Assessing gene-level translational control from ribosome profiling

Adam B. Olshen1,2, Andrew C. Hsieh2,3,4, Craig R. Stumpf3,4, Richard A. Olshen5, Davide Ruggero2,4 and Barry S. Taylor1,2,3,*

1Department of Epidemiology and Biostatistics, 2Helen Diller Family Comprehensive Cancer Center, 3Department of Medicine and 4Department of Urology, University of California, San Francisco, CA 94158, USA and 5Department of Health Research and Policy, Stanford University School of Medicine, Stanford, CA 94305, USA

ABSTRACT

Motivation: The translational landscape of diverse cellular systems remains largely uncharacterized. A detailed understanding of the control of gene expression at the level of messenger RNA translation is vital to elucidating a systems-level view of complex molecular programs in the cell. Establishing the degree to which such post-transcriptional regulation can mediate specific phenotypes is similarly critical to elucidating the molecular pathogenesis of diseases such as cancer. Recently, methods for massively parallel sequencing of ribosome-bound fragments of messenger RNA have begun to uncover genome-wide translational control at codon resolution. Despite its promise for deeply characterizing mammalian proteomes, few analytical methods exist for the comprehensive analysis of this paired RNA and ribosome data.

Results: We describe the Babel framework, an analytical methodology for assessing the significance of changes in translational regulation within cells and between conditions. This approach facilitates the analysis of translation genome-wide while allowing statistically principled gene-level inference. Babel is based on an errors-in-variables regression model that uses the negative binomial distribution and draws inference using a parametric bootstrap approach. We demonstrate the operating characteristics of Babel on simulated data and use its gene-level inference to extend prior analyses significantly, discovering new translationally regulated modules under mammalian target of rapamycin (mTOR) pathway signaling control.

Availability: The Babel framework is freely available as source code at http://taylorlab.ucsf.edu/software_data.html.

Contact: barry.taylor@ucsf.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on July 7, 2013; revised on August 19, 2013; accepted on September 7, 2013

1 INTRODUCTION

The translation of cellular messenger RNAs (mRNAs) is the all-important final product of gene expression. Nevertheless, a complete understanding of gene expression at the level of mRNA translation is generally lacking. As mRNA levels explain only a modest fraction of protein abundance, we must establish the landscape of mRNA translation and determine the mechanisms by which it is regulated to fully elucidate diverse cellular systems. Nevertheless, genome-scale characterization of translational changes has lagged behind the development of similar methods for exploring mammalian transcriptomes. Recent advances in profiling ribosome occupancy with deep sequencing (Ingolia et al., 2009), however, now allows a near codon resolution view of ribosome-bound mRNAs. Briefly, ribosome profiling entails isolating fragments of such mRNAs, which results in ~30 bp ribosome-protected fragments (RPFs) that are purified and processed for massively parallel sequencing. Simultaneously, and from the same cells, poly(A)+ mRNA is purified, and concurrent sequencing of both ribosome and RNA libraries generates short sequence reads that identify either the position of a bound ribosome or the expression of the cognate transcript. These methods have been used to explore translation in diverse genomes including those from yeast, zebrafish, murine models and human cancers (Bazzini et al., 2012; Guo et al., 2010; Hsieh et al., 2012; Ingolia et al., 2011; Thoreen et al., 2012).

These studies have revealed a complex and multifaceted translational landscape in mammalian transcriptomes. Although open reading frames (ORFs) that initiate with an AUG codon are canonical and are perhaps best characterized, a variety of non-canonical structures also contribute to the diversity and complexity of translation. These include upstream ORFs, non-AUG initiation of translation, internal ribosome entry sites, translational re-initiation and frameshift, leaky scanning and the ever-expanding world of small RNAs: expressed short and long non-coding sequences that lack an ORF.

