Supplementary Information: Riborex: Fast and flexible identification of differential translation from Ribo-seq data

Wenzheng Li Weili Wang Philip J Uren Luiz OF Penalva Andrew D Smith

In this supplement we provide a more detailed description of our approach and our simulation method. We also include additional simulations that could not be included within the manuscript, both simulated and actual data, to provide a broader picture of the performance of methods, and to indicate the relative behavior of Riborex using the various engines.

1 Methods

Detailed approach: We are interested in studying translation across a set of treatment conditions. We assume a set of m treatment conditions, and a set of n genes. We let λ_{gj} denote the expected proportion of reads that originate from gene g in RNA-seq experiment for treatment condition j. We call λ_{gj} the expression level of gene g in condition j. So for all j,

$$\sum_{g=1}^{n} \lambda_{gj} = 1 \quad \text{and} \quad 0 \leq \lambda_{gj} \leq 1.$$

We similarly define π_{gj} as the expected proportion of reads for gene g in Ribo-seq data from condition j. So π_{gj} is considered as the translation level of gene g and for any j,

$$\sum_{g=1}^{n} \pi_{gj} = 1, \quad \text{and} \quad 0 \leq \pi_{gj} \leq 1.$$

We define *translation efficiency* as the ratio of these two levels:

$$t_{gj} = \pi_{gj} / \lambda_{gj}.$$

The intent of this definition is to give us a means of quantifying, for a given condition, the magnitude of translation regulation relative to some other condition. A direct physical interpretation for t_{gj} is difficult to state directly, as the intuitive translation level is altered both by the number of mRNA molecules interacting with a ribosome, and by the amount of ribosomes interacting with an individual molecule for that gene. This difficulty seems not to be present when attaching intuition to expression levels, as there is only one copy of a gene being transcribed.

In the regression framework that has become standard for analyzing gene expression, the RNA-seq read count y_{gi} for gene g in sample i is assumed to be distributed according to a negative binomial distribution:

$$y_{gi} \sim \mathrm{NB}(\mu_{gi}, \phi_g),$$

where μ_{gi} is the expected value of y_{gi}. The parameter ϕ_g is the gene-specific dispersion. The expected read count is related to treatment condition as

$$\log(\mathrm{E}(y_{gi})) = \log(\mu_{gi}) = \log(\lambda_{gj}) + \log(N_i) = x_j^T \beta_g + \log(N_i),$$

where N_i is the total counted reads. The vector x_j indicates treatment conditions for sample i, and β_g is the corresponding vector of gene-specific coefficients.

Hypothesis testing in this context is done to determine whether the expression level differs between conditions, and is typically stated in terms of model coefficients. For example, in an experiment of two conditions, we have $\beta_g = (\beta_{g1}, \beta_{g2})$ with β_{g2} indicating whether or not some perturbation has been applied, the differential expression can be detected by testing

$$H_0 : \beta_{g2} = 0 \quad \text{and} \quad H_1 : \beta_{g2} \neq 0.$$

We introduce r_{gi} to denote the Ribo-seq read count for gene g in sample i:

$$r_{gi} \sim \mathrm{NB}(\eta_{gi}, \kappa_g).$$

We use R_i to denote the sum of those Ribo-seq read counts over all genes for sample i.

Read counts of Ribo-seq experiment depends on both translation efficiency and mRNA level, so we must account for the mRNA level while modeling Ribo-seq read count. The expected count of Ribo-seq reads for gene g in sample i is then

$$\begin{aligned}\log(\mathrm{E}(r_{gi})) = \log(\eta_{gi}) = \log(\pi_{gj}) + \log(R_i) &= x_j^T \alpha_g + \log(R_i) + \log \lambda_{gj} \\ &= x_j^T (\alpha_g + \beta_g) + \log(R_i).\end{aligned}$$

With the definition of *translation efficiency*, we have

$$\log(t_{gj}) = x_j^T \alpha_g$$

Differential translation efficiency will then be captured as coefficients vector α_g. The coefficient vectors α_g and β_g can be estimated simultaneously by constructing the design matrix:

$$\begin{aligned}x_j^{\mathrm{RNA}} &= (x_{j1}, \ldots, x_{jm}, 0, \ldots, 0)^T, \\ x_j^{\mathrm{Ribo}} &= (x_{j1}, \ldots, x_{jm}, x_{j1}, \ldots, x_{jm})^T.\end{aligned}$$

And similarly the coefficient vector is reorganized as

$$\gamma_g = (\beta_{g1}, \ldots, \beta_{gm}, \alpha_{g1}, \ldots, \alpha_{gm}).$$

With these modifications, detecting genes with differential efficiency of translation can be done using established frameworks for gene expression. The hypothesis tests now focus on α. This strategy can be used in the exact same way even if one drops the assumption that read counts follow a negative binomial distribution. For example, the Voom software (Law *et al.*, 2014) does not assume a negative binomial, but still assumes a GLM to describe the data and test hypotheses. Law *et al.*, 2014 provide rationale for why their assumptions may lead to more accurate results, and we expect that in some circumstances an approach like that used in Voom will have value in analysis of Ribo-seq data.

One important consideration when applying the above strategy with existing RNA-seq analysis frameworks (*e.g.* edgeR, DESeq2 and Voom; see below) whether or not the RNA-seq and Ribo-seq read counts

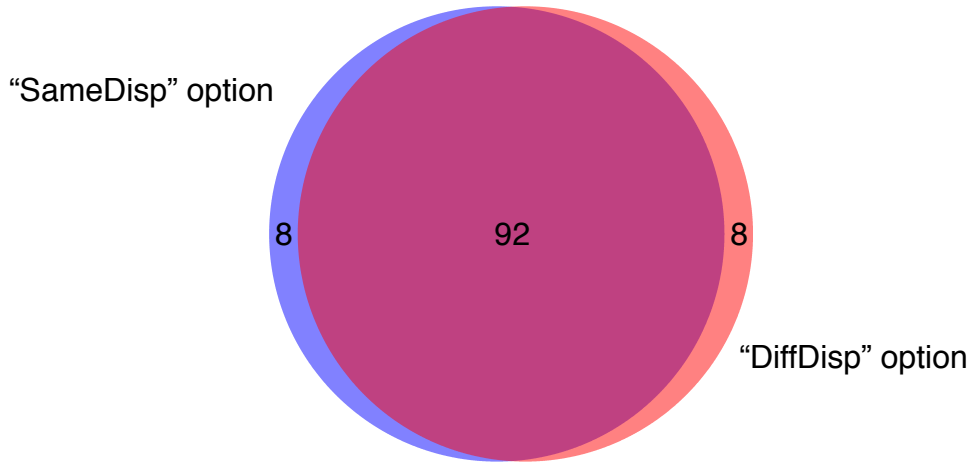

Figure S1: **Similar genes are identified as top differentially translated under different dispersion assumptions.** Riborex implements two assumptions on dispersion when using edgeR as the engine. The parameter setting "SameDisp" estimates a single dispersion for both RNA-seq and Ribo-seq data, while the "DiffDisp" parameter estimates dispersions separately for RNA-seq and Ribo-seq. Among the 100 differentially translated genes detected in the PC3 data set, 92 genes are detected under both parameter settings.

