ANalysis Of Translational Activity (anota)

Ola Larsson <ola.larsson@ki.se>,
Nahum Sonenberg <nahum.sonenberg@mcgill.ca>,
Robert Nadon <robert.nadon@mcgill.ca>

April 24, 2017

Contents

1 Introduction

Translational control of gene expression is a mechanism that affects the relationship between mRNA level and protein level. Translation is commonly deregulated in human diseases such as cancer and understanding which mRNAs are targeted for translational deregulation and the mechanisms that mediate such effects is of high importance. Estimates of global translational activity has primarily been generated using the poly(ribo)some microarray approach (reviewed in [1]) but sequencing based methods have also recently been developed [2]. Both approaches are based on a parallel quantification of cytosolic mRNA level and the level of those mRNAs that are actively utilized for translation. During poly(ribo)some preparations, the cytosolic mRNA is isolated and separated based on the number of associated ribosomes. Fractions of mRNAs associated with several ribosomes are pooled and designated the translationally active pool. A parallel cytosolic mRNA sample, which is used to assess the cytosolic mRNA level, is also collected. More details are found in [1]. The recent sequencing

based method utilizes the ribosome protection assay where the mRNA part that is bound to a ribosome is resistant to mild RNA degradation treatment and mRNA fragments, which correspond to the expected size that is protected from degradation (hence bound to ribosomes and actively translated), are isolated. This sample is similar to the actively translated sample from the polysome microarray approach. Analogous to the cytosolic mRNA sample that is isolated during the polysome microarray approach, a parallel sample which has been processed similarly but without applying the protection assay is also collected. Thus, both approaches generate two data types, one from cytosolic mRNA and a second from actively translating mRNA.

After isolation of cytosolic mRNA and translationally active mRNA, both samples are labelled and probed with microarrays in the case of the polysome microarray approach or subjected to deep sequencing in the case of the ribosome profiling approach. The result for both methods is, sample per sample, data for cytosolic mRNA level and data for translationally active mRNA. In the simplest case, changes in translational control between two sample classes could be assessed by comparing the data obtained from the translationally active mRNA. However, a change in transcription, mRNA transport from the nucleus to the cytosol or mRNA stability would be expected to also lead to changes in the data derived from translationally active mRNA despite being unrelated to translational activity. It is therefore necessary to correct for differences in the cytosolic mRNA pool when comparing data from translationally active mRNA. To date, correction has primarily been performed by creating per sample differences (log scale) between the translationally active mRNA level and the cytosolic mRNA level [1]. These have then been compared directly between sample classes. However, as discussed in [3] the corrected values often show spurious correlation to the data derived from the cytosolic mRNA. There are also several examples of where the log ratio approach will lead to false conclusions [3].

Most of these problems can be solved by using regression analysis between the translationally active mRNA levels and the cytosolic mRNA levels. Such analysis produces residuals that are uncorrelated with the cytosolic mRNA levels and differential translation can be identified using Analysis of Partial Variance (APV) [3]. However, to apply linear regressions and APV, various assumptions need to be fulfilled for tens of thousands of genes, thus offering substantial challenges. Due to the high dimensionality of the data, anota 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 anota.

The first issue that needs consideration is the appearance of highly influential data points which may cause errors in the regression analyses. 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 analyses performed. Thus we attempt to establish if we, when considering all analysed genes, 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 analyses in which they are found. For this reason, anota provides an output that can be used to flag these results so that they can be examined in more detail if desired.

For detection of influential data points, anota 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, anota 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 cytosolic mRNA level. The translationally active mRNA levels are obtained by sampling data points from a normal distribution with a mean of the cytosolic mRNA level for each cytosolic mRNA level data point. Ten different such data sets are obtained with different variances when sampling the translationally active 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. This level of quality control is performed in the `anotaPerformQc` function.

