An improved analysis methodology for translational profiling by microarray

THOMAS SBARRATO,[1,2,3,4] RUTH V. SPRIGGS,[1] LINDSAY WILSON,[1] CAROLYN JONES,[1] KATE DUDEK,[1] AMANDINE BASTIDE,[1] XAVIER PICHON,[1] TUIJA PÖYRY,[1] and ANNE E. WILLIS[1]

[1]Medical Research Council Toxicology Unit, Leicester LE1 9HN, United Kingdom
[2]Aix Marseille Université, LAI UM 61, Marseille F-13288, France
[3]Inserm, UMR_S 1067, Marseille F-13288, France
[4]CNRS, UMR 7333, Marseille F-13288, France

ABSTRACT

Translational regulation plays a central role in the global gene expression of a cell, and detection of such regulation has allowed deciphering of critical biological mechanisms. Genome-wide studies of the regulation of translation (translatome) performed on microarrays represent a substantial proportion of studies, alongside with recent advances in deep-sequencing methods. However, there has been a lack of development in specific processing methodologies that deal with the distinct nature of translatome array data. In this study, we confirm that polysome profiling yields skewed data and thus violates the conventional transcriptome analysis assumptions. Using a comprehensive simulation of translatome array data varying the percentage and symmetry of deregulation, we show that conventional analysis methods (Quantile and LOESS normalizations) and statistical tests failed, respectively, to correctly normalize the data and to identify correctly deregulated genes (DEGs). We thus propose a novel analysis methodology available as a CRAN package; Internal Control Analysis of Translatome (INCATome) based on a normalization tied to a group of invariant controls. We confirm that INCATome outperforms the other normalization methods and allows a stringent identification of DEGs. More importantly, INCATome implementation on a biological translatome data set (cells silenced for splicing factor PSF) resulted in the best normalization performance and an improved validation concordance for identification of true positive DEGs. Finally, we provide evidence that INCATome is able to infer novel biological pathways with superior discovery potential, thus confirming the benefits for researchers of implementing INCATome for future translatome studies as well as for existing data sets to generate novel avenues for research.

Keywords: translational regulation; polysome profiling; translatome; microarray analysis

INTRODUCTION

Microarrays have been widely used in past decades in order to identify genes differentially expressed under varying conditions. Landmark studies have allowed the identification of prognostic markers and disease subtypes (van 't Veer et al. 2002; Sorlie et al. 2003). The approach relies on drawing inferences from a set of biological samples, i.e., deducing properties of a given condition by testing hypotheses on a sampled population. The biggest advantage of this technique, the comprehensive gene set, is also its main pitfall as the imbalance between a large number of genes and a handful of biological samples transforms the analysis step into the bottleneck of the microarray technology (Yang and Speed 2002). Furthermore, technical variations (dye bias and array variability) and biological noise (signal variations) challenge accurate analysis (Bala et al. 2010).

To improve the robustness of data, analysis standards have been defined in order to minimize inferential systematic errors represented by false positives (type I errors) and false negatives (type II errors) and to measure error rate and confidence. Briefly, this entails (i) pre-processing steps (background correction) to adjust intensity variations due to hybridization defects, (ii) data normalization to remove systematic bias from the data and across the samples, and (iii) statistical testing of a given hypothesis to identify deregulated genes (DEGs) (Dudoit et al. 2002; Huber et al. 2002; Smyth et al. 2002). Importantly, the normalization procedures are an essential first step to correct for systematic errors, with an intra-array normalization to adjust for spatial or intensity variations and an inter-array normalization to ensure a constant variance across the dynamic range. Several normalization methods have been proposed, largely nonparametric and based on transformation by regression,

Corresponding authors: thomas.sbarrato@inserm.fr, rvs3@leicester.ac.uk

Article is online at http://www.rnajournal.org/cgi/doi/10.1261/rna.060525.116. Freely available online through the *RNA* Open Access option.

such as, e.g., LOESS (smoothing by local regression) or Quantile (quantile normalization toward identical distribution) (Yang et al. 2002; Bolstad et al. 2003). In a second step, a statistical test is applied to ascertain the working hypothesis (differential response with treatment). Numerous statistical methods have been suggested that take either a linear models or rank-based (nonparametric) approach. These include: modified *T*-Test, Rankprod, SAM, and LIMMA (Kerr et al. 2000; Thomas et al. 2001; Lin and Zou 2004; Smyth 2004).