should have different dispersions. Intuitively, since the RNA-seq and Ribo-seq experiments are done separately using different steps in their protocols, one can easily rationalize why the dispersions should be estimated separately for these two data types. However, we have no prior empirical evidence that a single dispersion cannot appropriately explain sources of variation in Ribo-seq and RNA-seq simultaneously. To directly address this question, we applied our strategy using the three selected RNA-seq analysis frameworks as the computational engine. For edgeR, we implemented our strategy with both assumptions: estimating a single dispersion ($\phi_g = \kappa_g$) and two distinct dispersions (allowing $\phi_g \neq \kappa_g$). As shown in Figure S1, these two different assumptions about dispersion show very high overlap among the top differentially translated genes. Also, applied on simulated data, we found both dispersion estimation methods to have very similar performance (data not shown).

The three engines used by Riborex: Riborex uses three existing RNA-seq data analysis frameworks: DESeq2 (Love *et al.*, 2014), edgeR (Robinson *et al.*, 2010) and Voom (Law *et al.*, 2014) as engines for detecting differentially translated genes from Ribo-seq data. Among them, DESeq2 and edgeR take raw read counts as input and model them as assuming the counts follow a negative binomial distribution, which is often used to describe count data with over-dispersion. Both DESeq2 and edgeR estimate the dispersion by sharing information across genes to overcome the limitations of typical small numbers of replicates in RNA-seq experiments. But they differ in the details of how the dispersion is estimated. DESeq2 treats each gene separately, estimates gene-wise dispersion using maximum likelihood and then fits a smooth curve for the relationship between dispersion and mean read count. The final dispersion is estimated by shrinking the gene-wise dispersion estimates toward the values predicted by the curve. edgeR uses a weighted likelihood empirical Bayes approach to estimate the gene-wise dispersion (McCarthy *et al.*, 2012). Voom is based on limma (Ritchie *et al.*, 2015) which was originally designed for analyzing micro-array data. Voom extended this approach for RNA-seq data, and transforms the raw read counts into counts per million (cpm) and uses

log-cpm within the limma context. Substantial research has advanced our understanding of how dispersion should be estimated, but this research remains highly active. In implementing Riborex, we adopted the following perspectives: (1) There are merits to each of the engines we selected. (2) The best way to choose between engines depends less on the specifics of Ribo-seq, and more on the experimental design. (3) As a consequence, we advise a user of Riborex to select an engine that they would use for an RNA-seq experiment with a similar experimental design. We have therefore implemented Riborex to allow researchers to choose between edgeR, DESeq2 and Voom using the same criteria as they would for RNA-seq analysis, whether based on specific technical criteria, or simply familiarity.

We note an additional consideration. A common challenge in RNA-seq data analysis is the high variance of log fold change estimated for genes with low read count. Diament and Tuller, 2016 have exposed this issue as having a potentially higher impact on variability and reproducibility in Ribo-seq analysis, as the hypotheses we are testing involve fold change across not only conditions, but also data types (RNA-seq and Ribo-seq). The experimental design is then necessarily more intricate. Since Riborex is built upon existing engines, it inherits their strategies for overcoming difficulties caused by low read counts. Specifically, DESeq2 shrinks the log fold change estimates towards zero when little information is available for a gene because of low read counts. When genes have very low read counts, edgeR squeezes their gene-wise dispersions towards the overall trend, as there is little statistical information available to estimating their own dispersions. This issue warrants future study, and ongoing investigation in the context of RNA-seq will be relevant to analysis of Ribo-seq data.

Differences between Riborex and existing methods: Xtail (Xiao *et al.*, 2016), RiboDiff (Zhong *et al.*, 2016) and Babel (Olshen *et al.*, 2013) take raw read counts as input and model them as following a negative binomial distribution. They all use theory that emerged in the context of gene expression analysis. When DESeq2 and edgeR are used as engines, Riborex also models the raw read counts as following a negative binomial distribution. When Voom is used, the raw read counts are transformed into log-cpm and used in a normal linear model. RiboDiff adopted the same GLM approach used by DESeq2 for dispersion estimation, so when DESeq2 is used as the Riborex engine, the modeling approach is nearly mathematically equivalent to that of RiboDiff, even though we have different explanations for the variables involved. RiboDiff and Riborex detect differentially translated genes by fitting RNA-seq and Ribo-seq read counts in a single GLM with a modified design matrix. RiboDiff provides two options for estimating either the same or different dispersions for RNA-seq and Ribo-seq data. Riborex only allows this when using edgeR as the engine. This similarity of the underlying model, and of the estimation procedures, makes it surprising that RiboDiff is both (a) much slower and (b) less accurate than Riborex. Xtail directly uses DESeq2 to estimate the mean and dispersion of RNA-seq and Ribo-seq counts separately, which means fitting two completely independent GLMs. Xtail then derives a discrete probability distribution for the log2 fold changes estimated from RNA-seq and Ribo-seq (also done separately), and then a joint distribution matrix is generated from those two probability density distributions to test if the log2 fold changes from RNA-seq and Ribo-seq data are significantly different. Our evaluations were not comprehensive, but we have found no advantages to this additional complexity in Xtail, and the main consequence is an extreme computational burden. Babel applies an errors-in-variables regression to assess the significance of changes within and between conditions for translation regulation, and depends on edgeR for dispersion estimation.

All existing RNA-seq data analysis tools (DESeq2, edgeR, Voom) need to normalize the read counts to account for different sequencing depths. The normalization methods used by edgeR and Voom are based on the assumption that most genes show similar expression between different samples (Robinson and Oshlack, 2010). DESeq2 uses the median-of-ratios method already used in DESeq (Anders and Huber, 2010). Hard-

castle and Kelly, 2010 showed that if there is a large proportion of genes that are differentially expressed in a single direction, DESeq2 will fail to capture those changes. As Riborex inherits the normalization methods used in its engines, for both RNA-seq and Ribo-seq, if there is a global change impacting translation efficiency across the transcriptome, then Riborex might fail to capture many of those changes, resulting in substantial false negatives. This would be a common failure for Xtail, RiboDiff and Babel, because of all these tools rely on the same assumptions in normalization.