A second issue concerns the APV assumption that the slopes of the regressions from each sample class are the same so that using the common slope is valid. This assumption postulates that the relationship between the translationally active mRNA level and the cytosolic mRNA level shows the same slope for each sample class, i.e., sample class and cytosolic mRNA levels do not interact in predicting translation mRNA levels. Again, because we analyse tens of 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 NULL distribution. Thus the second level of quality control compares the distribution of the interaction significances as well as the distribution after adjusting the interaction p-values for multiple testing. This level of quality control is performed in the `anotaPerformQc` function.

The third issue relates to the significance testing within the APV framework which assumes that the residuals from the regressions are normally distributed. The `anotaResidOutlierTest` function assesses whether the residuals from the linear regressions (gene by gene) of translationally active mRNA level cytosolic mRNA level are normally distributed. anota 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, anota calculates "envelopes" based on a set of samplings from the normal distribution using the same number of data points as for the true data [5]. 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.

The slopes that are used within anota can take unrealistic values that will influence the analysis of differential translation. These are random events that are likely to be more common when fewer samples and fewer sample classes are analysed. anota therefore tests whether slopes that are <0 (representing unlikely but not impossible translational control [3]) or >1 (slopes >1 are not realistic) differ from 0 and 1 respectively and reports a p-value in the output of the `anotaPerformQc` and `anotaGetSigGenes` functions. This p-value can be used to filter or flag genes with unrealistic slopes.

While anota enables testing of the issues discussed above it is left to the user to decide whether it is possible to use anota to identify differential translation. 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 translation data or the cytosolic 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. One essential component during the polysome preparations is the consistent isolation of the same stratum of the polysomes across all samples (i.e., so that the >n ribosome threshold is met when pooling fractions, not >n+1 or >n-1) because the cut off point will influence the slope. A systematic error in the cut off could cause a high abundance of interactions. If one retrospectively can go back and assess which samples may have error in the cut off one could try to either remove these or use established methods to remove systematic bias.

3. If the resulting residuals deviate strongly from normality an alternative normalization method could be tested.

2 Data set quality control in anota

The `anotaPerformQc` checks whether the data set shows the expected number of highly influential data points and whether there are more significant interactions compared to what is expected by chance. `anotaPerformQc` can also output a set of identifier per identifier regressions (not default) which may be a good approach to see how well regressions seem to work (1).

Further, anota generates an output from the influential data point analysis where the obtained number of influential data points using several suggested thresholds are compared to a simulated data set 2.