The choice of approach to use for any given experiment can be difficult, and the assumptions on which methodologies depend, while clearly set out in original publications, are often overlooked by users (Dabney and Storey 2007). Indeed, these methods assume that (i) only a small fraction of genes are significantly deregulated, and (ii) this deregulation occurs symmetrically around zero (i.e., there is an equal number of up- and down-regulated genes) (Yang et al. 2002). While these assumptions may hold true for transcriptional profiling (for which all these methods were initially developed), they can be easily violated with skewed experiments (due to the study design and/or to the nature of the biological samples) such as enrichment experiments (e.g., CHromatin Immuno Precipitation) or sensitive systems where deregulation is global and/or asymmetrical, such as in many studies of the regulation of translation (changes in the "translatome") (Jeanmougin et al. 2010; Landfors et al. 2011). When these assumptions are violated, general microarray methods will introduce inferential errors leading to spurious deregulation (false positives) and/or censored biological changes (false negatives) (Zhao and Pan 2003; Dabney and Storey 2007). By attempting to correct data bias on skewed experiments, microarray methods may alter the biological profile of interest and ultimately guide the user toward false leads in terms of validation and biological significance. Consequently, one can extrapolate that many experiments may have failed to identify and/or validate correctly the treatment response due to flawed data analysis and leave unexplored several avenues of investigation.

There is consequently a long-standing need for alternative methods where skewness is representative of the biological context and needs to be maintained throughout the analysis to retain the true differential expression upon treatment. This can be achieved by using normalization approaches that make appropriate use of within-array housekeeping genes (selected by the user), internal controls (introduced on the array), or Spike-In (added to the samples) to scale the arrays and to compare the observed values to expected ones. Examples of these applications are the use of Y chromosome-linked genes, bacterial operons, or commercial Spike-In (Li and Wong 2001; Galfalvy et al. 2003; Irizarry et al. 2003a,b; Harr and Schlotterer 2006; Pelz et al. 2008). Other published strategies include tools to detect skewness and propose new normalization methods (Dabney and Storey 2007; Landfors et al. 2011).

Here, we have examined and compared the different methods available to analyze any microarray data in which there is a strong skew, i.e., a strong bias toward up- or down-regulation of the data set, at the two critical steps of normalization and statistical testing. As a model, we have used a system measuring global change in mRNA translational regulation termed translational profiling. Perturbation of biological systems, particularly involving disruption of the cell cycle or induction of DNA damage, can lead frequently to global translatome change, as the cell adjusts to dramatic changes in growth rate and protein synthesis demand (Polunovsky and Bitterman 2006; Spriggs et al. 2010): In such cases, determining on a genome-wide scale which changes are key to controlling particular processes can be challenging and requires confidence in the statistical methodology. Improving analysis of both new and existing data sets is of potentially great value, since a number of studies have provided evidence for widespread translational deregulation upon various treatments as well as in the etiology and progression of cancer (Le Quesne et al. 2010; Blagden and Willis 2011). And yet, to date, apart from a small number of studies seeking to refine methodology to analyze specific combinations of deep-sequencing transcriptome and translatome studies (Ingolia 2010, 2016; Larsson et al. 2010), there has been no robust comparison of the different methods available for the analysis of translational profiling by microarray.

We present simulated microarray data to show for the first time that, under common conditions, polysome profiling is a skewed experiment refractory to general microarray analysis methods. We then devised a novel approach for normalization of translatome studies, internal control analysis of translatome (INCATome), and demonstrated that this delivers the best performance across the spectra of parameters tested. Next we propose a new statistical workflow to avoid interference of data skewness in the identification of DEGs. Simulated results were validated in a biological data set comparing untreated HeLa cells to those in which the protein PSF, critical for mRNA splicing and DNA damage repair, has been knocked down. Overall, we show that our novel method allows for more accurate characterization of translatome change, and as an example present interesting findings relating to cell response to PSF silencing. Importantly, the refined methodology set out here may allow existing data sets to be reanalyzed to reveal significant new insights for follow-up.

