Generally applicable transcriptome-wide analysis of translational efficiency using anota2seq

Christian Oertlin, Julie Lorent, Ola Larsson *

*ola.larsson@ki.se

April 30, 2018

Contents

Introduction

Gene expression is a multi-step process including transcription, mRNA-transport, -stability and -translation. Dysregulated mRNA translation is commonly observed in human diseases such as cancer and understanding which mRNAs are differentially translated and the mechanisms that mediate such effects is therefore of high importance. Estimates of transcriptome-wide translational efficiency can be obtained using polysome-profiling and ribosome-profiling. Both approaches are based on isolation of translated mRNA (polysome-associated mRNA or Ribosome Protected Fragments [RPF]) followed by quantification using DNA-microarrays or RNA sequencing (RNAseq). A parallel total mRNA sample is also isolated and quantified in order to identify *bona fide* changes in translational efficiency. More details are found in [1, 2].

During analysis of the resulting data, three regulatory modes can be observed: changes in mRNA abundance (i.e. similar changes in total mRNA levels and levels of translated mRNA) and changes in translational efficiency leading to changes in protein levels (a change in the amount of translated mRNA that is not explained by a change in total mRNA) or buffering which maintains constant levels of translated mRNA (and hence also protein levels) despite altered levels of total mRNA. Efficient separation of these regulatory modes is necessary to elucidate underlying regulatory mechanisms [3]. Studies of changes in translational efficiency commonly apply per sample differences (log scale) between levels of translated mRNA and total mRNA [1] that are compared between treatments. However, as discussed in [4] such translational efficiency scores and outputs from methods that use such scores will show spurious correlations leading to elevated false positive findings [4].

This bias from spurious correlations can be solved by using per-identifier regression-based analysis between levels of translated mRNA and total mRNA. Such analysis produces residuals that are uncorrelated with the total mRNA levels and changes in translational efficiency leading to altered protein levels or buffering can be identified using Analysis of Partial Variance (APV) [4]. Anota2seq allows for identification of all three regulatory modes from polysome- or ribosome- profiling data quantified by DNA-microarrays or RNAseq. It uses APV and thereby eliminates spurious correlation bias. Here we illustrate the use of the anota2seq package.

1 Workflow

Analysis of translational activity using anota2seq includes the following steps:

1. Initialize an Anota2seqDataSet and pre-process RNA sequencing data using `anota2seqDataSetFromMatrix` or `anota2seqDataSetFromSE`. See section 3.2
2. Assessment of model assumptions using `anota2seqPerformQC` and `anota2seqResidOutlierTest`. See section 3.3
3. Analysis of changes in mRNA abundance and translational efficiency leading to altered protein levels or buffering using `anota2seqAnalyze`. See section 3.4
4. Selection of identifiers and classification into different regulatory modes of gene expression using `anota2seqSelSigGenes` and `anota2seqRegModes`. See section 3.4.4 and 3.5 respectively.

5. Visualize the results using `anota2seqPlotPvalues`, `anota2seqPlotFC` and `anota2seqPlotGenes` See sections 3.5.1, 3.4.2 and 3.4.4, respectively.

2 Getting started

anota2seq provides a wrapper function called `anota2seqRun` which performs all analysis steps with relevant default parameters. Here we show an overview of the whole workflow using this function. We illustrate an analysis using count data from a RNA sequencing experiment on total mRNA (called anota2seq_data_T here) and on translated mRNA (polysome-associated mRNA or Ribosome Protected Fragments, data called anota2seq_data_P) on at least 2 conditions ("ctrl" and "treatment" here) and a vector of sample annotation (called anota2seq_pheno_vec). The following code performs normalization, assesses model assumptions and performs the analysis for the default contrast (treatment vs. control in this case):

```
library(anota2seq)
data(anota2seq_data)
```

```
ads <- anota2seqDataSetFromMatrix(
    dataP = anota2seq_data_P[1:1000,],
    dataT = anota2seq_data_T[1:1000,],
    phenoVec = anota2seq_pheno_vec,
    dataType = "RNAseq",
    normalize = TRUE)
ads <- anota2seqRun(ads)
```

The regulatory modes can be quickly visualized:

```
anota2seqPlotFC(ads, selContrast = 1, plotToFile = FALSE)
```

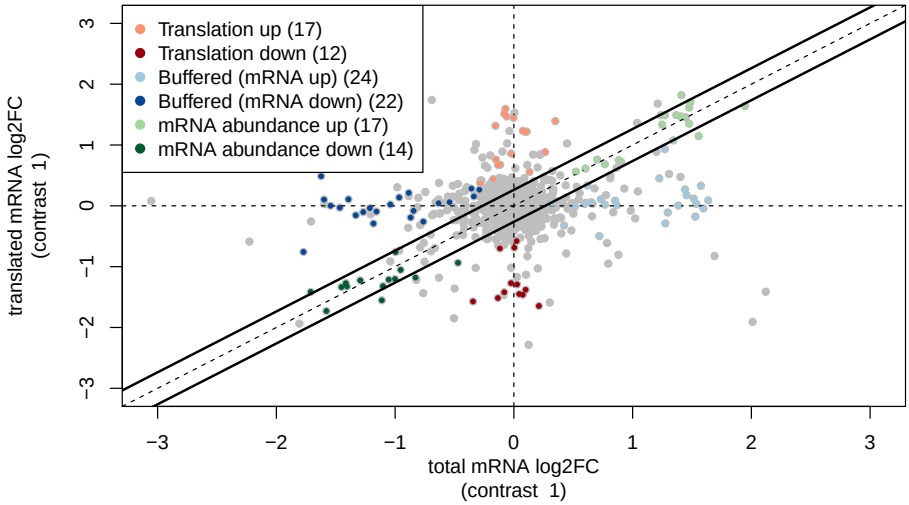

Figure 1: Visualization of the different regulatory modes

The following code chunk illustrates how to access the top list of significant changes in translational efficiency leading to altered protein levels (effect, adjusted p-value, regulatory mode):

```
head(
    anota2seqGetOutput(
        ads, analysis = "translation", output = "selected", getRVM = TRUE,
        selContrast = 1)[, c("apvEff", "apvRvmPAdj", "singleRegMode")])
```

```
##                 apvEff  apvRvmPAdj singleRegMode
## simGenes_464  1.476348 0.002236209   translation
## simGenes_819 -1.481481 0.002236209   translation
## simGenes_723  1.447539 0.004926436   translation
## simGenes_713  1.615780 0.004926436   translation
## simGenes_201 -1.525266 0.004926436   translation
## simGenes_480  1.182835 0.004926436   translation
```

This provided an overview of the features of the package. Each step of the analysis are detailed in the next section.

3 Transcriptome-wide analysis of translational efficiency using anota2seq

3.1 Input Data

anota2seq can analyze data from both ribosome-profiling and polysome-profiling quantified by RNAseq or DNA-microarrays. anota2seq cannot use data from competitive two channel experiments when the polysome- associated mRNA is directly compared to total mRNA as these do not allow independent estimates of polysome- associated mRNA and total mRNA levels[1]. anota2seq requires 3 replicate experiments per group if there is 2 treatments. If there is more than two treatments, two replicates is sufficient but will result in reduced statistical power as compared to three replicates. We recommend three replicates in most cases.

[1] A two-channel reference design should be applicable although we have not tested this data type.