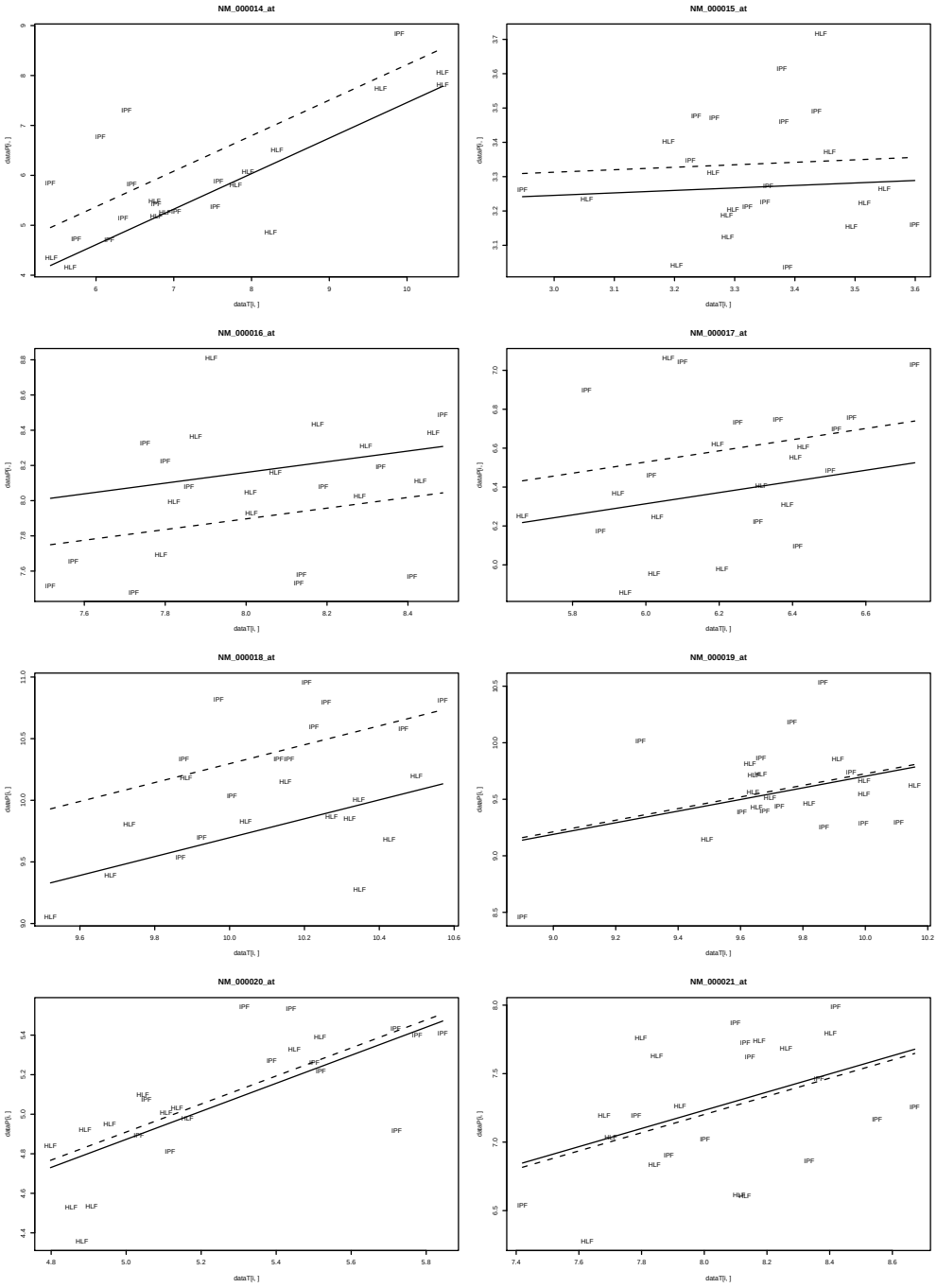

Figure 1: anota can be set to output identifier per identifier regressions between translationally active and cytosolic mRNA levels. Plotting symbols are taken from the `phenoVec` argument and the lines are the regressions lines per samples class