2 Simulations

Obtaining and pre-processing data: All raw data was downloaded in SRA format from the NCBI SRA database, and FASTQ files were extracted using the SRA toolkit. We used `fastqc` to ensure data quality. We used `fastx-clipper` to trim adaptors for both RNA-seq and Ribo-seq reads. For Ribo-seq we retained only reads longer than 25bp after trimming. We used the GENCODE reference transcriptome. The reference genome for human is hg19 and reference genome for mouse is mm10. We mapped reads for both RNA-seq and Ribo-seq using `TopHat`, allowing at most two mismatches each alignment. Read counts for genes included in the reference transcriptome were obtained by `htseq-count`.

Simulations based on real data: Here we explain details of the simulations that appear in Figure 1, and Figures S2-S6 (which we discuss below).

We simulated data from a published Ribo-seq data-set (SRA accession number PRJEB7498) (Schafer *et al.*, 2015) which investigated strain-specific translation regulation in the heart and liver of spontaneously hypertensive rats. For both the heart and liver data-set, this paper included five replicates for the NB-*lx* reference strain for both RNA-seq and Ribo-seq experiments. We performed our first simulation experiment based on the heart data and we refer to this as the rat heart data-set hereafter. In each simulation, we randomly chose two replicates to use as control samples and another two to use as treated samples. We omitted those genes with zero sum of read counts across all samples, leaving 14622 expressed genes. To broadly understand the performance of different methods, we explored their behaviors under four different scenarios, varying number of true differentially translated genes and different magnitudes of the change in the translation level.

In the first scenario, we randomly sampled 1% of the genes, and assigned 2-fold change to both RNA-seq and Ribo-seq data in one of the conditions to serve as negative controls. Then we randomly sampled another 1% of the genes (not explicitly preventing overlap with the first 1%) and with equal probability randomly chose to either: (1) assign 2-fold change to RNA-seq and 4-fold change to Ribo-seq, or (2) assign 4-fold change to RNA-seq and 2-fold change to Ribo-seq. By doing this, 1% of genes are implanted with a 2-fold change in translation efficiency and the direction of the change is random. In the second scenario, we assigned a fold change of 4 to the 1% genes that are randomly selected as negative controls in both Ribo-seq and RNA-seq data. We altered the magnitude of the fold change used for the 1% genes that are differentially translated: we assigned 2-fold and 8-fold change to RNA-seq and Ribo-seq, respectively, or 8-fold and 2-fold to RNA-seq and Ribo-seq, respectively. This simulates a 4-fold change in translation efficiency for those 1% of genes, with a random direction of change. For the third scenario, we randomly sampled 10% of the genes as negative controls, and another 10% as truly differentially translated genes, and the fold change assignments remain the same as in the first scenario to implant a 2-fold change in translation efficiency. Lastly, for the fourth scenario, we randomly sampled 10% of the genes as negative controls, and another 10% as truly differentially translated genes, and the fold change assignments remain the same as in the second scenario to implant a 4-fold change in translation efficiency. In all of these simulations, unchanged

genes are considered true negatives, along with those genes that are assigned the same fold change in both RNA-seq and Ribo-seq. The rationale for including negative controls that have both RNA-seq and Ribo-seq levels altered is that any method that does not properly weight the contribution of the two types of data might have difficulty if both levels are changed between conditions.

We also applied the above simulation using data from a second public ribosome profiling data-set (GEO session number GSE62134) (Diaz-Muñoz *et al.*, 2015) which studies the post-transcriptional regulation of mRNA by the RNA-binding protein HuR in mouse B cells (we call this the mouse HuR data-set). This data-set has four replicates and we simulated data-sets under the four scenarios as described with the rat heart data-set. For the mouse HuR data-set, the total number of expressed genes in each simulated data-set is 11876.

3 Results

Obtaining different tools: Riborex (version 1.2.3) was used during the experiments and is available at `https://github.com/smithlabcode/riborex`. For the engines, the version of DESeq2 was 1.12.3, the version of edgeR was 3.14.0 and the version of limma (for Voom) was 3.28.11. We installed Xtail (version 1.1.5) from `https://github.com/xryanglab/xtail` and RiboDiff (version 0.2.1) from `https://github.com/ratschlab/RiboDiff`, and we installed Babel (version 0.3-0) from Bioconductor.

Comparing methods on simulated data-sets: We applied Riborex (using DESeq2, edgeR, Voom) and other existing methods on the two sets of data-sets simulated from rat heart and mouse HuR data. For results generated by each method, we put the FDR threshold at 0.05. We compared different methods in terms of three metrics: sensitivity, positive predictive value (PPV) and F score. We prefer PPV over specificity here because we assume most genes are true negatives, so specificity saturates near 1.0. The average results for different methods, in terms of each metric, were calculated to measure accuracy. Error bars in the figures show one standard deviation based on the 100 simulations.

When the magnitude of change in translation efficiency is set to $4\times$, all methods have extremely similar performance (see Figure 1). Riborex and Xtail have a slight advantage over RiboDiff and Babel. From Figure S2, with difference in translation efficiency set to $2\times$, under both scenarios where 1% or 10% genes are differentially translated, Riborex using the DESeq2 engine and Xtail have similar results. In these cases, both are much better than RiboDiff and Babel. When edgeR or Voom is used as the engine, Riborex is more sensitive to the number of implanted differentially translated genes, but accuracy is much higher for the 10% implanted genes compared with the 1%. RiboDiff performs particularly poorly under this scenario.

We further explored the behaviors of different methods in terms of sensitivity and positive predictive values. Figures S3 and S4 show sensitivity of the methods in the same simulations, for a $2\times$ and $4\times$ difference in translation efficiency, respectively. The strongest trend in these results is for Riborex (DESeq2 and edgeR engines) and Xtail to show accuracy consistently at the top.

The results for positive predictive value are shown in Figures S5 and S6, for difference in translation efficiency of $2\times$ and $4\times$, respectively. At $2\times$ difference, RiboDiff shows higher PPV than sensitivity (Figure S3), but the s.d. of PPV is also high. The other methods have performance similar to what they showed for sensitivity, indicating a general balance of false-positives and false-negatives. From Figure S6, where the magnitude of difference in translation is set to 4-fold, each method has significantly improved accuracy.