RESULTS

Violations of method assumptions in translatome studies by microarrays

Commonly used microarray analysis tools rely on several assumptions, for instance low percentage (PDE) and equal symmetry (SYM) of deregulation, that may easily be violated, particularly when microarray is applied to systems (such as the translatome) subject to rapid and large changes (Dembélé 2013). Separation of actively translating mRNAs is naturally skewed by the capacity of the biological system to respond quite drastically to perturbations in terms of PDE and SYM. For instance, a transient depletion of splicing factor PSF (*SFPQ* gene) in HeLa cells caused a marked reduction in the amount of actively translating ribosomes (Fig. 1Ai, fractions 6–11). Similarly, the amount of RNA purified from each fraction of the polysome separation showed a reduction of 20% in collected material associated with polysomes, amounting to a global fold change of 0.43 (Fig. 1Aii). Thus, the changes in proportion of active translatome confer a non-negligible skew when comparing translation between conditions (Fig. 1Aii). Based on these facts, we developed a new analysis pipeline for translatome studies by microarray, combining a new normalization method (INCATome) permissive to extreme PDE and SYM conditions, as well as an improved methodology for identification of DEGs.

Our novel approach to translatome analysis is based on the root mean square deviation (RMSD) of internal controls (Supplemental Fig. 1). These can be represented by either the use of spike-in controls that are independent of the sample and of known concentrations or by the use of Internal References chosen by the user and experimentally validated. The main advantage of this implementation is that the expected values for these given probes are already at hand to the user before the experiment is performed [INCATome (SI)—spike-in expected values given by spike-in concentration ratios or INCATome(IR)—internal references expected values given by at least two Northern blotting/qPCR quality controls for subpolysomal and polysomal associations i.e., ACTB and PABP, respectively]. Importantly, the user can choose any given mRNA as internal reference for the procedure, as long as the expected polysomal distribution has been experimentally validated. Thus, in the case where treatments/conditions might change the polysomal association for these reference mRNAs, the experimental validation will inform the user of the expected fold change and thus be correctly taken into account during the INCATome procedure. As a consequence, the RMSD values can be computed between expected and observed values for these probes in order to normalize the data. This procedure results in a within sample normalization (to the expected levels of the given INCATome probes for each sample) as well as a general scaling method across the samples (all tied to the same set of INCATome probes).