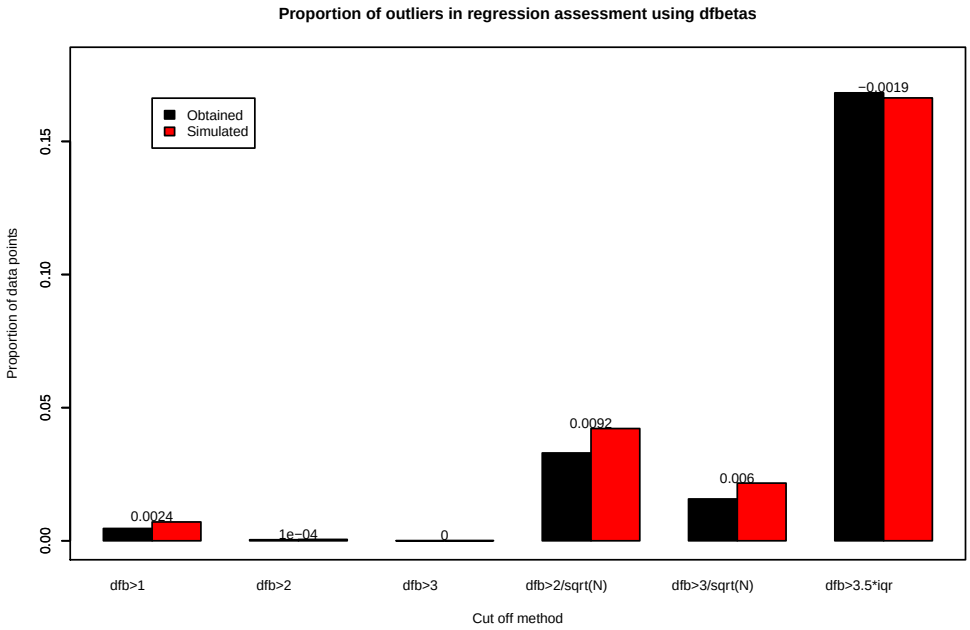

Figure 2: A bar graph showing the obtained and expected (based on a simulation) number of influential data points as judged by different thresholds. For each threshold the difference between the obtained and the simulated frequency of outliers is shown.

anota also generates an output from the analysis of interactions 3. Here the distribution of the obtained interaction significances is assessed to examine if these follow the uniform NULL distribution.

anota also allows for examination of the residuals from the linear regression within the `anotaResidOutlierTest`. As discussed in the introduction, the test allows for examination if the residuals are approximately normally distributed. The single identifier output from `anotaResidOutlierTest` is shown in figure 4 and the summary output output is shown in figure 5.

Finally, as described above, anota reports p-values for unlikely or unrealistic slopes which can be used to filter or flag genes.

3 Using anota to identify differential translation

Once the data set has been validated as suitable for analysis, significant translational regulation is identified. `anotaPerformQc` performs an omnibus group effect test when there are more than 2 sample classes. Within the `anotaGetSigGenes` the user can set custom contrasts to identify significantly differentially translated genes. The output from `anotaGetSigGenes` can be visualized and filtered using the `anotaPlotSigGenes` function to generate both a summary table and per gene plots. The graphical output shows both the graphical inter-