To provide a baseline for understanding our results in the context of the more comprehensive comparison done by Xiao *et al.*, 2016, we used a simulated data-set provided by the authors. This simulated data-set

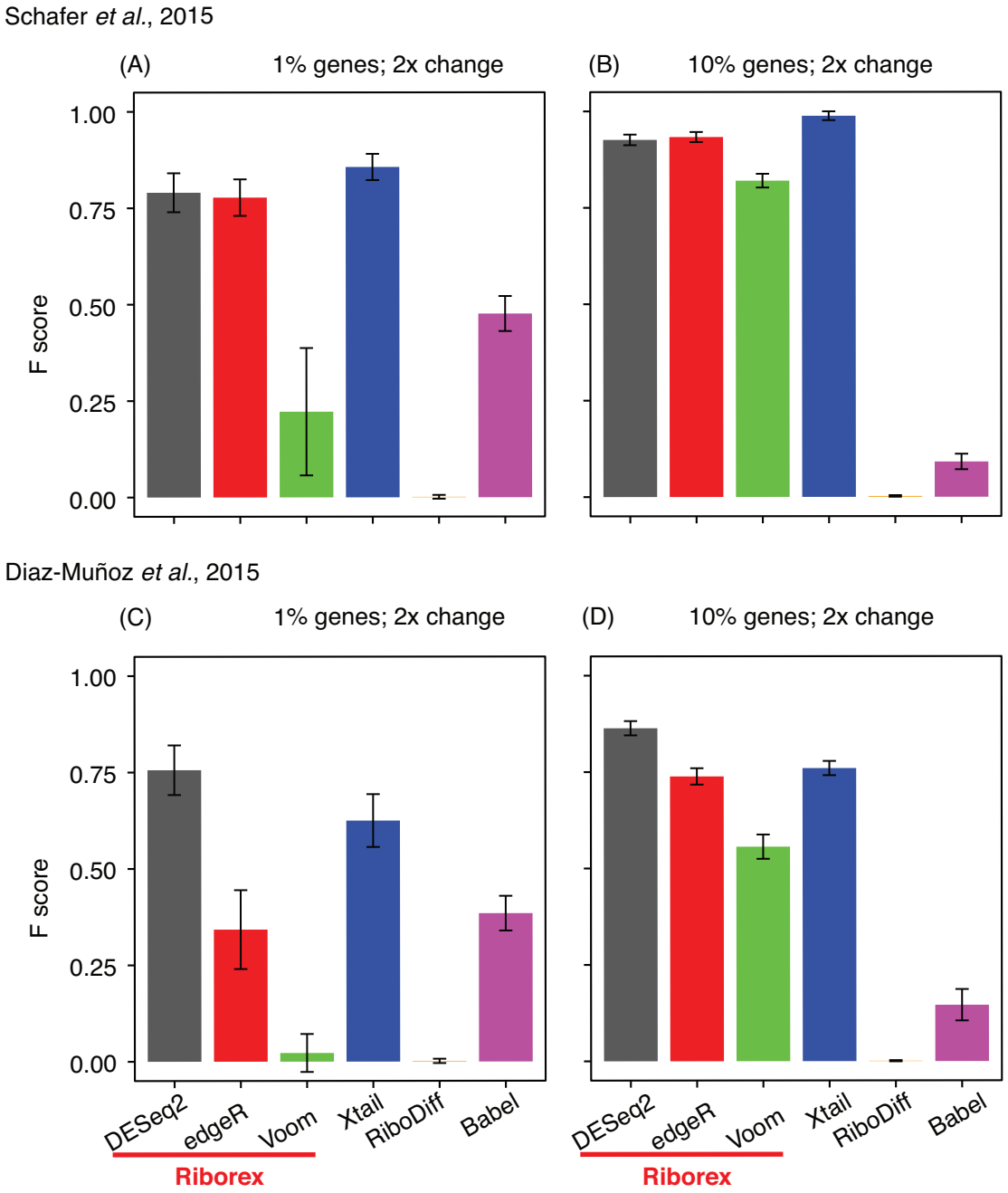

Figure S2: **F scores for different methods identifying implanted 2-fold change to translation efficiency.** Different numbers of differentially translated genes were implanted in the rat heart (panels A and B) and mouse HuR data sets (panels C and D). (A) and (C): 1% of genes were implanted with 2-fold change in translation efficiency. (B) and (D): 10% of genes were implanted with 2-fold change in translation efficiency. Barplots give average F score from 100 simulations and error bars represent the standard deviation.

is based on a published data-set (GEO session number GSE62134; Diaz-Muñoz *et al.*, 2015), with fold