In this vignette, we will use simulated data provided with the package to illustrate how to perform each step of the analysis. These data originate from the study by Oertlin et al. [3] which compared methods for analysis of translatomes quantified by RNAseq. Eight samples were simulated from 2 sample classes ("control" and "treatment"); both total mRNA (anota2seq_data_T, raw RNAseq counts) and paired translated mRNA (anota2seq_data_P, raw RNAseq counts) are provided together with a sample class vector (anota2seq_pheno_vec).

```
data(anota2seq_data)
# Polysome-associated mRNA and total mRNA columns must follow the same order
head(anota2seq_data_P, n = 2)
```

```
##            control_1_poly control_2_poly control_3_poly control_4_poly
## simGenes_1          11819          11262          10267          10757
## simGenes_2           3668           4573           2207           8745
##            treatment_1_poly treatment_2_poly treatment_3_poly
## simGenes_1             9448             9574            11964
## simGenes_2             3058             2970             8500
##            treatment_4_poly
## simGenes_1            11202
## simGenes_2             4695
```

```
head(anota2seq_data_T, n = 2)
```

```
##            control_1_cyto control_2_cyto control_3_cyto control_4_cyto
## simGenes_1           6612           6832           6041           5920
## simGenes_2           1054           1012            486            597
##            treatment_1_cyto treatment_2_cyto treatment_3_cyto
## simGenes_1             3823             6880             3374
## simGenes_2              915              661              372
##            treatment_4_cyto
## simGenes_1             5934
## simGenes_2             1018
```

```
# phenoVec must describe the sample class for corresponding columns
# in dataT and dataP
anota2seq_pheno_vec
```

```
## [1] "ctrl"      "ctrl"      "ctrl"      "ctrl"      "treatment" "treatment"
```

```
## [7] "treatment" "treatment"
```

3.2 Normalization and transformation of the raw data

The anota2seq performance will vary depending on normalization and transformation of the data. We therefore recommend that the user tries several different transformations and normalization approaches while monitoring the quality control plots (the influential data points, the interactions and the normality of the residuals) and the RVM F-distribution fit plot if RVM is used (see sections 3.3 and 3.4.1).

anota2seq gives the options to supply normalized DNA-microarrays data, normalized and transformed RNAseq data or raw RNAseq data for both translated mRNA (i.e. polysome-associated mRNA or RPF) and total mRNA. As anota2seq requires data on a continuous log scale, raw RNAseq data (count data) will be pre-processed to ensure efficient analysis.

In general, RMA is an efficient normalization for DNA-microarray data (for Affymetrix GeneChips) while TMM-log2 normalization [5, 6] is efficient for RNAseq data (this is the default method in anota2seq when raw RNAseq data is provided as input). The rlog algorithm from *DESeq2* [7] can also be used within anota2seq.

Normalization and transformation of RNAseq data are performed at the step of initialization of an *Anota2seqDataSet* object in the `anota2seqDataSetFromMatrix` and `anota2seqDataSetFromSE` functions. Filtering of identifiers with 0 counts in at least one sample is also available when raw RNAseq data is provided (parameter `filterZeroGenes`). Additionally, users can filter the dataset to remove identifiers with no variance in each mRNA source prior to analysis. This filtering prevents an APV analysis without variance which will result in an error and a halt in the analysis (parameter `varCutOff`).

```
ads <- anota2seqDataSetFromMatrix(
    dataP = anota2seq_data_P[1:1000,],
    dataT = anota2seq_data_T[1:1000,],
    phenoVec = anota2seq_pheno_vec,
    dataType = "RNAseq",
    filterZeroGenes = TRUE,
    normalize = TRUE,
    transformation = "TMM-log2",
    varCutOff = NULL)
```

Similarly, an *Anota2seqDataSet* object can be initialized from a *SummarizedExperiment* object using `anota2seqDataSetFromSE` as follows[2]:

```
adsFromSE <- anota2seqDataSetFromSE(
    se = mySummarizedExperiment,
    assayNum = 1, # Position of the count data in assays(mySummarizedExperiment)
    dataType = "RNAseq",
    normalize = TRUE,
    transformation = "TMM-log2")
```

[2] see `help(anota2seqDataSetFromSE)` for details on required `colData` formatting

3.3 Assessment of model assumptions

To apply APV, multiple assumptions need to be fulfilled for tens of thousands of identifiers, which is a substantial challenge to evaluate. Due to the high dimensionality of the data, anota2seq takes multiple testing into account when assessing assumption violations. If we observe the same number of problematic features as expected, we assume that we can apply anota2seq.

Using the following code, anota2seq performs quality control checks and outputs diagnostic plots (Fig. 2 to 4) which are further described below.

```
ads <- anota2seqPerformQC(Anota2seqDataSet = ads,
                          generateSingleGenePlots = TRUE)
```

Highly influential data points may cause errors in the regression analyzes. On the one hand, we expect that a number of highly influential data points will appear merely by chance because of the large number of analyzes performed. Thus anota2seq attempts to establish if we, when considering all analyzed identifiers, observe more influential data points compared to what would be expected by chance. If the answer is no, then there are no concerns with the overall analysis. On the other hand, influential data points may nonetheless affect the specific APV analyzes in which they are found. For this reason, anota2seq provides an output (Fig. 2) that can be used to flag these identifiers so that they can be examined in more detail if desired.

For detection of influential data points, anota2seq uses standardized dfbeta for the slope of the regression and several thresholds to determine whether or not a data point is highly influential. As there is no known distribution of the dfbetas when the underlying data are normally distributed, anota2seq simulates data sets to obtain estimates of the expected number of outliers. The simulation is performed by sampling N (corresponding to the number of samples in the analysis) data points from the normal distribution and calling these data points the translated mRNA level. In detail, such translated mRNA levels are obtained by sampling data points from a normal distribution with a mean of the corresponding total mRNA level data point. Ten different such data sets are obtained with different variances when sampling translated mRNA level data. These data sets are then merged and frequencies of outlier dfbetas are calculated and compared to the frequencies of outlier dfbetas from the analyzed data (Fig. 3).

APV assumes that the slopes of the regressions from each treatment are the same so that using the common slope is valid. This assumption postulates that the relationship between the translated mRNA level and the total mRNA level shows the same slope for each treatment, i.e., treatment and total mRNA levels do not interact in predicting translated mRNA levels. Again, because we analyze thousands of regressions, we expect that a number of interactions will arise simply due to chance. If the number of interactions does not exceed what is expected by chance, their p-values should follow a uniform distribution. Thus anota2seq provides an output allowing to compare the distribution of the interaction significances as well as their distribution after adjusting for multiple testing (Fig. 4).

Significance testing within the APV framework assumes that the residuals from the regressions are normally distributed. The `anota2seqResidOutlierTest` function assesses whether the residuals from the linear regressions (identifier by identifier) of translated mRNA level total mRNA level are normally distributed.