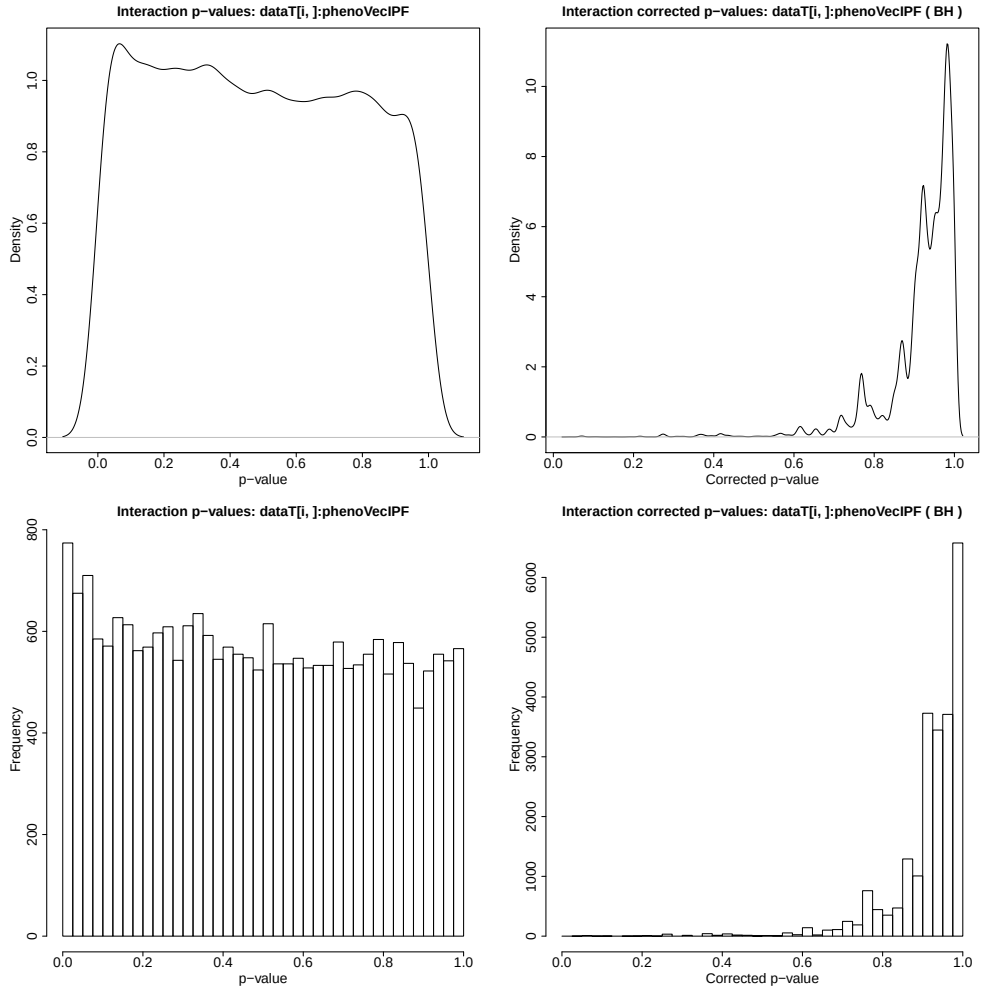

Figure 3: Assessment of whether the significances for the interactions follow the uniform NULL distribution. Shown are both density plots and histograms of the nominal and adjusted p-values (in this case adjusted using Benjamini-Hochberg FDR).