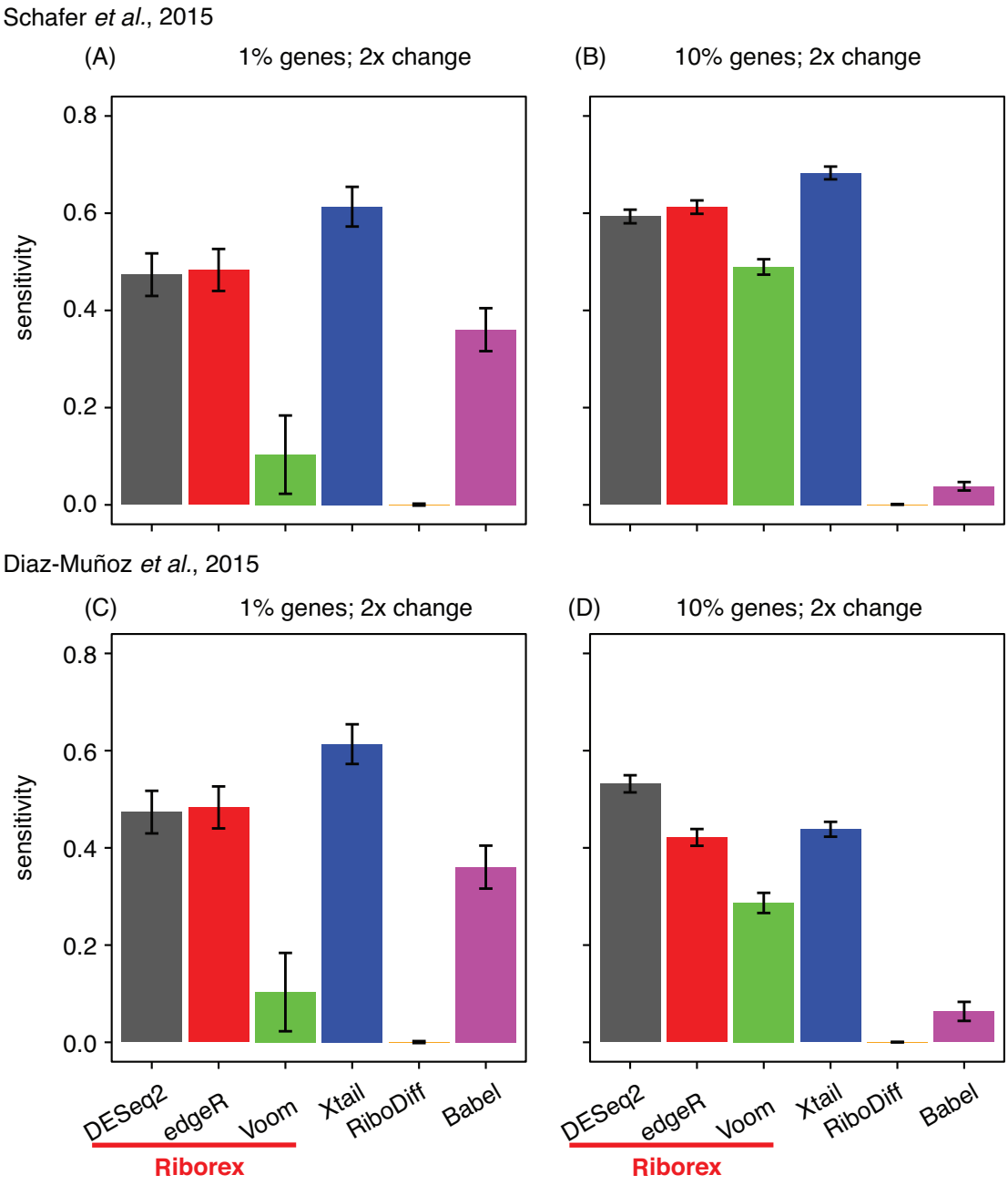

Figure S3: **Sensitivity of different methods identifying implanted 2-fold change to translation efficiency.** Different numbers of differentially translated genes were implanted in the rat heart (panels A and B) and mouse HuR data sets (panels C and D). (A) and (C): 1% of genes were implanted with 2-fold change in translation efficiency. (B) and (D): 10% of genes were implanted with 2-fold change in translation efficiency. Barplots give average sensitivity from 100 simulations and error bars represent the standard deviation.

changes assigned following a gamma distribution (shape=0.6, scale=0.5) (our relatively limited comparison used constant fold change, ensuring that more control on magnitude of genes selected to differ). In this

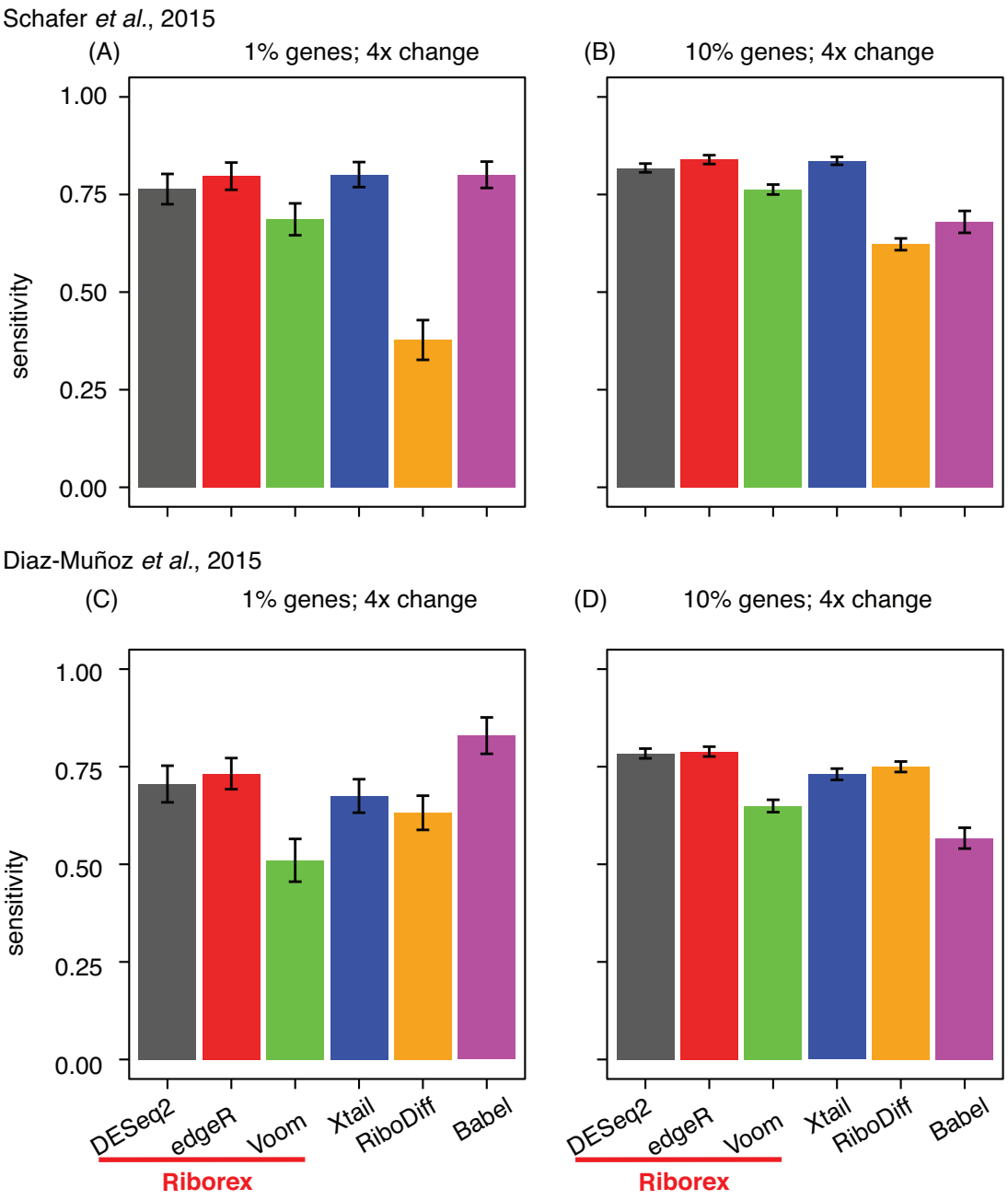

Figure S4: **Sensitivity of different methods identifying implanted 4-fold change to translation efficiency.** Different numbers of differentially translated genes were implanted in the rat heart (panels A and B) and mouse HuR data sets (panels C and D). (A) and (C): 1% of genes were implanted with 4-fold change in translation efficiency. (B) and (D): 10% of genes were implanted with 4-fold change in translation efficiency. Barplots give average sensitivity from 100 simulations and error bars represent the standard deviation.

simulated data-set, roughly 19% genes are selected as truly differentially translated and the total number of genes is 9911. The identity of truly differentially translated genes is indicated within the file provided for