So, although ribosome profiling promises a new window into the nature and dynamics of normal and anomalous mRNA translation, progress is predicated on the development of robust methodologies to facilitate statistically principled analysis. Prior analyses have described the global features of translational efficiency in different systems, species and upon multiple perturbations by comparing the log of scaled ribosome counts to the log of scaled mRNA counts for groups of genes (Bazzini et al., 2012; Guo et al., 2010; Hsieh et al., 2012; Ingolia et al., 2009, 2011; Thoreen et al., 2012). However, it is unclear how such a measure would lead naturally to inference about the biological significance of individual changes in genes, especially in the presence of few replicates. There is, therefore, a great need for robust methods for drawing gene-level inference on the basis of changes in translational regulation or even subtle shifts in patterns of ribosome occupancy. The importance of such quantitative methods will only grow as ribosome profiling is adopted more widely to characterize translational abnormalities that drive diverse pathologies including human cancers. Such efforts would serve...
as a vital companion to extensive genetic characterization efforts already underway, such as The Cancer Genome Atlas (http://cancergenome.nih.gov/).

Here, to facilitate such efforts, we describe a statistical framework, **Babel**, for determining genes with unexpected ribosome occupancy: those insufficiently explained by their intrinsic transcript abundance within a condition, and those whose ribosome association changes in magnitude or direction between conditions (Fig. 1A). We demonstrate its statistical properties through simulations, and its applicability through the re-analysis of ribosome profiling data interrogating translational control downstream of the mTOR signaling pathway.

**2 METHODS**

**2.1 Outline of methods**

The relationship between the level of ribosome occupancy and mRNA expression transcriptome-wide can be described quantitatively from sequencing data. We sought to exploit the monotonic relationship between ribosome association and mRNA levels as the basis for the gene-level inference described here (Fig. 1B). For a transcript with a given abundance, the expected level of ribosome occupancy can be inferred so that translationally regulated genes are those for which the level of bound ribosome deviates significantly from that expected value. To estimate this expected level, we developed an errors-in-variables regression model. Unlike typical regression, our model treats the mRNA level (predictor) as measured with error rather than as fixed. As a first step, mRNA levels are modeled by the negative binomial (NB) distribution, as is convention (Robinson and Smyth, 2008), because the variance of the counts is greater than the mean. As a second step, because ribosome occupancy involves counts for which there is again extra Poisson variation (Fig. 1C), we also model the level of bound ribosome given mRNA abundance as NB. To estimate the mean in the second part of the model, we tested multiple regression forms (Fig. 1B). A trimmed least-squares approach (for which a fraction of genes with outlying mRNA levels were excluded before model fitting) was most stable over all experiments (data not shown). This second over-dispersion parameter is modeled using an iterative algorithm to prevent overestimation. Subsequent inference under
Fig. 2. Combining multiple tests. (A) Fisher’s method for combining two independent tests. (B) Our alternative approach based on the arithmetic mean of P-values. All equally significant pairs define a line passing through (d, 0) and (0, d), whereas more significant pairs correspond to a line with d’ < d (dark blue). Thus, the combined P-value corresponds to the area of the triangle. (C) One-sided P-values are used; therefore, the opposite triangle is used for pairs of consistently large P-values. (D) A practical example of the difference between approaches.

2.2 Sequencing and alignment

The ribosome profiling strategy, described elsewhere (Ingolia et al., 2009), produces short single-end sequencing reads of between 36 and 50 bp. Data analyzed here were generated from libraries prepared and sequenced as previously described (Hsieh et al., 2012) and downloaded from the NCBI Gene Expression Omnibus under the accession number GSE35469. Because the ribosome-bound mRNA fragment may be a variable-length component of each sequencing read followed by a known adapter sequence (CGTATGGCACCTCATGAA), raw sequencing reads in FASTQ format were first clipped of the adaptor and remaining 3’ sequence with the FASTX-Toolkit, retaining only reads of 24 bp post-clipping length or longer. Reads were mapped to the human assembly (NCBI build 36) with Burrows-Wheeler Alignment (ver. 0.5.9-r16, default parameters) (Li and Durbin, 2009). Unaligned reads were then mapped to known splice junctions with TopHat (ver. 1.4.1) (Trapnell et al., 2009) using a transcriptome index created from version 11 of the Gencode gene set (Harrow et al., 2006). Provisional merged BAM files (Li et al., 2009) were created from aligned and unaligned reads from either the genome or spliced alignment, and read groups were enforced across lanes and experiments. BAM files were then indexed, coordinate sorted and had alignment metrics determined, all with the Picard suite (http://picard.sourceforge.net/).

