BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 10 2011, pages 1440–1441
doi:10.1093/bioinformatics/btr146

Gene expression Advance Access publication March 21, 2011

anota: analysis of differential translation in genome-wide studies

Ola Larsson[1,*,†] Nahum Sonenberg[1] and Robert Nadon[2,3,*]

[1]Department of Biochemistry, McGill University, Montreal, Quebec H3A 1A3, [2]Department of Human Genetics, McGill University, Montreal, Quebec H3A 1B1 and [3]McGill University and Genome Quebec Innovation Centre, Montreal, Quebec H3A 1A4, Canada

Associate Editor: Trey Ideker

ABSTRACT

Summary: Translational control of gene expression has emerged as a major mechanism that regulates many biological processes and shows dysregulation in human diseases including cancer. When studying differential translation, levels of both actively translating mRNAs and total cytosolic mRNAs are obtained where the latter is used to correct for a possible contribution of differential cytosolic mRNA levels to the observed differential levels of actively translated mRNAs. We have recently shown that analysis of partial variance (APV) corrects for cytosolic mRNA levels more effectively than the commonly applied log ratio approach. APV provides a high degree of specificity and sensitivity for detecting biologically meaningful translation changes, especially when combined with a variance shrinkage method for estimating random error. Here we describe the *anota* (analysis of translational activity) R-package which implements APV, allows scrutiny of associated statistical assumptions and provides biologically motivated filters for analysis of genome wide datasets. Although the package was developed for analysis of differential translation in polysome microarray or ribosome-profiling datasets, any high-dimensional data that result in paired controls, such as RNP immunoprecipitation-microarray (RIP-CHIP) datasets, can be successfully analyzed with *anota*.

Availability: The *anota* Bioconductor package, www.bioconductor.org.

Contact: ola.larsson@ki.se; robert.nadon@mcgill.ca

Received on January 25, 2011; revised on March 2, 2011; accepted on March 9, 2011

1 INTRODUCTION

Translational control of gene expression acts primarily at the initiation step of translation prior to peptide bond formation and controls how many ribosomes are initiated per mRNA (Mathews *et al*., 2007), thereby enabling instantaneous control of gene expression. Studies in many biological model systems and human diseases have documented that such regulation can be mRNA specific and can have large biological impact (Silvera *et al*., 2010; Sonenberg and Hinnebusch, 2009). Genome-wide patterns of differential translation nonetheless remain understudied, in part for lack of adequate statistical analysis procedures (Larsson and Nadon, 2008).

Genome-wide translational control data are currently generated using two approaches: polysome microarrays which involve isolation of actively translating mRNAs that are probed with DNA microarrays (Larsson and Nadon, 2008), and ribosome profiling where RNA fragments that are protected by ribosomes are isolated and sequenced (Ingolia *et al*., 2009). Neither of these data types, however, is independent of total cytosolic mRNA levels. In order to study differential translation *per se*, it is therefore necessary to correct the translationally active mRNA data for confounding total cytosolic mRNA levels. We have recently shown that analysis of partial variance (APV) in combination with the Random Variance Model (RVM) (Wright and Simon, 2003) is an effective method for such analysis that circumvents problems with the commonly applied log ratio approach in combination with *t*-tests or analysis of variance (ANOVAs) (Larsson *et al*., 2010). In APV, a hierarchical per gene linear regression is performed on translational activity data, with cytosolic mRNA data and contrast vectors (which define membership in sample classes such as genotype, treatment or disease) as covariates. Class comparison effects are estimated by calculating differences between class intercepts (Larsson *et al*., 2010).

As with all statistical procedures, it is important to evaluate model assumptions. An interpretive problem arises in high-throughput contexts, however, because of the large number of individual statistical tests. In genome-wide analysis of translational activity, for example, a number of genes are expected to fall outside of the assumption space simply by chance. If the number of assumption violations does not exceed chance levels, the data are behaving as expected of a random variable and analysis can safely proceed. Individual genes with serious violations of assumptions, however, should nonetheless be interpreted with caution. Thus, genome-wide application of APV to identify differential translation requires a set of methods that assess the genome-wide performance of APV, relates it to chance outcomes and provides filtering capabilities based on statistical and biological criteria. Here we provide the *anota* (analysis of translational activity) statistical package, which is implemented in the R statistical computing language and contains the analytical approaches and diagnostic outputs to guide such analysis.

2 IMPLEMENTATION

Identification of differential translation within *anota* is a two-step process. Model assumptions are assessed, followed by identification of differential translation.

2.1 Assessment of model assumptions

Model assumptions are examined first to determine if the data are appropriate for *anota* (Fig. 1A). If there are more violations of

[*] To whom correspondence should be addressed.

[†] Present address: Department of Oncology-Pathology, Cancer Center Karolinska, Karolinska Institute, R8:01, Stockholm 171 76, Sweden.