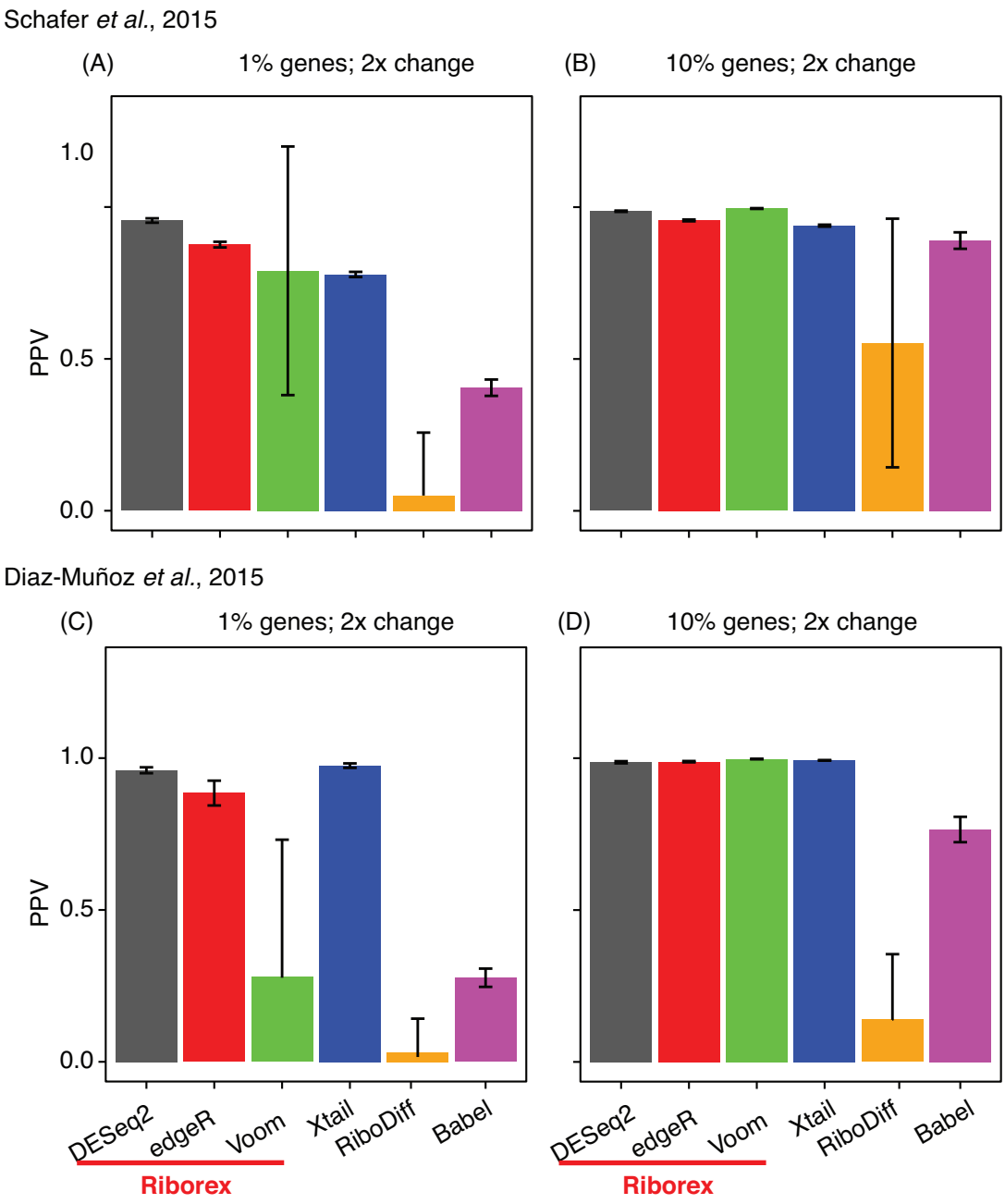

Figure S5: **Positive predictive value for different methods identifying implanted 2-fold change to translation efficiency.** Different numbers of differentially translated genes were implanted in the rat heart (panels A and B) and mouse HuR data sets (panels C and D). (A) and (C): 1% of genes were implanted with 2-fold change in translation efficiency. (B) and (D): 10% of genes were implanted with 2-fold change in translation efficiency. Barplots give average PPV from 100 simulations and error bars represent the standard deviation.

download by Xiao *et al.*, 2016. ROC curves are plotted to show the general performance of each method, as was done by Xiao *et al.*, 2016. From Figure S7, we can see Riborex, Xtail and RiboDiff all have extremely