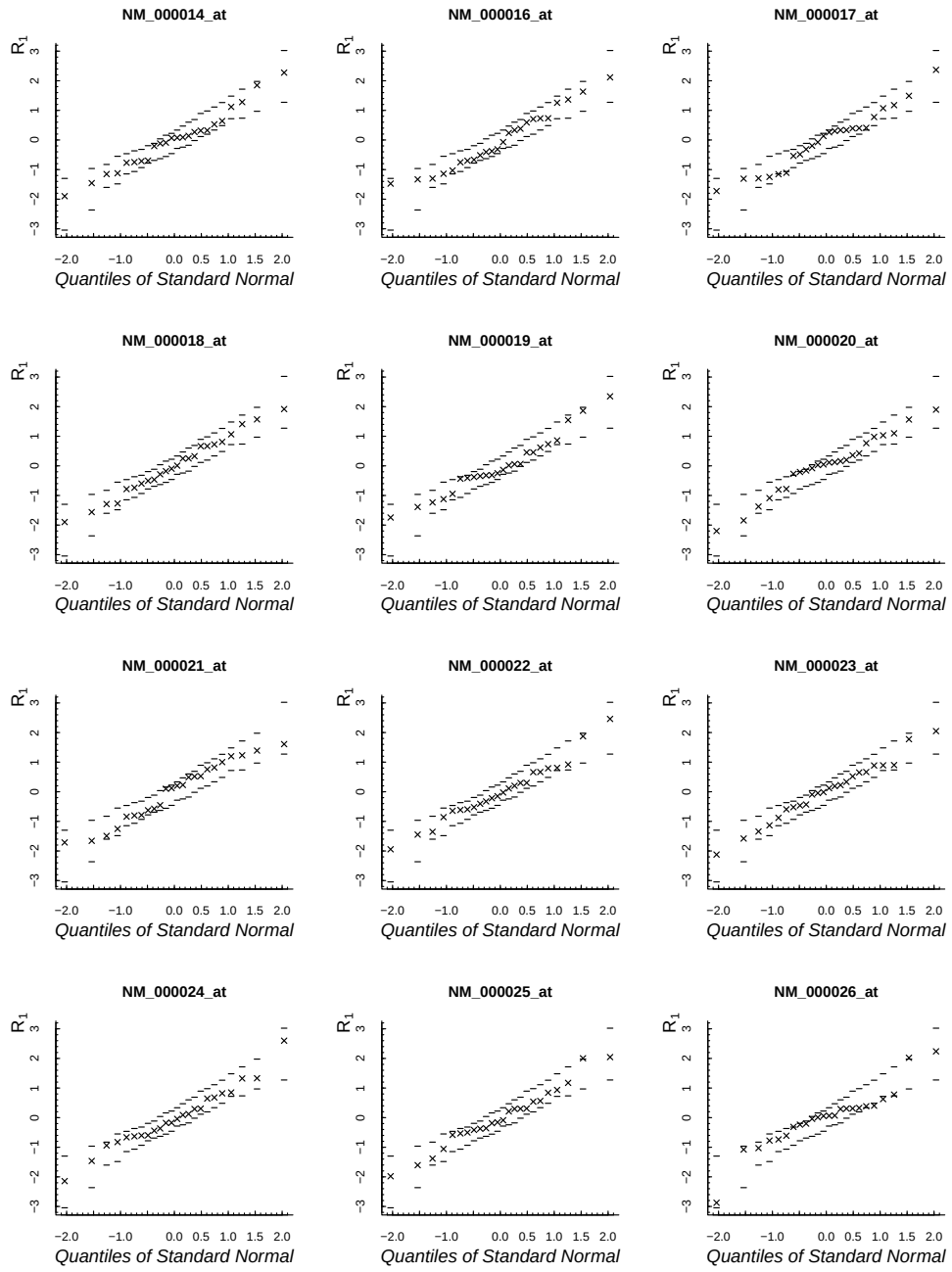

Figure 4: Assessment of whether the residuals are approximately normally distributed. Shown is the output from the single identifier alternative within `anotaResidOutlierTest`. The Q-Q plot for the identifier is compared to the outer limits of a set of Q-Q plots generated by sampling from the normal distribution (described further in the introduction).

