Regulation of gene expression through translational control is a fundamental mechanism implicated in many biological processes ranging from memory formation to innate immunity and whose dysregulation contributes to human diseases. Genome wide analyses of translational control strive to identify differential translation independent of cytosolic mRNA levels. For this reason, most studies measure genes’ translation levels as log ratios (translation levels divided by corresponding cytosolic mRNA levels obtained in parallel). Countervariantly, arising from a mathematical necessity, these log ratios tend to be highly correlated with the cytosolic mRNA levels. Accordingly, they do not effectively correct for cytosolic mRNA level and generate substantial numbers of biological false positives and false negatives. We show that analysis of partial variance, which produces estimates of translational activity that are independent of cytosolic mRNA levels, is a superior alternative. When combined with a variance shrinkage method for estimating error variance, analysis of partial variance has the additional benefit of having greater statistical power and identifying fewer genes as translationally regulated resulting merely from unrealistically low variance estimates rather than from large changes in translational activity. In contrast to log ratios, this formal analytical approach estimates translation effects in a statistically rigorous manner, eliminates the need for inefficient and error-prone heuristics, and produces results that agree with biological function. The method is applicable to datasets obtained from both the commonly used polysome microarray method and the sequencing-based ribosome profiling method.

Identification of differential translation in genome wide studies

Ola Larsson1,a,2, Nahum Sonenberg3, and Robert Nadon4,a,2

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

Edited by Peter J. Bickel, University of California, Berkeley, CA, and approved October 26, 2010 (received for review May 20, 2010)

Results and Discussion

Common Issues in Genome Wide Analysis of Translational Control. Two types of data are produced from each sample when studying translational activity: actively translated mRNAs (“translational activity data”) and cytosolic mRNAs (“cytosolic mRNA data”) obtained in parallel. Typically, log (translational activity data/ cytosolic mRNA data) ratios are calculated (i.e., logged cytosolic mRNA data is subtracted from its associated logged translational activity data) with the idea of obtaining a cytosolic mRNA data-corrected estimate of translation activity and compared between classes (Fig. 1A). Because log ratios are differences between logged values, class comparison effects are estimated from difference scores. Log difference scores can be correlated with cytosolic mRNA data, however, leading to incorrect biological conclusions. Such is the case when translational activity data and cytosolic mRNA data are uncorrelated for technical or biological reasons (Fig. 1 B and C). For example, false positives can arise from log ratios when mRNAs fail to reach the predetermined threshold for the number of ribosomes necessary to join the pool of actively translated mRNAs or when short poorly translated mRNAs, but not their paired cytosolic mRNAs, fail to achieve counts for protected mRNA pieces above the noise level. Log ratios for such mRNAs tend to produce false positives as they appear to be under translational control (Fig. 1B). False negatives can be produced by log ratios when translational activity is regulated independently of the cytosolic mRNA level (Fig. 1C). The phenomenon of a difference score (Y − Z) correlating with each of its terms (Y and Z) was first described by Pearson in 1897 (7) who labeled it “spurious correlation” because of the frequent practice of interpreting such correlations as substantive rather than artifactual. We will also refer to these correlations as spurious, although our focus will be on the inadequacy of difference scores to control for cytosolic mRNA data and as a consequence fail to correctly estimate class effects (e.g., genotype, treatment, disease). The following equation illustrates the mathematical

Author contributions: O.L., N.S., and R.N. designed research; O.L. performed research; O.L. analyzed data; and O.L., N.S., and R.N. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

1Present address: Department of Oncology-Pathology, Cancer Center Karolinska, Karolinska Institute, B103, 171 76 Stockholm, Sweden.

2To whom correspondence may be addressed. E-mail: ola.larsson@ki.se or robert.nadon@mcgill.ca.