2.3 Gene-level assessment

For all gene-level inference, only those protein-coding transcripts of levels 1 or 2 of Gencode support (verified or manually curated loci, respectively) were used. We initially assessed two candidate gene models. The first retained only the longest transcript of multi-isofrom genes. The second was a unified gene model in which the union of coding sequence across all isoforms of a given gene was collapsed into a single model per gene. As no significant difference in read mapping or expression (as reads per kilobase of exon model per million mapped reads; rpkM) was observed between these two models, all subsequent analyses used the unified model. The gene-level inference of ribosome-given-mRNA levels described here evaluated only those genes that we estimated to be expressed in the studied transcriptome. To determine expressed genes, we first mapped aligned reads to their coding sequences, treated spliced reads between exons as a single fragment and excluded ambiguous reads (those aligning to overlapping coding sequence of two distinct genes: 1.52–2.79% of coding reads to their coding sequences, treated spliced reads between exons as a single fragment and excluded ambiguous reads (those aligning to overlapping coding sequence of two distinct genes: 1.52–2.79% of coding

2.4 Errors-in-variables regression of ribosome association on mRNA abundance

The NB distribution is used throughout the Babel statistical model (see Fig. 1). One parameterization of the NB distribution is with mean λ and variance λ(1 + λφ), where φ represents the parameter for over-dispersion (variance above what would be expected from the Poisson distribution). When φ is zero, the variance is equal to the mean, and the NB distribution is the same as the Poisson distribution.

All analyses conducted here, unless otherwise specified, were performed using the R statistical language, version 2.11.
Let \( x_{g,i} \) and \( y_{g,i} \) represent the mRNA level for the \( g \)th gene and the number of reads for the \( g \)th gene, respectively. We assume that these values follow a normal distribution with mean \( x_{g} \) and variance \( s_{g}^2 \), where \( x_{g} \) is estimated from the observed data and \( s_{g} \) is estimated from the null hypothesis. We use the parametric bootstrap and the modeling described previously to estimate \( \hat{\phi} \), the optimal value for the over-dispersion parameter. For every gene from the NB distribution, \( x_{g} \) is estimated from the observed data and \( \hat{\phi} \) is estimated by minimizing the squared error across bins:

\[
\hat{\phi}_g = \arg \min_{\phi} \sum_{i=1}^{N} (y_{g,i} - \phi)^2.
\]

Because there exists a set of genes for which ribosome levels are different than expected from their intrinsic mRNA levels, which are in fact the very genes we are trying to identify, we developed an iterative process to avoid overestimating \( \hat{\phi} \). Here, we define an initial value \( \hat{\phi}_0 = 0.2 \) and eliminate genes in the variance calculations with \( P \)-values < 0.01 under the NB model with mean equal to the product of \( \hat{\phi} \) and the median mRNA abundance in the bin. To calculate the final \( \hat{\phi} \), this process is repeated, seeded with \( \hat{\phi}_0 \) (again, excluding outlier genes from the calculation of variance).