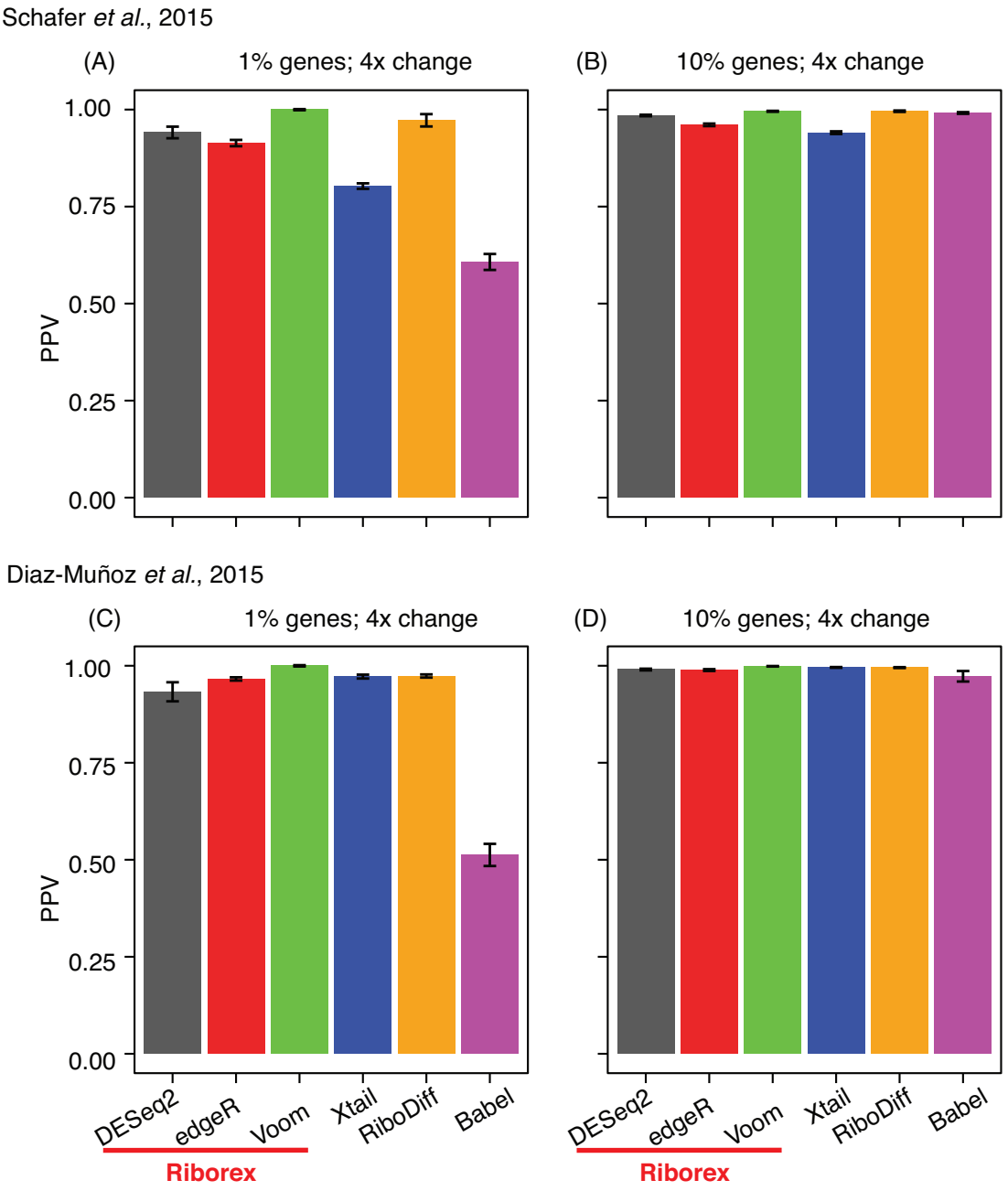

Figure S6: **Positive predictive value for different methods identifying implanted 4-fold change to translation efficiency.** Different numbers of differentially translated genes were implanted in the rat heart (panels A and B) and mouse HuR data sets (panels C and D). (A) and (C): 1% of genes were implanted with 4-fold change in translation efficiency. (B) and (D): 10% of genes were implanted with 4-fold change in translation efficiency. Barplots give average PPV from 100 simulations and error bars represent the standard deviation.

similar performance while Babel is less accurate. This simulation setting is most similar to our scenario #3, and shows consistent results. Our application of RiboDiff to this data shows significantly better accuracy

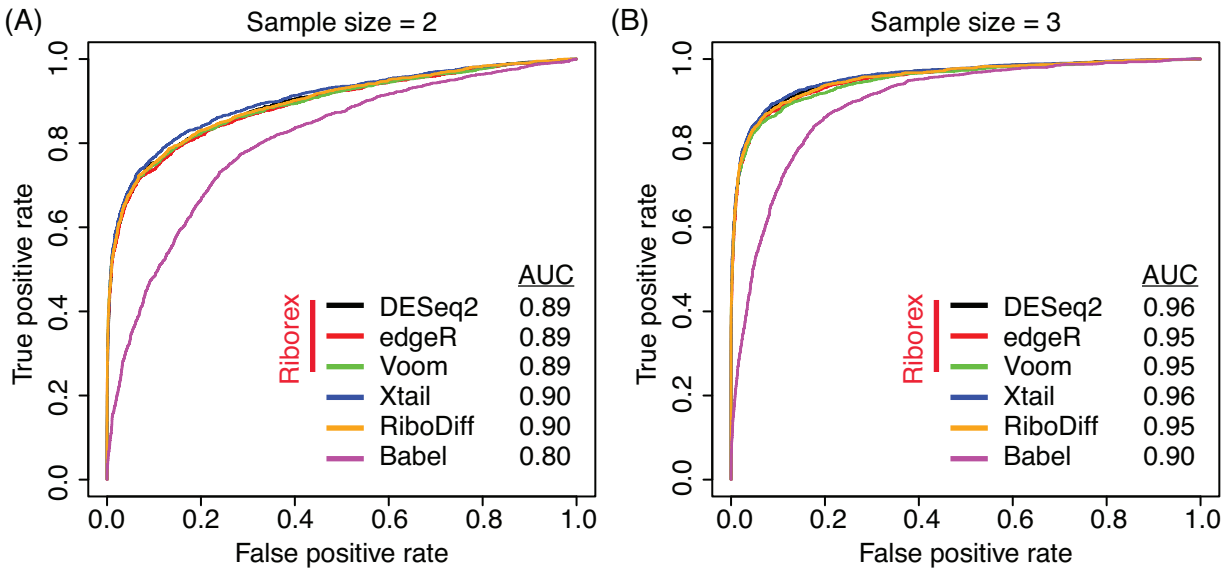

Figure S7: **ROC curves for different methods applied to semi-simulated data of Xiao et al. (2016).** This data set is based on the mouse HuR data with 19% implanted true differentially translated genes. Riborex and existing methods were applied on two simulated ribosome profiling data with 2 (A) and 3 (B) samples in each condition. ROC curves were plotted to show the general performance of different methods. This simulated data set was generated by Xiao et al. (2016) and used for comparing existing differential translation analysis tools.

than observed in the tests of Xiao *et al.*, 2016 on the same data set, with RiboDiff now showing comparable results to Xtail and Riborex. We attribute this change to our use of a more recent version of RiboDiff. We used R package ROCR (Sing *et al.*, 2005) for plotting those ROC curves. For results from all methods, we substituted the p-values of "nan" or "NA" with 1.

In summary, Riborex using DESeq2 engine has accuracy that is consistently very similar to that of Xtail under a variety of scenarios. RiboDiff and Babel seem to exhibit problems in certain important scenarios. Based on these results, at present we suggest users first consider using the DESeq2 engine unless additional information is available. These recommendations will be maintained upon releases as new versions of the engines are accommodated. At present the Voom engine seems to have less value overall compared with DESeq2 and edgeR, but the principles within Voom are sufficiently different from DESeq2 and edgeR that we anticipate specific use-cases emerging as it matures.

Applications on published data sets: We applied Riborex (DESeq2, edgeR, Voom as engines) and the other existing methods on a published data-set that studies human PC3 cells in response to mTOR signaling perturbation (Hsieh *et al.*, 2012). We will call this data-set PC3 data-set hereafter. Since mTOR is a major regulator of protein synthesis, the treatment of mTOR ATP inhibitor PP242 is expected to lead to substantial translational repression. Among the top 100 differentially translated genes detected, Riborex with DESeq2 or edgeR as engine detects only 1 up-regulated gene. The other 99 are exclusively down-regulated. Xtail also only detects 1 up-regulated gene. With Voom as the engine, Riborex detects 2 up-regulated genes among the top 100 differentially translated genes. RiboDiff and Babel detect 6 and 10 up-regulated genes, respectively. The overlap among the most significant genes detected by different methods is summarized