This article contains supporting information online at www.pnas.orglookup/suppl/doi:10.1073/pnas.1006821107/-/DCSupplemental.
Fig. 1. (A–D) Common problems when analyzing translational activity data. Data were simulated to generate two sample classes, each with five measurements of paired translational activity data (indicated as translation on each y-axis) and cytosolic mRNA data (indicated as transcription on each x-axis). The examples were generated to illustrate different scenarios of translational control analysis. For each example, a two-tailed t-test was performed comparing the sample classes using the translational activity data only or the log-ratio data (p-values as indicated). anota was also performed on each example (p-value as indicated). Lines represent the regression lines estimated by anota for the two sample categories. (E) Common spurious correlations in published datasets of translational control. Shown are boxplots of the spurious correlations that emerge between the log (translational activity data/cytosolic mRNA data) ratios and log cytosolic mRNA data. (F) Small overlap between the log-ratio approach and anota. The top 5% genes ranked by significance from the log-ratio approach and anota were collected. The overlap between the log-ratio approach and anota is visualized using Venn diagrams for the three example studies.
necessity underlying the problem (8):

\[
r_{Y-Z} = \frac{\text{rYZ} \cdot s_Y \cdot s_Z - s_{YZ}^2}{s_Y \sqrt{s_Y^2 + s_Z^2} - 2 \text{rYZ} \cdot s_Y \cdot s_Z}
\]

where \( Y \) is a vector of translational activity data for a specific mRNA, \( Z \) is a vector of paired cytosolic mRNA data for the same mRNA, \( r \) is the Pearson correlation coefficient, and \( s \) is the sample standard deviation. When translational and cytosolic levels are uncorrelated, Eq. 1 then simplifies to:

\[
r_{Y-Z} = \frac{-s_Z}{\sqrt{s_Y^2 + s_Z^2}}.
\]

Eq. 2 makes the origin of these false negatives and false positives clear; under this circumstance the correlation between the log ratios and their corresponding log cytosolic mRNA data is a function of the standard deviations of translational activity data and cytosolic mRNA data replicates. When the standard deviations are also equal, Eq. 2 yields a correlation of \((-0.71) \approx (-1/\sqrt{2})\). That is, half \((-0.71^2)\) of the variance associated with the log ratio would be due to cytosolic mRNA levels. A similar difficulty with log ratios arises in the more typical situation in which the correlation between translational activity data and cytosolic mRNA is nonzero. When standard deviations are also equal, Eq. 1 then simplifies to:

\[
r_{Y-Z} = \frac{1}{\sqrt{2}} \cdot \frac{(\text{rYZ} - 1)}{\sqrt{1 - \text{rYZ}}}.
\]

When the correlation between translational activity and paired cytosolic mRNA is 0.60, Eq. 3 yields a correlation of \(-0.45\). Thus, under various realistic scenarios, the correlation between log ratios and cytosolic mRNA is nontrivial and gives rise to biological false positives and negatives.

To examine the practical implications of problems with log ratios, we calculated the per-gene correlations between logged translational activity data/cytosolic mRNA data ratios and log cytosolic mRNA data per gene across all samples in a set of published studies (9–28). The median of the per dataset medians for the spurious correlations was \(-0.61\) with values covering almost the entire range between \(+1\) and \(-1\) (Fig. 1E). Although these results do not imply low data quality, they demonstrate that log ratios are inappropriate for controlling the confounding effects of cytosolic mRNA levels. It is noteworthy that using the union from the translational activity data only analysis and the log ratio approach resolves the false negative example (Fig. 1C), but leads to increased false positives (Fig. 1A and B). Using the intersection of the translational activity data only analysis and the log ratio approach limits identification of differential translation to situations that are shown in Fig. 1D and may fail to identify instances of translational regulation that are independent of cytosolic mRNA levels (Fig. 1C). One might argue that a set of heuristic thresholds for differences at the translational activity data and cytosolic mRNA data could be used either alone or in combination with the translational activity data only or the log ratio approach. Such heuristics suffer from numerous failings, however, including the absence of objective statistical thresholds and that any choice of subjective threshold generates unknown rates of false positives and false negatives. Thus, from both theoretical and practical standpoints, current approaches are inadequate for analysis of differential translation.

Anota Outperforms Current Approaches for Analysis of Differential Translation. APV (29) provides a more suitable analysis of translational activity data because APV-corrected translation measurements do not, by definition, show spurious correlations. We have implemented this approach in the anota (analysis of translational activity) R-package. In anota, a common slope for all sample categories is identified for each gene from the least squares linear regression of translational activity data on cytosolic mRNA data. Class comparison effects are estimated by calculating differences between sample category intercepts; the sum of squares error for these comparisons is reduced by the sum of squares associated with the covariance between the translational activity data and the cytosolic mRNA data. We found that anota potentially correctly analyzes the different scenarios outlined in Fig. 1A–D. Anota corrects for confounding cytosolic mRNA levels (Fig. 1A), handles false positives arising from technical issues (Fig. 1B), identifies translational regulation that is independent of cytosolic mRNA levels (Fig. 1C), and identifies translational control when there is some confounding cytosolic mRNA difference (Fig. 1D). In summary, anota handles different problematic scenarios by automatically adapting the analysis to the situation at hand.