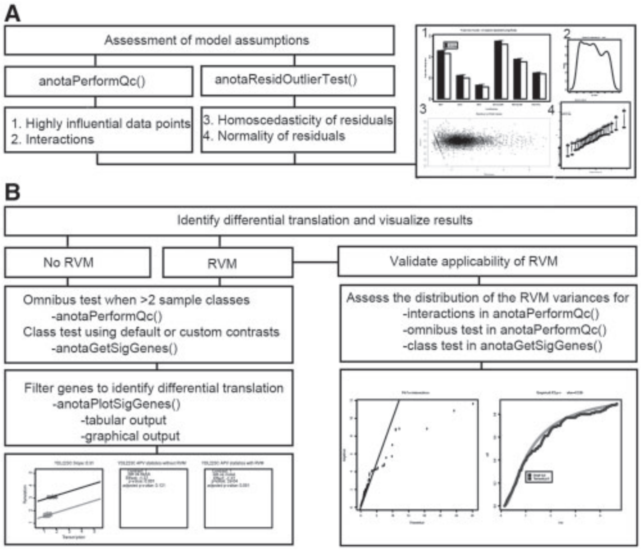
A
Assessment of model assumptions
anotaPerformQc()
anotaResidOutlierTest()
1. Highly influential data points
2. Interactions
3. Homoscedasticity of residuals
4. Normality of residuals
1
2
3
4
B
Identify differential translation and visualize results
No RVM
RVM
Validate applicability of RVM
Omnibus test when >2 sample classes
-anotaPerformQc()
Class test using default or custom contrasts
-anotaGetSigGenes()
Assess the distribution of the RVM variances for
-interactions in anotaPerformQc()
-omnibus test in anotaPerformQc()
-class test in anotaGetSigGenes()
Filter genes to identify differential translation
-anotaPlotSigGenes()
-tabular output
-graphical output

Fig. 1. An overview of assessment of model assumptions (**A**) and data analysis (**B**) within *anota*.

assumptions than expected of a random variable, conclusions based on *anota* may be suspect. We note, however, that we have used *anota* with eight datasets, each of which met model assumption requirements. These analyses were done on three published (Larsson *et al*., 2010) and three yet to be published datasets on translational activity and two datasets from RIP-CHIP experiments. Assumptions that can be examined in *anota* include:

(i) Highly influential data points which may unduly affect parameter estimates in the linear model. We compare the frequency of highly influential data points (defined by standardized *dfBeta* [*dfbetas*]) to values generated in simulated data. For the simulation, translationally active mRNA data and total cytosolic mRNA data are sampled from a normal distribution.

(ii) In APV it is assumed that the sample classes share a common slope. *anota* allows the user to assess this assumption by generating a distribution of all cytosolic mRNA × group interaction *P*-values which test for differences among slopes. The *P*-values should approximately follow a uniform distribution when the classes share common slopes.

(iii) APV assumes homoscedasticity of residuals. This can be evaluated within *anota* by examining residuals as a function of signal intensity.

(iv) The APV model assumes that per gene residuals are normally distributed. This assumption can be examined in *anota* by comparing per gene residuals to a simulation of residuals from a normal distribution. Percentage of outliers is compared to what is expected by chance.

If these four criteria are fulfilled, *anota* can be used to identify differential translation. More details regarding the simulations and guidance for interpreting outputs are found in the *anota* manual.

2.2 Analysis of differential translation

Identification of differential translation within *anota* can be performed in an omnibus mode (testing differences in translational activity among three or more sample classes) and/or by setting custom contrasts within the anotaGetSigGenes function to specify particular class comparisons (Fig. 1B). Both the omnibus and the custom contrasts analyses are performed using APV with or without RVM. We recommend the former when there are few replicates for its superior statistical power (Larsson *et al*., 2010). When applied, RVM's assumption that gene variances are random variables from an inverse gamma distribution should be examined in *anota* for both the omnibus and the contrast tests. The RVM *P*-values for the interactions described above (Point ii) should also follow a uniform distribution under the assumption of a common slope among the classes.

There are two additional biologically motivated slope considerations when using *anota*. Slopes >1 should not occur inasmuch as the actively translated mRNAs are a fraction of all mRNAs. Similarly, although slopes <0 are biologically plausible, they indicate complex regulatory mechanisms which are usually not the primary focus. To address this, *anota* tests whether slopes are significantly >1 or <0 and reports a *P*-value that can be used for filtering. Unrealistic slopes will likely be more common in datasets with few replicates/sample classes and it is therefore especially important to consider the slope when evaluating results from such studies. Filtering of genes based on slope and many other user-defined criteria is done using the anotaPlotSigGenes function which generates both a tabular summary and per mRNA graphical representations of the APV analysis together with key statistics. This output can easily be used to examine the performance of APV for individual genes of interest. The output also includes intercepts for each class which could be used, for example, in clustering of translational activity levels. Nominal and various types of false discovery rate (FDR)-adjusted *P*-values can also be produced and exported for further examination within R or other statistical software.

2.3 Work flow

Translationally active mRNA data and cytosolic mRNA data will typically be analyzed within *anota* using four main functions: anotaPerformQc, anotaResidOutlierTest, anotaGetSigGenes and anotaPlotSigGenes (Fig. 1A and B).

ACKNOWLEDGEMENTS

We thank Richard Simon and George Wright for permission to incorporate some of their original RVM code into *anota*.

Funding: O.L. was initially supported by a fellowship from the Knut and Alice Wallenberg foundation. Partial support was provided by Le Fonds Québécois de la Recherche sur la Nature et les Technologies (FQRNT) Grant 119258 (to R.N.).

Conflict of interest: none declared.

REFERENCES

Ingolia,N.T. *et al.* (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, **324**, 218–223.

Larsson,O. and Nadon,R. (2008) Gene expression - time to change point of view? *Biotechnol. Genet. Eng. Rev.*, **25**, 77–92.

Larsson,O. *et al.* (2010) Identification of differential translation in genome wide studies. *Proc. Natl Acad. Sci. USA*, **107**, 21487–21492.

Mathews,M. *et al.* (2007) *Translational Control in Biology and Medicine*. CSHL Press, New York.

Silvera,D. *et al.* (2010) Translational control in cancer. *Nat. Rev. Cancer*, **10**, 254–266.

Sonenberg,N. and Hinnebusch,A.G. (2009) Regulation of translation initiation in eukaryotes: mechanisms and biological targets. *Cell*, **136**, 731–745.

Wright,G.W. and Simon,R.M. (2003) A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics*, **19**, 2448–2455.