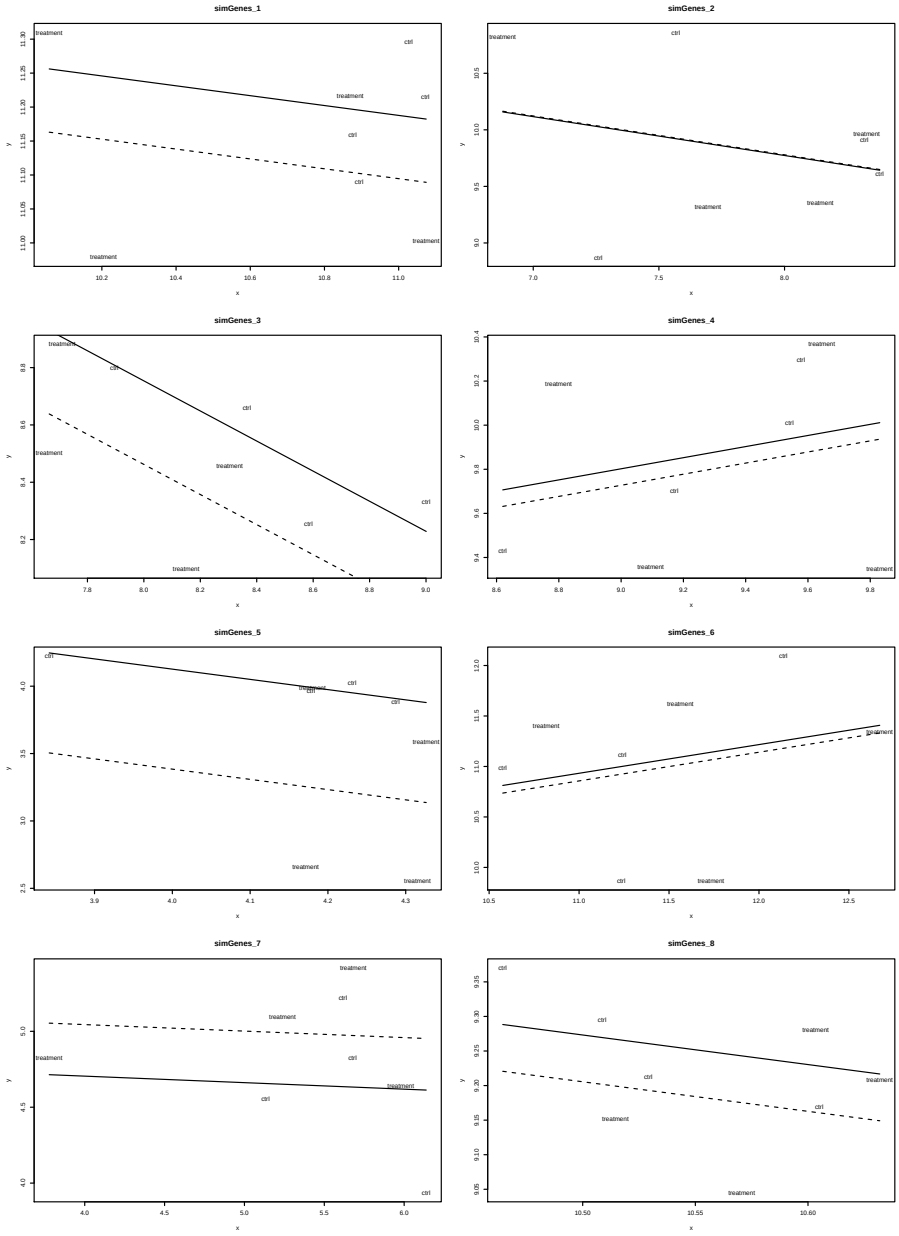

Figure 2: anota2seq can be set to output identifier per identifier regressions between translated mRNA and total mRNA levels
Plotting symbols are taken from the `phenoVec` argument and the lines are the regression lines per samples class

```
ads <- anota2seqResidOutlierTest(ads)
```

anota2seq generates normal Q-Q plots of the residuals. If the residuals are normally distributed, the data quantiles will form a straight diagonal line from bottom left to top right. Because there are typically relatively few data points, anota2seq calculates "envelopes" based on a set of samplings from the normal distribution using the same number of data points

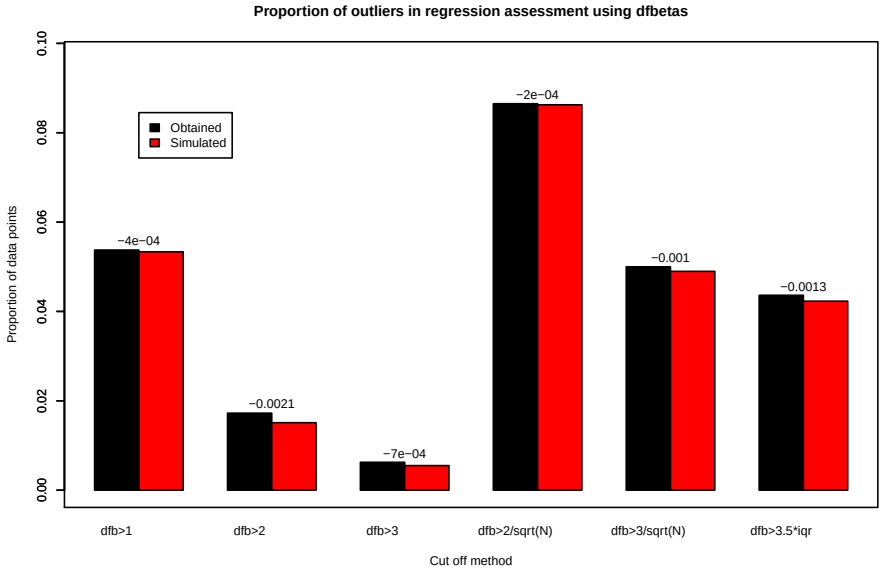

Figure 3: A bar graph showing the obtained and expected (based on a simulation) number of influential data points as judged by different thresholds
For each threshold the difference between the obtained and the simulated frequency of outliers is shown.

as for the true data [8]. To enable a comparison, both the true and the sampled data are scaled (variance=1) and centered (mean=0). The samples (both true and sampled) are then sorted and the true sample is compared to the envelopes of the sampled series at each sort position. The result is presented as a Q-Q plot of the true data where the envelopes of the sampled series are indicated. If there are 99 samplings we expect that 1/100 values should be outside the range obtained from the samplings. Thus it is possible to assess if approximately the expected number of outlier residuals are obtained. anota2seq provides a single identifier output (Fig. 5) as well as a summary output (Fig. 6).

While anota2seq enables testing of the issues discussed above, it is left to the user to decide whether it is possible to use anota2seq to identify changes in translational efficiency affecting protein levels or buffering. A few issues that may cause problems in the quality control are:

1. Outlier samples. One or a few outlier samples in the analysis (either from the translated mRNA data or the total mRNA data) could give rise to many influential data points. Thus, if there are more influential data points than would be expected, a careful quality control of the data followed by identification and exclusion of outlier samples might be needed to resolve such issues.
2. More significant interactions compared to what is expected by chance could be caused by bias in the data set.
3. If the resulting residuals deviate strongly from normality an alternative normalization method could be tested.