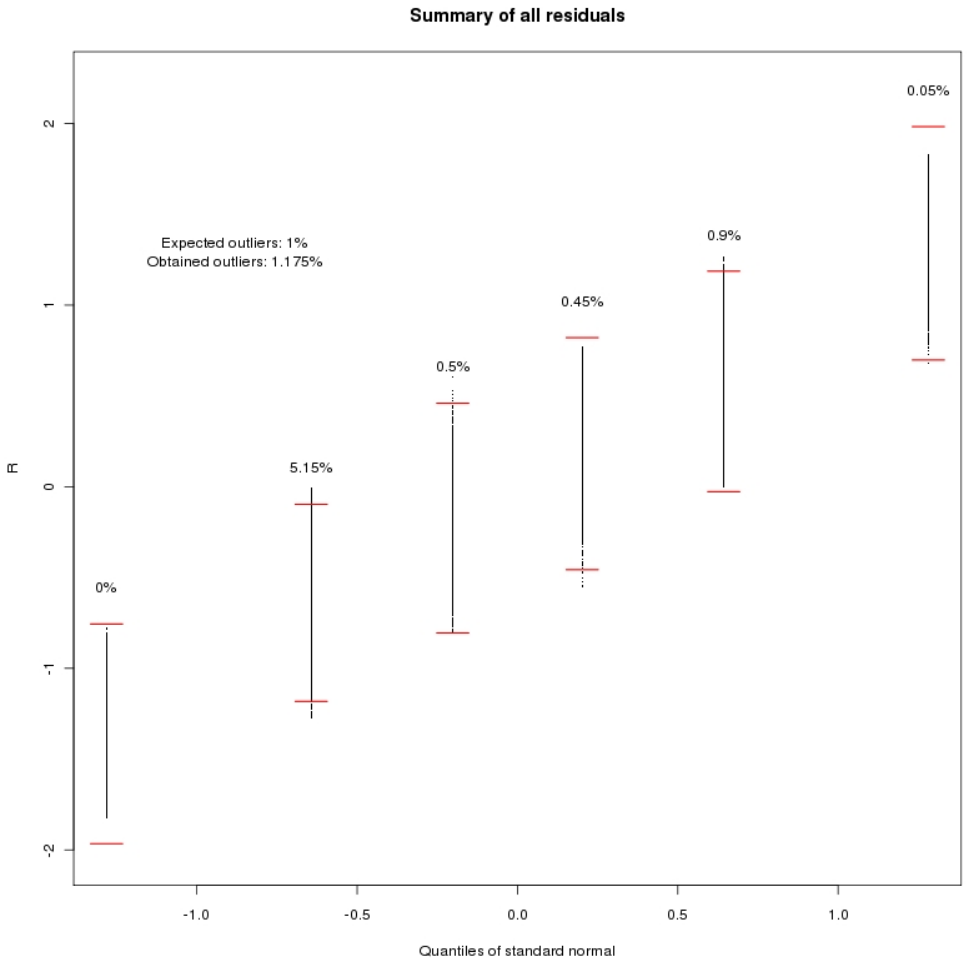

Figure 5: Assessment of whether the residuals are approximately normally distributed. Shown is the output from all identifiers using the `anotaResidOutlierTest` function. The Q-Q plot for the identifiers is compared to the outer limits of a set of Q-Q plots generated by sampling from the normal distribution (described further in the introduction). The obtained and expected percentage of outliers is indicated at each rank position and combined.

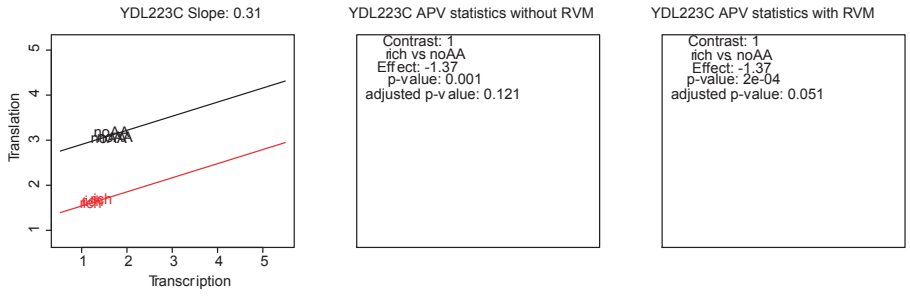

Figure 6: The output from the `anotaPlotSigGenes` function. The left graph shows the identifier per identifier regressions between translationally active and cytosolic mRNA levels. Plotting symbols are taken from the `phenoVec` argument supplied to the `anotaGetSigGenes` function and the lines are the regression lines per sample class 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. When there is more than one contrast all contrasts will be shown but any filterings defined within the `anotaPlotSigGenes` function will be applied to the selected contrast only.

pretation of the APV analysis and the key statistics from both the standard and the RVM based analysis 6. See the function descriptions within the anota R package for each function for more details.

4 Random variance model (RVM) to improve power in detection of differential translational activity within anota

RVM is an empirical Bayes method which has been shown to increase statistical power for small N analysis [6]. In RVM, the within gene variance is adjusted using the variance obtained from an inverse gamma distribution derived from the variances of all genes. A key assumption in RVM is that the resulting variances follow a theoretical F-distribution. anota test this for the analysis of omnibus interactions, omnibus group effects and the identification of differential translational activity. Each of these analyses generates a comparison of the obtained empirical distribution compared to the theoretical distribution (similarity assessed using a KS test which NULL hypothesis should not 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. It is necessary to validate that application of RVM does not influence the distribution of the interaction p-values. Figure 7 shows the output from the test of the fit.

