gloria A Brar, Silvia Rouskin, Anna M McGeachy, and Jonathan S Weissman. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mrna fragments. *Nat Protoc*, 7(8):1534–50, Aug 2012.

[7] Nicholas T Ingolia, Liana F Lareau, and Jonathan S Weissman. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell*, 147(4):789–802, Nov 2011.

[8] 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.

[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] Adam B Olshen, Andrew C Hsieh, Craig R Stumpf, Richard A Olshen, Davide Ruggero, and Barry S Taylor. Assessing gene-level translational control from ribosome profiling. *Bioinformatics*, 29(23):2995–3002, Dec 2013.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] Yi Zhong, Phillip Drewe, Andrew L Wolfe, Kamini Singh, Hans-Guido Wendel, and Gunnar Rätsch. Protein translational control and its contribution to oncogenesis revealed by computational methods. *BMC Bioinformatics*, 16(Suppl 2):A6, Jan 2015.