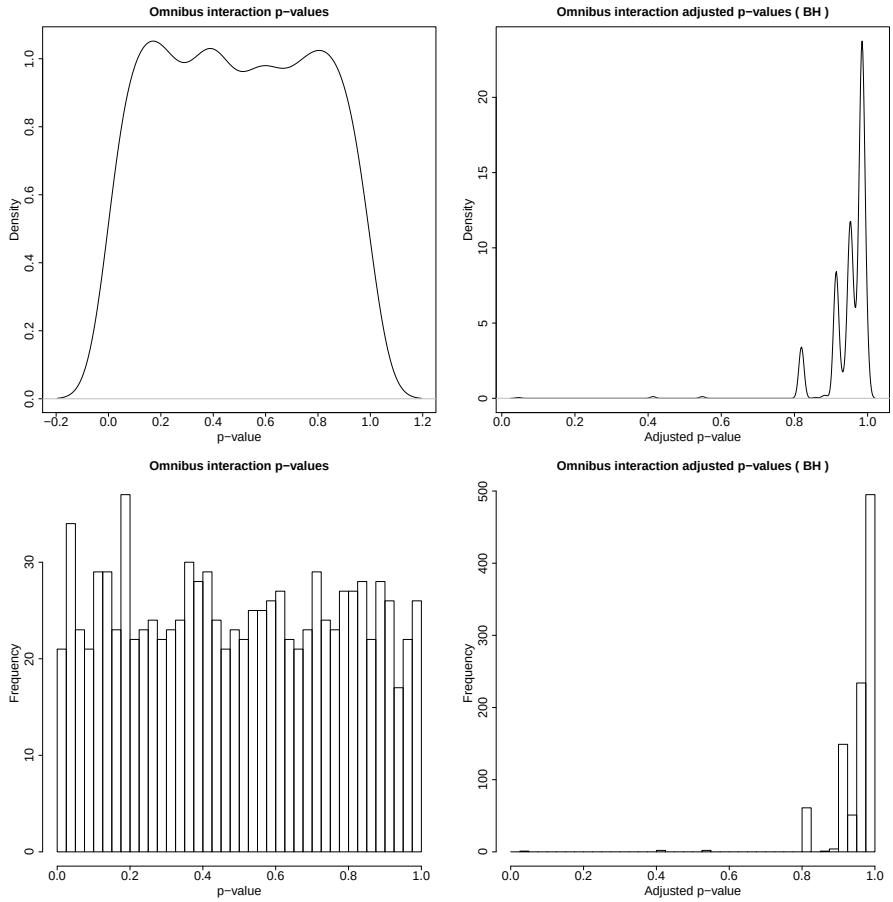

Figure 4: Assessment of whether the significances for the interactions follow the uniform NULL distribution
Shown are both density plots and histograms of the nominal and adjusted p-values (in this case adjusted using Benjamini-Hochberg FDR).

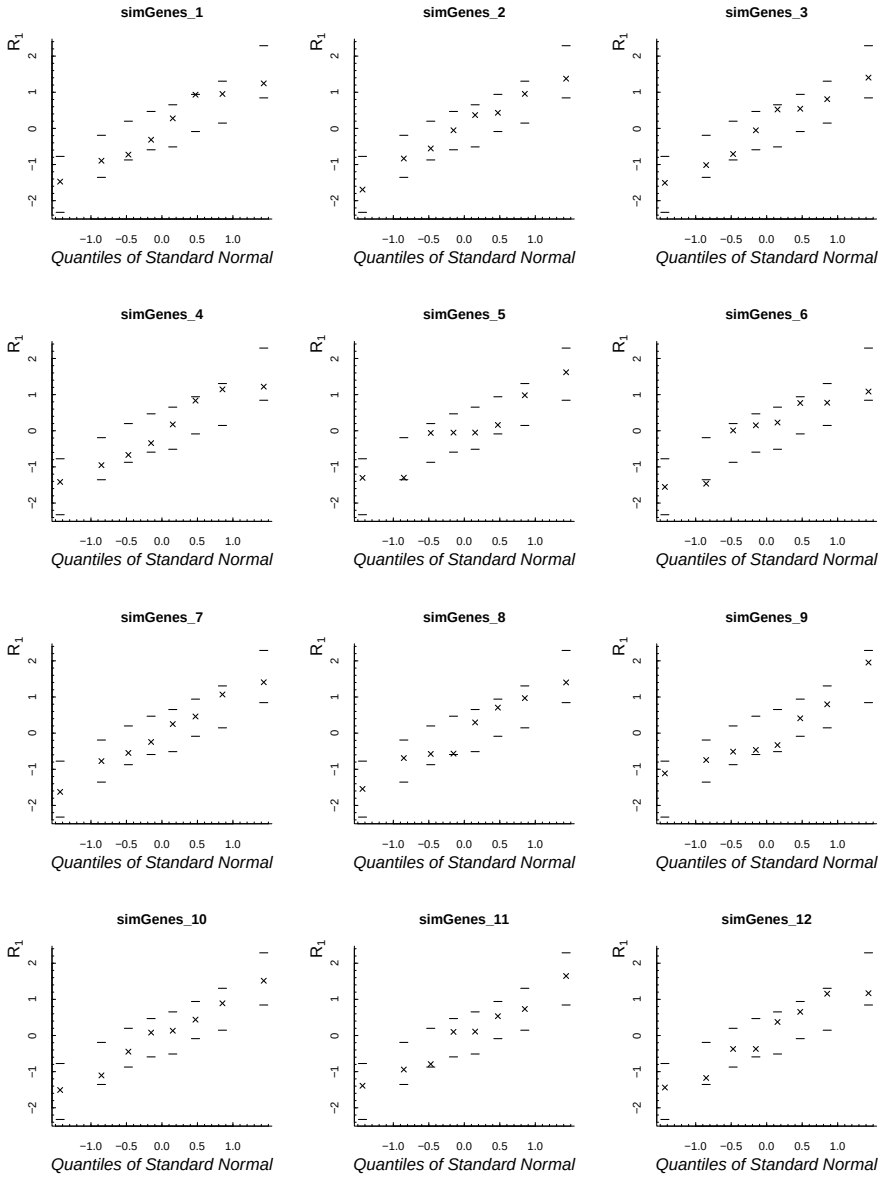

Figure 5: Assessment of whether the residuals are approximately normally distributed
Shown is the output from the single identifier alternative within `anota2seqResidOutlierTest`. The Q-Q plot for the identifier is compared to the outer limits of a set of Q-Q plots generated by sampling from the normal distribution.

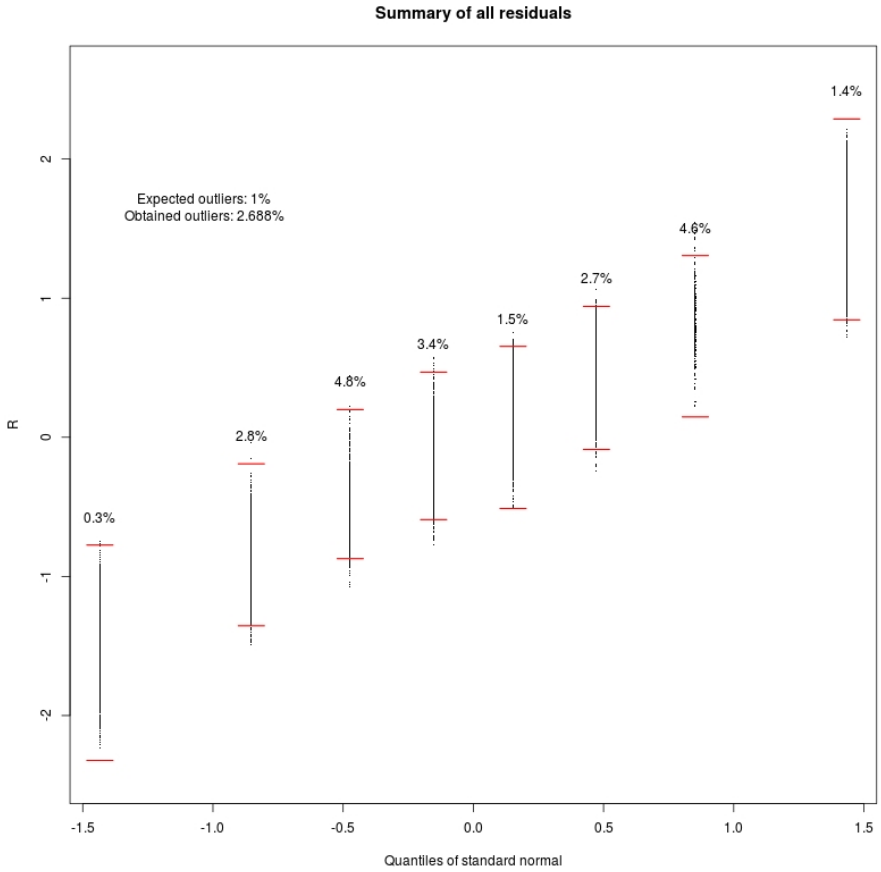

Figure 6: Assessment of whether the residuals are approximately normally distributed
Shown is the output from all identifiers using the `anota2seqResidOutlierTest` function. The Q-Q plot for the identifiers is compared to the outer limits of a set of Q-Q plots generated by sampling from the normal distribution. The obtained and expected percentage of outliers is indicated at each rank position and combined.