To further assess anota’s performance, we analyzed genome wide data from both the polysome microarray approach and the sequencing-based approach in three studies [Ingolia et al. (24), Kitamura et al. (21), and Otulakowski et al.(9); see Fig. 1E]. Two sample categories were identified in each study. We compared results obtained from anota with those obtained from a t-test between the log ratios of the two sample classes. Scatter plots of effects and of p-values reveal large discrepancies between the two approaches for many genes (Fig. S1). We investigated this further by classifying mRNAs as either translationally activated or inactivated and examining the overlap among the top 5% of differentially translated mRNAs (ranked by nominal p-values) between the two analytic approaches. As shown in Fig. 1F, the number of genes that were identified as translationally regulated by one method but not the other was substantially higher than the number of shared genes. Genes were then classified into one of three sets: identified by anota only, log ratio only, or by both log ratio and anota. The top two genes (ranked by p-value) from each set from the Ingolia et al. study (24) illustrate the differences between the methods (Fig. 2). Of the top two genes identified by the log ratio approach, one was of the type exemplified in Fig. 1B with cytosolic mRNA data differences combined with smaller translational activity data differences (Fig. 2, rank 2) whereas the other showed a larger difference at the cytosolic mRNA level compared to the translational activity level, leading to the conclusion that a gene with lower translational activity data is more translationally active. This could be a case of “over-correction” and hence represents a false positive or possibly a biologically interesting finding. However, it is commonly assumed that a gene that is translationally activated would also show higher protein levels and this gene would therefore not be a primary target for follow-up studies. The top two genes that were identified by anota only showed translational regulation that was independent of cytosolic mRNA levels and accordingly belonged to the type exemplified in Fig. 1C. The genes that were identified by both anota and the log ratio approach showed translational regulation independent of cytosolic mRNA level (as exemplified in Fig. 1C) or a small difference in cytosolic mRNA level associated with a larger difference in translational activity data (as exemplified in Fig. 1D). The two other studies analyzed in Fig. 1F showed similar results (Figs. S2 and S3).

To extend the observations from Fig. 2, we compared all genes from the three sets (identified by anota only, log ratios only, or both log ratios and anota). Within each set, genes were classified further into one of three modes of regulation based on between-group changes in translational activity data and cytosolic mRNA data (shown as delta translation and delta transcription respectively, in Fig. 2). The first mode, labeled “translation > transcription,” identifies those genes for which the translational
activity data difference was larger than the opposing cytosolic mRNA difference; i.e., for a gene that is translationally activated, the activation originated primarily from an activation at the translational activity level and not from a decrease at the cytosolic mRNA level (all genes identified by anota only or by both anota and log ratios in Fig. 2 fit into this group). The second mode, labeled “no translation,” identifies those genes for which there was no difference in the translational activity data; i.e., for a gene defined as activated, no increase in the translational activity data (delta translation ≤ 0) was associated with a decrease in cytosolic mRNA data (the log ratio only genes from Fig. 2 would belong to this mode). The third mode, labeled “transcription > translation,” identifies those genes that show a translational activity data difference but for which the cytosolic mRNA data difference in the opposite direction is of larger magnitude. Of these three modes, the transcription > translation mode corresponds to the examples shown in Fig. 1C and D and represents cases when translational regulation is plausible whereas the no translation and the transcription > translation modes are variants of the example in Fig. 1B when it is unclear if there is any translational regulation. Anota primarily identified genes belonging to the preferred translation > transcription mode whereas most of the genes identified by the log ratio approach belong to the undesired no translation or transcription > translation modes (Fig. 3). Interestingly, those genes that were identified by both anota and the log ratio approach showed a distribution among the three modes similar to that of anota and thus represent a subset of genes for which the log ratio approach obtained a result similar to anota. In summary, anota outperforms the log ratio approach and can be used both for microarray-based as well as sequencing-based datasets.