We use the parametric bootstrap and the modeling described previously to estimate \( P \)-values under the null hypothesis that the ribosome level is as expected from the mRNA level. Here, we begin by simulating mRNA values \( x_{g}^{1}, \ldots, x_{g}^{N} \), for every gene from the NB distribution with mean \( x_{g} \) and over-dispersion \( \hat{\phi} \). Then \( y_{g}^{1}, \ldots, y_{g}^{N} \) is simulated once from a NB distribution with mean \( x_{g} \hat{\phi} \) and over-dispersion \( \hat{\phi} \), where \( \hat{\phi} \) and \( \phi \) are as described earlier. A one-sided \( P \)-value is calculated from these simulations as:

\[
P_{g} = \frac{\sum_{k=1}^{N} I(y_{g,k} > y_{g})}{N},
\]

where \( I \) is the indicator function. This corresponds to the proportion of simulated read counts for bound ribosome greater than the observed count. Low \( P \)-values imply ribosome levels are greater than expected from a gene’s mRNA level, whereas high \( P \)-values imply the reverse. If there is only a single experiment, two-sided \( P \)-values are obtained as:

\[
P_{g} = 2 \min(1 - p_{g}, p_{g}).
\]

These \( P \)-values are corrected for multiple comparisons using the \( q \)-value method of Storey (Storey, 2002). Although the focus so far has been on the estimate of significance in a single sample, the next section details how to estimate both one- and two-sided \( P \)-values when there is more than one experiment in a condition.

### 2.5 Combining \( P \)-values within a condition

For every condition with two or more replicates, we would like a single consensus \( P \)-value to quantify how unexpected the ribosome levels are given the mRNA abundance of the gene in that condition. A common approach for combining independent \( P \)-values is Fisher’s method, which uses a test statistic \( S \):

\[
S = \sum_{i=1}^{n} -2 \log(p_{i}),
\]

where \( S \) is distributed as \( \chi^2 \) with \( 2n \) degrees of freedom. Here, \( S \) is a function of the product of the independent \( P \)-values, and thus it is proportional to the geometric mean of the \( P \)-values. Fisher’s method is sensitive to single \( P \)-values. We have used and developed an application of alternative methodology proposed by Edgington (Edgington, 1972). The method, which is based on the arithmetic mean of \( P \)-values, estimates a combined \( P \)-value that is more appropriate than Fisher’s method in this context because it is less sensitive to single outlier \( P \)-values. In the case of two replicates \( i \) and \( j \) in a given condition, we let \( d = p_{i} + p_{j} \) for \( p_{i} \leq p_{j} \). This defines a line that passes through \((d, 0)\) and \((0, 0)\) as shown in Fig. 2. Here, all pairs of \( P \)-values satisfying this condition are considered equally significant, whereas pairs satisfying \( d < d \) are considered more significant. Therefore, all significant pairs of \( P \)-values trace out a triangle to the left of \( d \). For \( p_{i} + p_{j} \leq 1 \), the combined \( P \)-value is thus:

\[
P = 0.5(p_{i} + p_{j})^{2},
\]

whereas for \( p_{i} + p_{j} > 1 \),

\[
P = 0.5 + (p_{i} + p_{j} - 1) \times \left( 1 - \frac{(p_{i} + p_{j} - 1)}{2} \right).
\]

The first term corresponds to the triangle for \( d = 1 \) and the second corresponds to the trapezoid to the right of this line. Critically, input for this method is the one-sided \( P \)-values of ribosome level on mRNA, such that pairs of small one-sided \( P \)-values give small combined \( P \)-values and pairs of large one-sided \( P \)-values give large combined \( P \)-values.

Additionally, this methodology can be generalized to any number of dimensions using the same principles. The assumption, however, underlying the appropriate sampling distribution used to estimate \( P \)-values is difficult to compute. It is the distribution of the sum of independent, identically distributed random numbers, each uniformly distributed on the unit interval \([0, 1] \). This is known as the Irwin–Hall distribution (Hall, 1927; Irwin, 1927), where the density of \( d \) is:

\[
f_{d}(\delta) = \frac{1}{(n-1)!} \sum_{k=0}^{n} (-1)^{k} \frac{n!}{k!} (\delta - k)^{n-1}.
\]