3.4 Analysis of changes in translational efficiency leading to altered protein levels or buffering

Once the data set has been validated as suitable for analysis, significant changes in translational efficiency affecting protein levels or buffering can be identified.

Translational buffering is a regulatory pattern which decouples mRNA levels from protein levels (despite altered levels of total mRNAs between conditions, translated mRNA levels remain constant; mRNAs under such a regulatory pattern are colored in dark and light blue in the example presented in Fig. 1) and which potentially holds important information regarding how gene expression is regulated. anota2seq allows to distinguish between changes in translational efficiency leading to altered protein levels (orange and red colored mRNAs in Fig. 1) and buffering. Both analyzes can be performed on our sample data using the following code:

```
ads <- anota2seqAnalyze(Anota2seqDataSet = ads,
                        analysis = c("translation", "buffering"))
```

While `anota2seqPerformQC` performs an omnibus treatment effect test when there are more than 2 treatments, `anota2seqAnalyze` allows the user to set custom contrasts using the `contrasts` parameter. In the example above, the default contrast ("treatment" vs. "control") is used.

3.4.1 Random variance model (RVM) to improve power in detection of changes in translational efficiency leading to altered protein levels or buffering

RVM is an empirical Bayes method which has been shown to increase statistical power for small N analysis [9]. In RVM, the variance of each identifier is adjusted using the variance obtained from an inverse gamma distribution derived from the variances of all identifiers. A key assumption in RVM is that the resulting variances follow a theoretical F-distribution. anota2seq tests this for the analysis of omnibus group effects (Fig. 7), omnibus interactions (not shown, output of `anota2seqPerformQC`), and the identification of changes in translational efficiency leading to altered protein levels and buffering (not shown, output of `anota2seqAnalyze`). Each of these analyzes generates a comparison of the obtained empirical distribution compared to the theoretical distribution (similarity is then assessed using a KS test whose alternative hypothesis should be rejected for a good fit). We have noticed that the normalization of the data can strongly influence the fit but that RVM seems to be applicable in most cases after identifying an efficient normalization/transformation. It is necessary to validate that application of RVM does not influence the distribution of the interaction p-values (not shown, output of `anota2seqPerformQC`). `anota2seqAnalyze` performs analyzes both with and without RVM ; we recommend using RVM as it improves the power to detect changes in translational efficiency leading to altered protein levels or buffering within anota2seq [4].

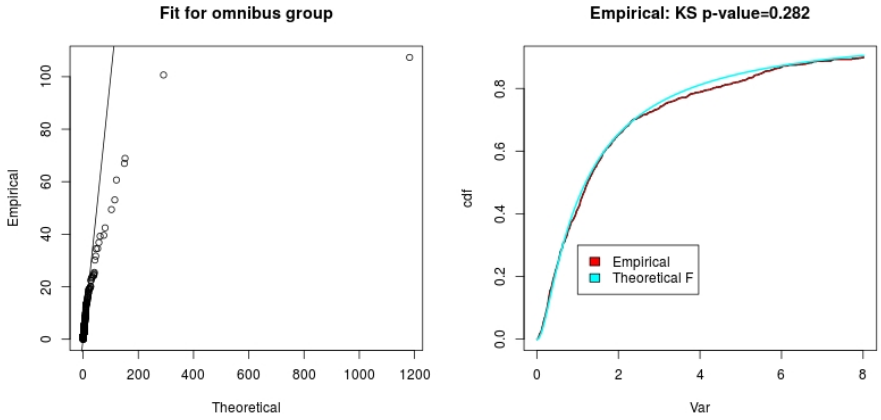

Figure 7: An output from the `anota2seqPerformQC` function (used with parameter `useRVM = TRUE`) comparing obtained variances to the theoretical F-distribution
RVM assumes that the empirical and the theoretical distributions are similar.

3.4.2 Visualization of the results from `anota2seqAnalyze`

`anota2seqAnalyze` outputs details of the tests for each identifier (information about slopes of the APV model, test statistics, effect, unadjusted and adjusted p-value):

```
head(anota2seqGetOutput(
    ads, analysis = "translation",
    output = "full",
    selContrast = 1,
    getRVM = TRUE))
```

```
##              apvSlope  apvSlopeP      apvEff apvRvmMSerror    apvRvmF
## simGenes_1 -0.07264945 0.33382264 -0.09313704    0.01820784 0.47641595
## simGenes_2 -0.34344681 0.27915442  0.00617025    0.25189874 0.00015114
## simGenes_3 -0.52515104 0.03676298 -0.29165877    0.03668045 2.31907866
## simGenes_4  0.25178079 1.00000000 -0.07486115    0.09942910 0.05636369
## simGenes_5 -0.75993233 0.30922928 -0.74205704    0.13453505 4.09297549
## simGenes_6  0.28424131 1.00000000 -0.07560520    0.34979134 0.01634159
##            residRvmDf    apvRvmP apvRvmPAdj
## simGenes_1   6.766864 0.51301921  0.9367994
## simGenes_2   6.766864 0.99054570  0.9971405
## simGenes_3   6.766864 0.17306945  0.7506550
## simGenes_4   6.766864 0.81936543  0.9806028
## simGenes_5   6.766864 0.08414756  0.6819360
## simGenes_6   6.766864 0.90199589  0.9933827
```

The density of p-values can be visualized using the `anota2seqPlotPvalues` function on the output of `anota2seqAnalyze` (Fig. 8).

```
par(mfrow = c(1, 2))
anota2seqPlotPvalues(ads, selContrast = 1, plotToFile = FALSE)
```

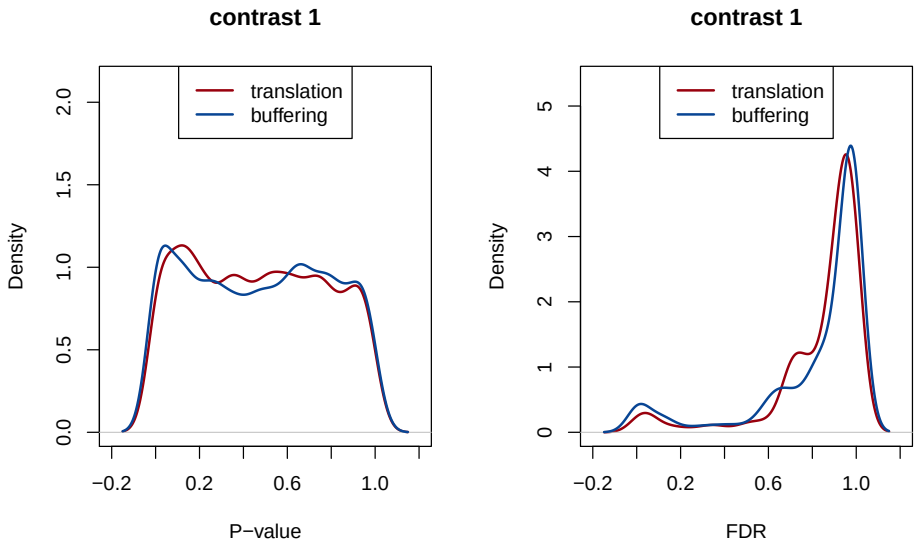

Figure 8: An output from the `anota2seqPlotPvalues` function
The left graph shows a P-value distribution of changes in translational efficiency leading to altered protein levels (designated "translation") and buffering for all analyzed identifiers. The right graph shows the corresponding adjusted P-value (FDR) distributions.