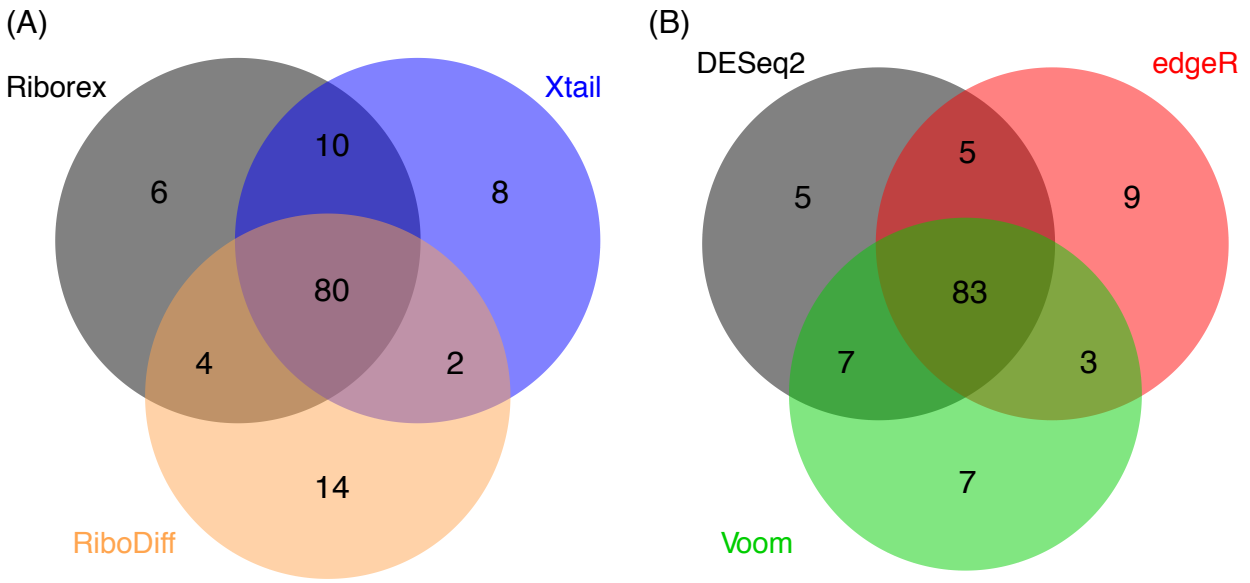

Figure S8: **Overlap of top differentially translated genes identified by different in the PC3 data.** (A) Among the top 100 differentially translated genes in PC3 data detected by Riborex (DESeq2 engine), RiboDiff and Xtail, the intersection contains 80 genes. (B) When Riborex was applied using different engines {DESeq2, edgeR, Voom} to identify the top 100 differentially translated genes, the intersection has size 83.

in Venn diagrams shown in Figure S8. Babel was not included in the Venn diagram, as it is an outlier with only $\sim$ 50 overlap relative to the other methods. In general, these results point to generally high overlap between all methods when applied to this data set, which may be explained by the expected high magnitude and breadth of translational response to perturbation mTOR.

Additional details on run time comparisons: We applied Riborex (using three engines), RiboDiff, Xtail and Babel on four published data-sets. The MSI2 data-set comes from a study of Musashi-2 (Msi2) RNA-binding protein which explored how Msi2 recognizes and regulates mRNA targets in mouse epithelial progenitor cells (Bennett *et al.*, 2016). The MSI2 data-set has 11004 genes with non-zero sum of read counts across samples, and 3 replicates in each condition. The Hela S3 data-set comes from a study that detects differentially translated genes between G1 and M phase of human cells during the cell-cycle and based on this analysis to further compare Ribo-seq with PUNCH-P which is another approach for genome-wide identification of protein synthesized at a given time (Zur *et al.*, 2016). The Hela S3 data-set has 13187 genes with non-zero sum of read counts across samples and 4 replicates in each condition. The mTOR data-set (Hsieh *et al.*, 2012) and the Liver data-set (Schafer *et al.*, 2015) have been introduced in previous sections. The mTOR data-set has 11392 genes with non-zero sum of read counts across samples and 2 replicates in each condition and the Liver data-set has 14032 genes with non-zero sum of read counts across samples and 5 replicates in each condition. The benchmark comparison was performed on a server with Intel Xeon CPU X5650@2.67GHz, 24GB of RAM and a single core was used for each method. The "wall time" (i.e. the total run time for completing the analysis) was recorded in Table 1.

References

Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biology,* **11** (10), R106.

Bennett,C.G. *et al.* (2016) Genome-wide analysis of Musashi-2 targets reveals novel functions in governing epithelial cell migration. *Nucleic Acids Research,* **1**, gkw207.

Diament,A. and Tuller,T. (2016) Estimation of ribosome profiling performance and reproducibility at various levels of resolution. *Biology direct,* **11**, 24.

Diaz-Muñoz,M.D. *et al.* (2015) The RNA-binding protein HuR is essential for the B cell antibody response. *Nature Immunology,* **16** (4), 415–425.

Hardcastle,T.J. and Kelly,K.A. (2010) baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics,* **11** (1), 422.

Hsieh,A.C. *et al.* (2012) The translational landscape of mTOR signalling steers cancer initiation and metastasis. *Nature,* **485** (7396), 55–61.

Law,C.W. *et al.* (2014) Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology,* **15** (2), R29.

Love,M.I. *et al.* (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology,* **15** (12), 1–21.

McCarthy,D.J. *et al.* (2012) Differential expression analysis of multifactor rna-seq experiments with respect to biological variation. *Nucleic acids research,* **1**, gks042.

Olshen,A.B., Hsieh,A.C., Stumpf,C.R., Olshen,R.A., Ruggero,D. and Taylor,B.S. (2013) Assessing gene-level translational control from ribosome profiling. *Bioinformatics,* **29** (23), 2995–3002.

Ritchie,M.E. *et al.* (2015) limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research,* **1**, gkv007.

Robinson,M.D. and Oshlack,A. (2010) A scaling normalization method for differential expression analysis of rna-seq data. *Genome biology,* **11** (3), 1.

Robinson,M.D. *et al.* (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics,* **26** (1), 139–140.

Schafer,S. *et al.* (2015) Translational regulation shapes the molecular landscape of complex disease phenotypes. *Nature Communications,* **6** (7200).

Sing,T., Sander,O., Beerenwinkel,N. and Lengauer,T. (2005) Rocr: visualizing classifier performance in r. *Bioinformatics,* **21** (20), 7881.

Xiao,Z. *et al.* (2016) Genome-wide assessment of differential translations with ribosome profiling data. *Nature Communications,* **7** (11194).

Zhong,Y. *et al.* (2016) RiboDiff: Detecting Changes of Translation Efficiency from Ribosome Footprints. *bioRxiv,* **1**, 017111.

Zur,H. *et al.* (2016) Complementary post transcriptional regulatory information is detected by punch-p and ribosome profiling. *Scientific reports,* **6**.