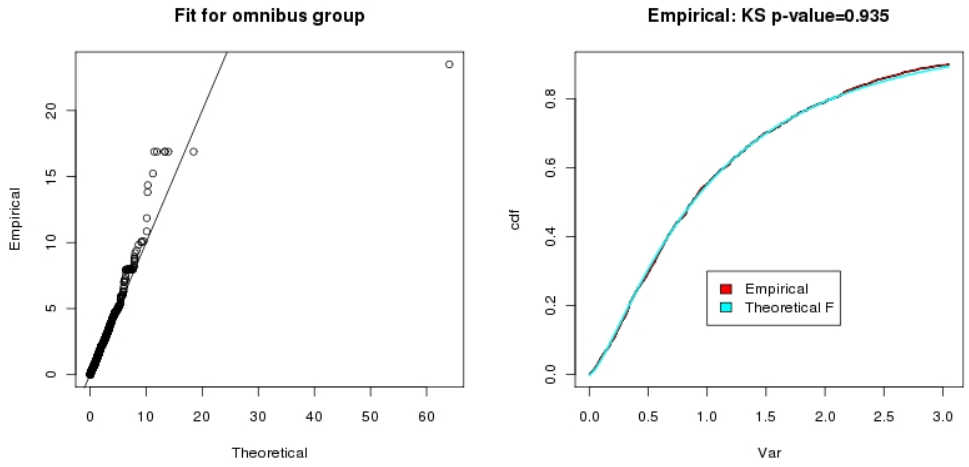

Figure 7: A comparison of the obtained variances to the theoretical F-distribution. RVM assumes that the empirical and the theoretical distributions are similar.

5 Data set requirements for application of anota

anota can analyse data from both sequencing based methods and the more standard polysome microarray method. anota cannot use data from competitive two channel experiments when the translationally active mRNA is directly compared to cytosolic mRNA as these do not allow independent estimates of the cytosolic and the translationally active mRNA levels. A two-channel reference design should be applicable although we have not tested this data type. anota requires 3 replicate experiments per group. The performance will vary depending on the normalization and the transformation of the data. We therefore recommend that the user tries several different transformations and normalization 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. We recommend using RVM as it improves the power to detect differential translation activity within anota [3].

6 Example

The example illustrates a typical analysis procedure using a part of the data set reported in [4]

```
> ##Loading the library and data set, perform quality control and identify significant
>  library("anota")
>  data(anotaDataSet)
>  anotaQcOut <- anotaPerformQc(dataT=anotaDataT[1:200,], dataP=anotaDataP[1:200,], ph

Running anotaPerformQc quality control
        Calculating omnibus interactions & effects and dfbetas
```

```
        Assessing dfbetas for model without interaction
        Performing dfbetas simulation


        Using RVM for omnibus interaction statistics
        The a and b parameters for the inverse gamma distribution are:
         a: 2.74032040787002 b: 22.626023943369
        Using RVM for omnibus group statistics
        The a and b parameters for the inverse gamma distribution are:
         a: 2.7630233823186 b: 20.1223737663144
        Adjusting p-values for multiple testing
```

```
> anotaResidOut <- anotaResidOutlierTest(anotaQcObj=anotaQcOut, useProgBar=FALSE)
```

```
Running anotaResidOutlierTest
```

```
> anotaSigOut <- anotaGetSigGenes(dataT=anotaDataT[1:200,], dataP=anotaDataP[1:200,],
```

```
Running anotaGetSigGenes
        Using default "treatment" contrasts between (custom contrasts can be set):
         noAA rich
        These contrasts will be evaluated:

     contrast 1
noAA         -1
rich          1
```

```
> anotaSelected <- anotaPlotSigGenes(anotaSigObj=anotaSigOut, selContr=1, maxP=0.1, mi
```

References

[1] Larsson, O. and Nadon, R. Gene expression: Time to change point of view? *Biotechnology and Genetic Engineering Reviews*, 2008, 25 p77-92.

[2] Ingolia, NT et al. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, 2009, 10;324(5924):218-23.

[3] Larsson, O. et al. Identification of differential translation in genome wide studies. *PNAS*, 2010, 107;50:21487-21492.

[4] Larsson, O. et al. Fibrotic myofibroblasts manifest genome-wide derangements of translational control. *PLoS One*, 2008, 3 (9) e3220.

[5] Venable, W.N. and Ripley, B.D. Modern applied statistics with S-PLUS. *Springer*, 1999.

[6] Writht, G.W. and Simon, R.M. A random variance model for detection of differential gene expression in small microarray expeirments. *Bioinformatics*, 2003, 12;19(19):2448-55.