Integration gives the cumulative distribution function of \( d \):

\[
F_{d}(\delta) = \frac{1}{n!} \sum_{k=0}^{n} (-1)^{k} \frac{n!}{k!} (\delta - k)^{n}.
\]

The one-sided \( P \)-value is then \( P = F_{d}(\delta) \). A two-sided combined \( P \)-value can be obtained as before (3) and corrected for multiple comparisons as described previously.
2.6 Assessing changes in translational regulation between conditions

To compare the translational regulation for every gene between pairs of conditions, we test for equality of the within-condition P-values, which allows mRNA levels to change while maintaining a valid test. P-values are converted to standardized Z-statistics using the Gaussian quantile function $G^{-1}$:

$$z = G^{-1}(p).$$

Let $p_{ij}$ represent the one-sided P-value for ribosome relative to mRNA level for the $i$th condition and $j$th replicate for one gene. Assuming two replicates, the corresponding Z-statistics are $z_{ij}, i = 1, 2, j = 1, 2$. The test statistic for every gene $g$ is therefore:

$$t_g = \frac{(z_{11} + z_{12}) - (z_{21} + z_{22})}{\sqrt{s_g}},$$

where $s_g$ is the estimated standard error of the numerator. Accurately estimating $s_g$ from a small number of values for every gene is challenging. Although previous methods developed for the analysis of small microarray studies have used moderated approaches (Smyth, 2004), we believe insufficient information exists within genes to derive an accurate estimate in this way. Therefore, we derive a single $s_g$ across genes. We calculate the variance-covariance matrix for the $Z$s, which leads to the estimate:

$$s_g = \frac{\sum_{i,j=1}^{2} s_{ij}^2}{2} + 2(s_{11,12} + s_{21,22}) - 2(s_{11,21} + s_{11,22} + s_{12,21} + s_{12,22}).$$

where the first term is for the variances and the second and third terms are for the covariances. Generally, if there exists other than two replicates, all within condition covariance terms are added, whereas all between condition covariance terms are subtracted. Now $s_g$ would be estimated ideally from only the genes where there has been no change between conditions, but these genes are unknown. Therefore, if there are $m$ total samples in the test, we screen out genes with numerators corresponding to Equation (10) greater than $3\sqrt{m}$ before estimating $s_g$.

Test statistics (10) can then be converted into two-sided P-values by comparison with the standard Gaussian distribution and corrected for multiple comparisons as described previously.

We note that although genes are correlated across conditions, this test uses only the P-values within conditions. Nevertheless, the denominator of the test includes the covariance of the statistics in the numerator, so this correlation is incorporated. Additionally, a more natural test may use ribosome occupancy counts directly instead of the Gaussian quantile transformation of the P-values that we use here. Nevertheless, because of our use of the errors-in-variables model and the skewness of the NB distribution, there was no clear formulation to us of such a test. We did examine test statistics based on differences and log ratios and none were used. Generally, if there exists other than two replicates, all within condition covariance terms are added, whereas all between condition covariance terms are subtracted. Now $s_g$ would be estimated ideally from only the genes where there has been no change between conditions, but these genes are unknown. Therefore, if there are $m$ total samples in the test, we screen out genes with numerators corresponding to Equation (10) greater than $3\sqrt{m}$ before estimating $s_g$.

Test statistics (10) can then be converted into two-sided P-values by comparison with the standard Gaussian distribution and corrected for multiple comparisons as described previously.

We also evaluated the power of Babel by altering 50 genes at multiple levels of mRNA expression, increasing or decreasing the ribosome association between 2- and 4-fold. Within a given condition, Babel had similar power for increases and decreases across different levels of the mRNA distribution, except for the expected reduction in power for identifying decreases of any magnitude in ribosome association at low levels of mRNA expression. At $\alpha = 0.05$, mRNAs with a 4-fold increase or decrease in ribosome association in two replicates were identified 87.4 and 92% of the time, respectively, in the middle 90% of the mRNA abundance distribution (Fig. 3B, see Supplementary Table S2). As expected, increasing either the fold change of ribosome-given mRNA levels or the number of replicates increased the power in all cases. For testing between conditions, additional replicates are needed to achieve similar power to the within test. For the
Results across simulated datasets indicate the type I error rates in Babel are as expected. The power of the Babel framework for detecting translationally regulated genes (2–4-fold increase in ribosome association) as a function of the number of replicates at significance levels of 0.05 and 0.001 among genes in the middle 90% of the distribution of mRNA abundance.