3.4.3 Unrealistic models of changes in translation efficiency

The slopes that are fitted in the anota2seq APV models can take unrealistic values that will influence the analysis of changes in translation efficiency leading to altered protein levels or buffering. anota2seq therefore tests whether slopes for analysis of changes in translational efficiency affecting protein levels that are >1 differ from 1 and slopes for analysis of changes in translational efficiency leading to buffering that are <-1 differ from -1. Furthermore, slopes < 0 for analysis of changes in translational efficiency affecting protein levels or > 0 for analysis of changes in translational efficiency leading to buffering, indicate unlikely but not impossible translational control so these events are also tested for. Results of these tests (p-values) are found in the output of `anota2seqPerformQC`, `anota2seqAnalyze` and `anota2seqRun` functions. These p-values can be used to filter or flag identifiers with unrealistic slopes or slopes revealing unlikely translational control.

3.4.4 Identifier selection and visualization of single gene regressions

The output from `anota2seqAnalyze` can be filtered using the `anota2seqSelSigGenes`. Identifiers can be selected based on several criteria:

- include only realistic slopes (see section 3.4.3), using parameters `minSlopeTranslation`, `maxSlopeTranslation`[3] and `slopeP`
- include a minimum effect threshold, using parameter `minEff`
- include only significant identifiers according to a defined p-value or adjusted p-value threshold (parameter `maxP` and `maxPAdj`)

[3] and similarly `minSlopeBuffering` and `maxSlopeBuffering`

An example of code to perform this filtering is as follows:

```
ads <- anota2seqSelSigGenes(Anota2seqDataSet = ads,
                            selContrast = 1,
                            minSlopeTranslation = -1,
                            maxSlopeTranslation = 2,
                            minSlopeBuffering = -2,
                            maxSlopeBuffering = 1,
                            maxPAdj = 0.05)
```

Once the `Anota2seqDataSet` object has been filtered, single gene regressions can be visualized using the `anota2seqPlotGenes` function (Fig. 9 and 10). The graphical output includes both the graphical interpretation of the APV analysis and the key statistics from both the standard and the RVM based analysis.

```
anota2seqPlotGenes(ads, selContrast = 1, analysis = "translation", plotToFile = FALSE)
```

```
anota2seqPlotGenes(ads, selContrast = 1, analysis = "buffering", plotToFile = FALSE)
```

3.4.5 Note about analysis of translational buffering

The APV model fitted in anota2seq for analysis of changes in translational efficiency leading to altered protein levels consists in a model with translated mRNA as independent variable and total mRNA and the sample class variable as dependent variables. In other words, a common slope for all sample categories is considered and the translational effect is defined as a difference in intercepts [4]. This regression model can be visualized in Fig. 9.

Translational buffering is defined as changes in total mRNA level that are not paralleled by changes in levels of translated mRNA. As such, performing analysis of buffering considers total mRNA as independent variable and translated mRNA as dependent variable (together with the sample class; as illustrated in Fig. 10).

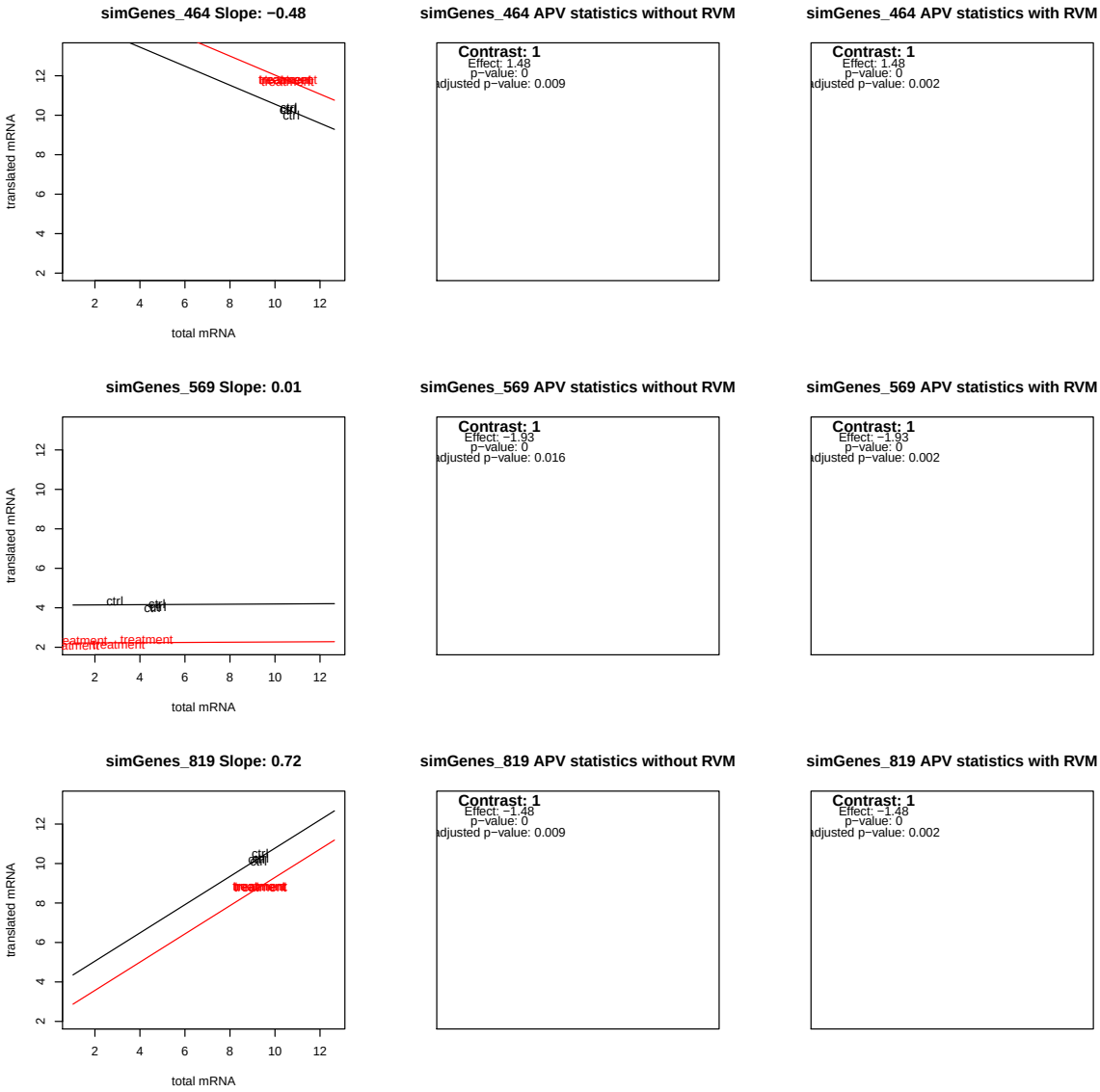

Figure 9: Visualization provided by the `anota2seqPlotGenes` function for analysis of changes in translational efficiency leading to altered protein levels

The left graph shows the identifier per identifier regressions between translated mRNA (in this case polysome-associated mRNA) and total mRNA levels. Plotting symbols are taken from the `phenoVec` argument supplied to the `anota2seqAnalyze` function and the lines are the regression lines per treatment using the common slope identified in APV (shown in the main title). The right and middle graphs show key statistics for the analyzed identifier with and without RVM, respectively. These graphs (shown here for only 3 identifiers) can be visualized for all identifiers selected in `anota2seqSelSigGenes`.