**Improving the Performance of Anota Using Variance Shrinkage.** In genome wide studies, gene-specific variances of many genes will be greatly under or overestimated due to chance factors. It is therefore possible that some of the genes shown in Fig. 2 have been identified as a result of unrealistically low variance estimates. This characteristic of high dimensional data has produced a consensus among methodologists to favor variance shrinkage methods that adjust error variances for statistical tests by weighting gene-specific variance estimates and a variance estimate obtained from all genes in the dataset (30). We have generalized one such method, the random variance model (RVM) (31), for APV within anota (referred to as “anota RVM”). We performed an analysis identical to that shown in Fig. 2 but with the RVM-adjusted error term. The top three genes that were selected by anota RVM but not by the log ratio approach in the Ingolia et al. study (24) are shown in Fig. 4A. These genes show larger effects than those genes identified without RVM (compare Fig. 4A to Fig. 2). To illustrate the effect across all genes, we generated “volcano” plots comparing anota and anota RVM. The volcano plots from the Ingolia et al. dataset show that smaller effects are less likely to be asso-
Conclusions

Regulation of translation is important for many biological processes and dysregulation is present in diseases such as cancer and fibrosis. Despite this, genome-wide studies of differential translation have been rare. Downstream of the technical challenges, data analysis has in the past relied on simple intuitive approaches that we show are ineffective for analysis of differential translation and have led to many false inferences that may have hampered the enthusiasm for genome-wide studies of translational control. Our approach to analysis of translational activity provides an advancement that is necessary to understand how mRNA specific translation is regulated on a translatome wide scale and shows promise for generating reproducible and more biologically sound findings.

We show that our approach for analysis of differential translation is versatile and can be used both on count data from sequencing as well as microarray data. Moreover, our approach is not limited to studies of translational control but could be used in, for example, ribonucleoprotein immunoprecipitation–microarray (RIP-chip) studies where the observed differential level of immunoprecipitated mRNA will be partially dependent on the total mRNA amount. In general, anota can be used profitably with any datasets containing paired-controls.

Defining which genes are under translational control only represents the first step toward mechanistic understanding of translational regulation and its underlying organization. Efforts are currently being made toward identifying tools and approaches that will be necessary to discover mechanisms that define the common regulatory patterns as described in the posttranscriptional regulon theory and by the originally proposed informosomes (33–36). We believe that our methodological improvements for analysis of genome-wide datasets are in this spirit and will help to understand how the translatome is regulated from a systems perspective. Such characterization will be necessary to understand how dysregulation of translational control leads to human diseases.

Materials and Methods

Sampling of Example Data. Data shown in Fig. 1 A–D were sampled from normal distributions. First, 5 cystolic mRNA data points were sampled and then a set of corresponding translational activity data points were sampled using the values of the sampled cystolic mRNA data as distribution means. A second sample class was generated in the same manner. The different examples in Fig. 1 A–D were generated by changing the means and the standard deviations of the normal distributions within the “norm” function in R.

A Survey of the Occurrence of Spurious Correlations in Published Datasets. We searched both the ArrayExpress (37) and the Gene Expression Omnibus (38) databases for a set of studies of translational activity. Datasets generated using Affymetrix GeneChips were normalized with the robust multiarray average method (39, 40) using updated probe set definitions when available (41, 42). For the cDNA dataset (11), we used the normalized data provided by the authors. For the sequencing dataset (24), we used the count data supplied by the authors, filtered for identifiers originating from the coding regions, and used quantile normalization and a transformation to stabilize the variance. For all studies, we identified paired translational activity and cystolic mRNA data and calculated correlations between the log2 difference scores and the cystolic mRNA data per gene across all samples in each study.

Analysis of Differential Translation Using Three Datasets. We identified two sample classes from each study to be used for the analysis of differential translation with anota RVM.
21. Kitamura H, et al. (2008) Genome-wide identification and characterization of genes of that gene in the empirical data. Correlation between the simulated and empirical data with a mean and a standard deviation derived from the data points was used to compare the log ratios between the sample categories.

ACKNOWLEDGMENTS. We thank Richard Simon and George Wright for permission to incorporate some of their original RVM code into anola. 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 (FCRNT) Grant 119258 (R.N.).