To assess the performance of Babel, we analyzed translational changes downstream of mTOR signaling. Ribosome profiling was previously performed on the PC3 prostate cancer cell line upon perturbation with one of two pharmacological inhibitors of mTOR: either the allosteric inhibitor rapamycin or the stringent adenosine triphosphate (ATP) site inhibitor PP242 (Hsieh et al., 2012). Each experiment was performed in duplicate as a biological replicate, and control experiments were performed with DMSO.

The original analysis identified 144 mRNAs that were decreased selectively in ribosome association upon PP242 inhibitor treatment. Of these, 114 were analyzed here (because of differences among gene definitions; see Methods). Notably, at \( \alpha = 0.05 \), Babel analysis identified 90 of these 114 genes from the original analysis (79%); \( P \)-value \( \sim 0 \), McNemar’s test), indicating strong concordance (Fig. 4A). These included a large number of the components of the translation machinery and a functionally validated pro-invasion gene signature, both consistent with upstream inhibition of mTOR signaling (Hsieh et al., 2012). Among 24 presumptive false negatives, those identified by the previous study but not significant in Babel, some had high variability between replicates (Supplementary Fig. S2), whereas others varied because of differences in the gene models used by the two analyses (see Methods).

Fig. 3. Characterizing Babel performance. (A) The proportion of \( P \)-values below the given cutoffs as a function of the number of replicates for simulated data in which there are no translationally regulated genes. Results across simulated datasets indicate the type I error rates in Babel are as expected. (B) The power of the Babel framework for detecting translationally regulated genes (2–4-fold increase in ribosome association) as a function of the number of replicates at significance levels of 0.05 and 0.001 among genes in the middle 90% of the distribution of mRNA abundance.

As mentioned, 114 genes were analyzed here (because of differences among gene definitions; see Methods). Notably, at a \( q \)-value cutoff of 0.25, Babel analysis identified 163 genes with a significant increase or decrease in ribosome association upon robust mTOR inhibition with PP242. Of those, 100 were not previously observed (Supplementary Table S3). Although many of these new findings were attributable to the broader range of mRNA expression Babel uses to credential genes for analysis, others revealed specific facets of mTOR regulation of translational control. Babel appeared particularly effective at identifying networks of functionally related genes. One such inter-connected module was comprised of multiple Rab family Ras-related GTPases involved in endocytosis. Although both the original study and Babel identified PP242-inhibitor related translational regulation of \( RABGGTC \), a Rab geranylgeranyltransferase, Babel also identified \( RAB3A \) and \( RAB13 \), indicating mTOR signaling may play a role in the translational regulation of multiple molecular components of endocytic trafficking.