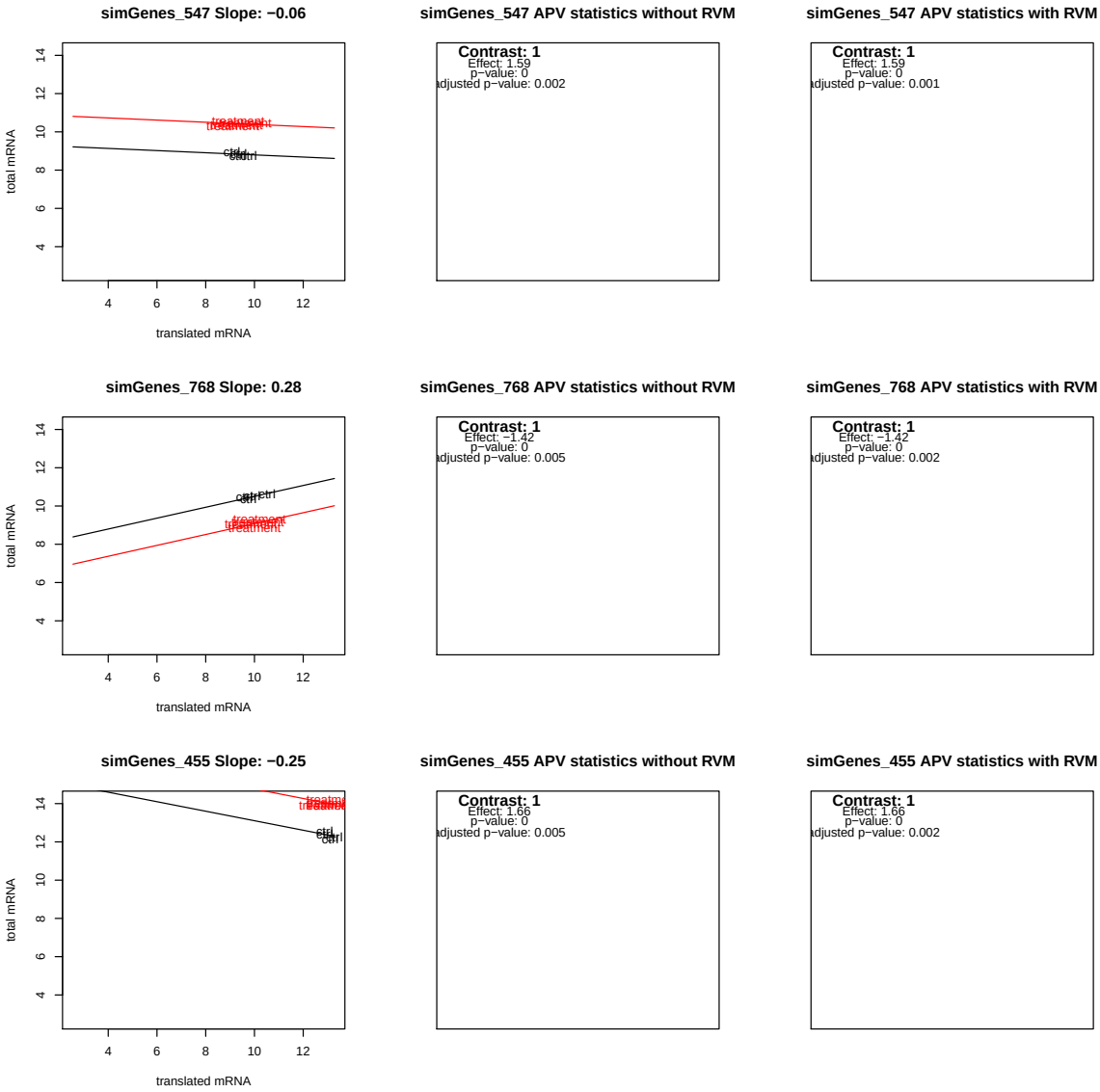

Figure 10: Visualization provided by the `anota2seqPlotGenes` function for analysis of changes in translational efficiency leading to buffering

The left graph shows the identifier per identifier regressions between total mRNA and translated mRNA (in this case polysome-associated mRNA) levels. Plotting symbols are taken from the `phenoVec` argument supplied to the `anota2seqAnalyze` function and the lines are the regression lines per treatment using the common slope identified in APV (shown in the main title). The right and middle graphs show key statistics for the analyzed gene with and without RVM respectively. These graphs (shown here for only 3 identifiers) can be visualized for all identifiers selected in `anota2seqSelSigGenes`.

3.5 Categorizing genes into regulatory modes

Polysome or ribosome profiling allows the user to distinguish between three regulatory modes: changes in mRNA abundance (i.e. similar changes in total mRNA levels and levels of translated mRNA) and translational efficiency leading to altered protein levels or buffering [3]. For that, the `anota2seqAnalyze` and `anota2seqSelSigGenes` functions have to be run with parameter `analysis` set to "translation" and "buffering" as shown above but analysis of differential expression of total mRNA and translated mRNA is also required. For that, the same functions can be used with `analysis` parameter set to "total mRNA" and "translated mRNA" as shown below:

```
ads <- anota2seqAnalyze(Anota2seqDataSet = ads,
                        analysis = c("total mRNA", "translated mRNA"))
ads <- anota2seqSelSigGenes(Anota2seqDataSet = ads,
                            analysis = c("total mRNA", "translated mRNA"),
                            selContrast = 1,
                            minSlopeTranslation = -1,
                            maxSlopeTranslation = 2,
                            minSlopeBuffering = -2,
                            maxSlopeBuffering = 1,
                            maxPAdj = 0.05)
```

Once all analyzes have been performed, all regulated identifiers can be categorized into one of these regulatory modes using the `anota2seqRegModes` function. This categorization into gene expression regulatory patterns might be of interest in order to elucidate underlying mechanisms.

```
ads <- anota2seqRegModes(ads)
```

Notably, there is a hierarchy such that mRNAs identified as changing their translational efficiency leading to altered protein levels will belong to the translation group and no other group; mRNAs that change their levels in the translated pool and total mRNA pool but are not identified as changing their translational efficiency leading to altered protein levels will be in the abundance group; and mRNAs that are identified as changing their translational efficiency leading to buffering and are not in the former two groups are allocated to the set of buffered mRNAs. Specifically, the `anota2seqRegModes` function adds a column named "singleRegModes" indicating the classification into regulatory modes in the data.frame containing gene by gene statistical results. This output can be accessed by using `anota2seqGetOutput` with `output` parameter set to "regModes".

```
head(anota2seqGetOutput(object = ads, output="regModes",
                        selContrast = 1, analysis="buffering",
                        getRVM = TRUE))[, c("apvSlope", "apvEff", "apvRvmP",
                                            "apvRvmPAdj", "singleRegMode")]

##                apvSlope     apvEff      apvRvmP  apvRvmPAdj singleRegMode
## simGenes_547 -0.05644675  1.5938688 1.420355e-06 0.001420355     buffering
## simGenes_768  0.28453551 -1.4234012 6.122129e-06 0.001748704     buffering
## simGenes_455 -0.24850186  1.6592623 6.344953e-06 0.001748704     buffering
## simGenes_858 -0.04197019 -0.8762414 9.543941e-06 0.001748704     buffering
## simGenes_318 -0.31283262 -1.3006886 1.205665e-05 0.001748704     buffering
## simGenes_431 -0.18482654 -1.3602127 1.224093e-05 0.001748704     buffering
```

3.5.1 Visualizing the different regulatory modes

anota2seq provides the `anota2seqPlotFC` function which plots the translated mRNA log Fold Change vs. the total mRNA log Fold Change and colors genes according to their regulatory mode (Fig. 11).

```
anota2seqPlotFC(ads, selContrast = 1, plotToFile = FALSE)
```

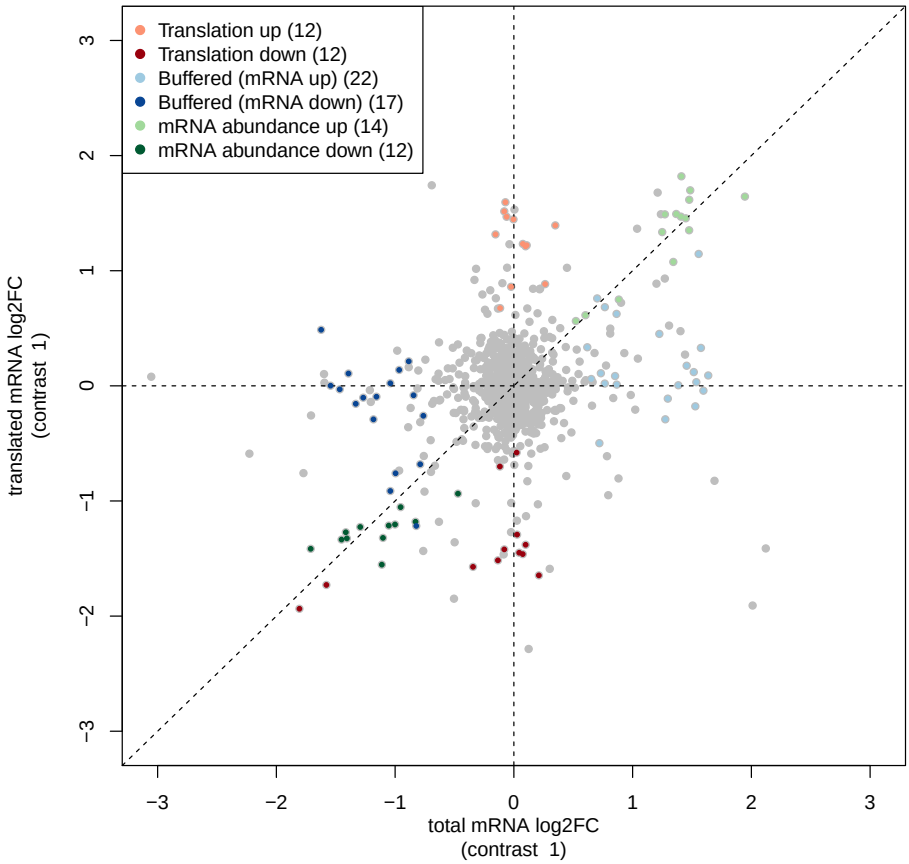

Figure 11: An output from the `anota2seqPlotFC` function
Shown is a scatter-plot (for all included identifiers) of fold-changes (treatment vs. control) for translated mRNA (in this case polysome-associated mRNA) and total mRNA. Identifiers filtered using the `anota2seqSelSigGenes` function have been categorized as either showing changes in abundance or changes in translational efficiency leading to altered protein levels (indicated as "translation" in the graph) or buffering ; these are indicated by colors.

3.6 One-step procedure with `anota2seqRun`

In addition to application of each of the functions within anota2seq which provides the maximum flexibility, the anota2seq package provides the option to perform a one-step analysis of translated mRNA, total mRNA and changes in translational efficiency leading to altered protein levels or buffering. This analysis performs quality control followed by analysis of changes in translational efficiency affecting protein levels or buffering. A filtering is also performed (as in `anota2seqSelSigGenes`) as well as categorization into regulatory modes.

```
ads <- anota2seqRun(
    Anota2seqDataSet = ads,
    thresholds = list(
        maxPAdj = 0.05,
        minEff = 1.5),
    performQC = TRUE,
    performROT = TRUE,
    useRVM = TRUE)
```

The output of the `anota2seqRun` function can be supplied to the `anota2seqPlotPvalues`, `anota2seqPlotGenes` and `anota2seqPlotFC` functions for similar visualization of the results as in Fig. 8, 9, 10 and 11.

4 Extending anota2seq to analysis of other data sources

In principle, any data source where the intention is to identify changes in a subset that is independent of a background can be analyzed (e.g. RIP-SEQ data).

5 New features in anota2seq compared to *anota*

The core models in anota2seq are similar to those in the *Bioconductor* package *anota*. However, there are many differences including:

- *anota* was designed to analyze data from DNA-microarrays platforms. anota2seq allows analysis of both DNA-microarrays and RNA sequencing (section 3.1)
- Implementation of analysis of translational buffering (section 3.4)
- anota2seq allows for batch adjustment (parameter `batchVec` of `anota2seqDataSetFromMatrix` and `anota2seqDataSetFromSE`)
- anota2seq provides additional functions in order to easily and consistently visualize the results of analyzes: `anota2seqPlotPvalues` (section 3.4.2) and `anota2seqPlotFC` (section 3.5.1)
- anota2seq provides a wrapper function which performs all steps of the workflow (section 3.6)
- anota2seq provides a classification of mRNAs into different gene expression regulatory modes: changes in mRNA abundance, or translational efficiency leading to altered protein levels or buffering (section 3.5)

References

[1] Ola Larsson and Robert Nadon. Gene expression - time to change point of view? *Biotechnology & Genetic Engineering Reviews*, 25:77–92, 2008.

[2] Ciriaco A. Piccirillo, Eva Bjur, Ivan Topisirovic, Nahum Sonenberg, and Ola Larsson. Translational control of immune responses: from transcripts to translatomes. *Nature Immunology*, 15(6):503–511, June 2014.

[3] Christian Oertlin, Julie Lorent, Valentina Gandin, Carl Murie, Laia Masvidal, Marie Cargnello, Luc Furic, Ivan Topisirovic, and Ola Larsson. Genome-wide analysis of differential translation and differential translational buffering using anota2seq. *bioRxiv*, 2017. URL: http://www.biorxiv.org/content/early/2017/02/08/106922.

[4] Ola Larsson, Nahum Sonenberg, and Robert Nadon. Identification of differential translation in genome wide studies. *Proceedings of the National Academy of Sciences of the United States of America*, 107(50):21487–21492, December 2010.

[5] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, 26(1):139–140, January 2010.

[6] Matthew E. Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47, April 2015.

[7] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, 2014.

[8] W.N. Venables and B.D. Ripley. *Modern Applied Statistics with S-PLUS*. Springer, 1999.

[9] George W. Wright and Richard M. Simon. A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics (Oxford, England)*, 19(18):2448–2455, December 2003.