Among core functional modules, Babel analysis identified five subunits of the eukaryotic initiation factor 3 (eIF3) that were significantly reduced in ribosome association upon mTOR-kinase inhibition (Supplementary Table S4). Although two non-conserved subunits eIF3l and eIF3h were identified previously (Hsieh et al., 2012; Thoreen et al., 2012), we found three additional components that had lower than expected ribosome-given-mRNA levels upon mTOR inhibition with PP242 that were absent upon rapamycin treatment, an effect that was eIF3-specific (Fig. 4B). Although the precise biochemical functions of eIF3 remains to be determined, it is known that eIF3 interacts with eIF4G to mediate the assembly of the 43S ribosome on individual mRNAs and its depletion reduces translation. Notably, all of the translationally repressed subunits, both conserved and non-conserved, have a direct physical interaction (Fig. 4B, inset), four of which (eIF3a, f, h and e) are also constituents of the functional core essential for eIF3 complex formation (Masutani et al., 2007; Siridechadiok et al., 2005; Zhou et al., 2008). These findings support a direct contribution of eIF3 loss to PP242, but not rapamycin-mediated inhibition of translation. Additionally, subunits dispensable for active eIF3 complex formation (Masutani et al., 2007) were also translationally repressed, indicating a regulatory function may
exist beyond complex formation that is controlled at the translational level rather than at the transcriptional level in response to upstream mTOR signals. Together, these results demonstrate eIF3 may be a key downstream effector of mTOR–kinase-mediated inhibition of translation.

4 DISCUSSION

The ability to draw statistically principled inferences about translationally regulated genes from ribosome profiling is necessary for the discovery of specific mechanisms of, and abnormalities in, translational control in diverse cellular systems. Here, we introduce the Babel framework for performing just such an analysis. Babel produces a natural and readily interpretable summary (P-value) of the level of ribosome-given-mRNA for individual genes within samples and between conditions. On simulated data, we demonstrated that Babel has proper type I error both within and between conditions and possesses good power that increases with sample and effect size, emphasizing the need for careful design of ribosome profiling experiments. A reanalysis of ribosome profiling in response to multifaceted mTOR inhibition both confirmed prior findings, and importantly, extended them to identify novel functional modules subject to translational control in prostate cancer cells.

Among the greatest strengths of Babel is model flexibility. As more is learned about the layers of overt post-transcriptional regulation that impact ribosome occupancy and may uncouple the abundance of a transcript from its level of active translation, quantitative measures of these underlying processes can be incorporated as covariates to improve the accuracy of the model. Babel also introduces a new approach for combining P-values across independent tests in this context that has applicability far beyond the analysis of ribosome profiling. Nevertheless, challenges remain. Our simulations and experience indicate that modeling RPFs for genes with low mRNA abundance (and, therefore, low RPFs) can produce inaccurate P-values based on what may be small stochastic fluctuations in read counts between samples (see Supplementary Note). Although beyond the scope of this work, improving this will require a joint estimate of RPF and mRNA counts sufficient to produce accurate P-values from the bootstrap in the left tail of the mRNA distribution. Furthermore, additional improvements to the ribosome profiling protocol are necessary to extend these analyses to isoform-specific translational patterns.

Despite these challenges, Babel establishes a statistical framework for inferring changes in translational regulation of cellular mRNAs. In conjunction with ribosome profiling, the Babel framework will help bridge the gap in our understanding of the translational control of normal and pathogenic phenotypes.

Funding: Prostate Cancer Foundation (PCF) Young Investigator Awards (to A.C.H. and B.S.T.); Burroughs Wellcome Fund Career Award for Medical Scientists (to A.C.H.); David H. Koch Foundation (to B.S.T.); National Institutes of Health [4R37EB002784-35 to R.A.O.; F32CA162634 to C.R.S.; 1K08CA175154-01 to A.C.H.; P30 CA82103 to A.B.O.; K01CA154916 and R01CA140456 to D.R.]; A.C.H. is a postdoctoral fellow of the American Cancer Society. D.R. is a Leukemia and Lymphoma Society Scholar.

Conflict of Interest: none declared.

REFERENCES

Hall,P. (1927) The distribution of means for samples of size n drawn from a population in which the variate takes values between 0 and 1, all such values being equally probable. Biometrika, 19, 240–245.
Harrow, J. et al. (2006) GENCODE: producing a reference annotation for ENCODE. *Genome Biol.*, 7(Suppl. 1), S4.1–S4.9.


Irwin, J.O. (1927) On the frequency distribution of the means of samples from a population having any law of frequency with finite moments, with special reference to Pearson’s Type II. *Biometrika*, 19, 225–239.