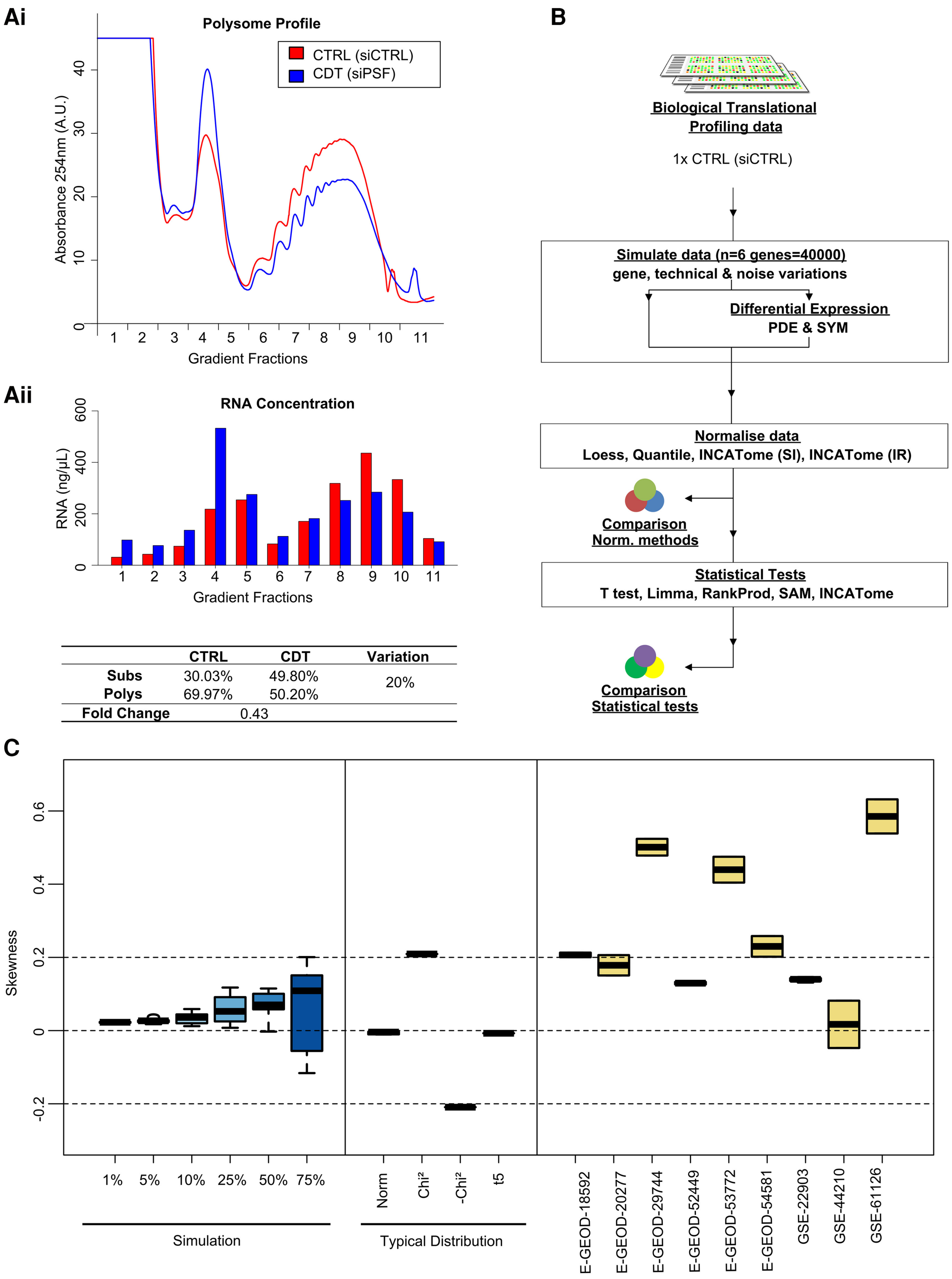
To assess the performance of INCATome, we simulated translatome data sets from a true biological sample (siCTRL in HeLa cells) with varying degrees of PDE and SYM (Fig. 1B). Based on modifications of Dembélé (2013), we generated six samples (three controls and three conditions) of 40,000 gene probes, while varying several parameters: PDE (1, 5, 10, 25, 50, and 75%), SYM (0 to 1 by 0.1 increments) and noise (0.1 low noise—data not shown—and 0.4 high noise) (Supplemental Fig. 2). All simulated DEGs were identified as true expected positives upon implementation of artificial fold change during simulation. A detailed pipeline can be found in the Materials and Methods section and in Supplemental Figure 3. To characterize the resulting simulated data sets, we studied the distribution characteristics (skewness, excess kurtosis, and tail heaviness) of the different simulations as well as standard distributions (normal distribution "Norm," skewed χ^2 distribution with three degrees of freedom "χ^2" and heavy tailed *t* distribution with five degrees of freedom "$t_{(5)}$"). With increasing PDE, the skewness of simulated data deviated from a normal distribution toward the non-normal distributions, with less effect on kurtosis and tail heaviness (Fig. 1C; Supplemental Fig. 4). Importantly, the simulated data sets exhibited intermediate skewness, kurtosis, and tail heaviness compared to published translatome data sets on microarrays from NCBI GEO and EBI ArrayExpress databases (Edgar et al. 2002; Kolesnikov et al. 2015).

Overall, we have shown that translatome studies result in genuinely skewed distributions by nature and violate general assumptions imposed by conventional microarray analysis schemes. By generating simulated data presenting intermediate skew compared to the deviation that can be found in published biological data sets, we chose to not overestimate the skewness and as a consequence set the simulation in a fair and realistic environment.

INCATome normalization outperforms other methods especially in extreme PDE and SYM settings on simulated data

To study the performance of INCATome in comparison to commonly used methods (Quantile and LOESS normalizations) for correct identification of DEGs in translatome studies, we subjected the simulated data sets to each normalization. Figure 2A shows that in extreme conditions (down-regulation SYM = 0.1 and up-regulation SYM = 0.9), Quantile and LOESS normalizations skew the fold-change distribution in cases of high percentage of deregulation (PDE 75%). On the contrary, both implementations of INCATome maintained the inherent skew due to the atypical distribution of translational fold change. More globally over the range of simulations, both Quantile and LOESS normalizations try to attenuate the inherent skewness and create additional tail heaviness for the distributions, unlike INCATome methods (Supplemental Fig. 5). These findings are confirmed when the simulated data are initially skewed by a large technical variation (Supplemental Fig. 6). Only INCATome implementations are able to correct for technical variation without inducing an unwanted skew.

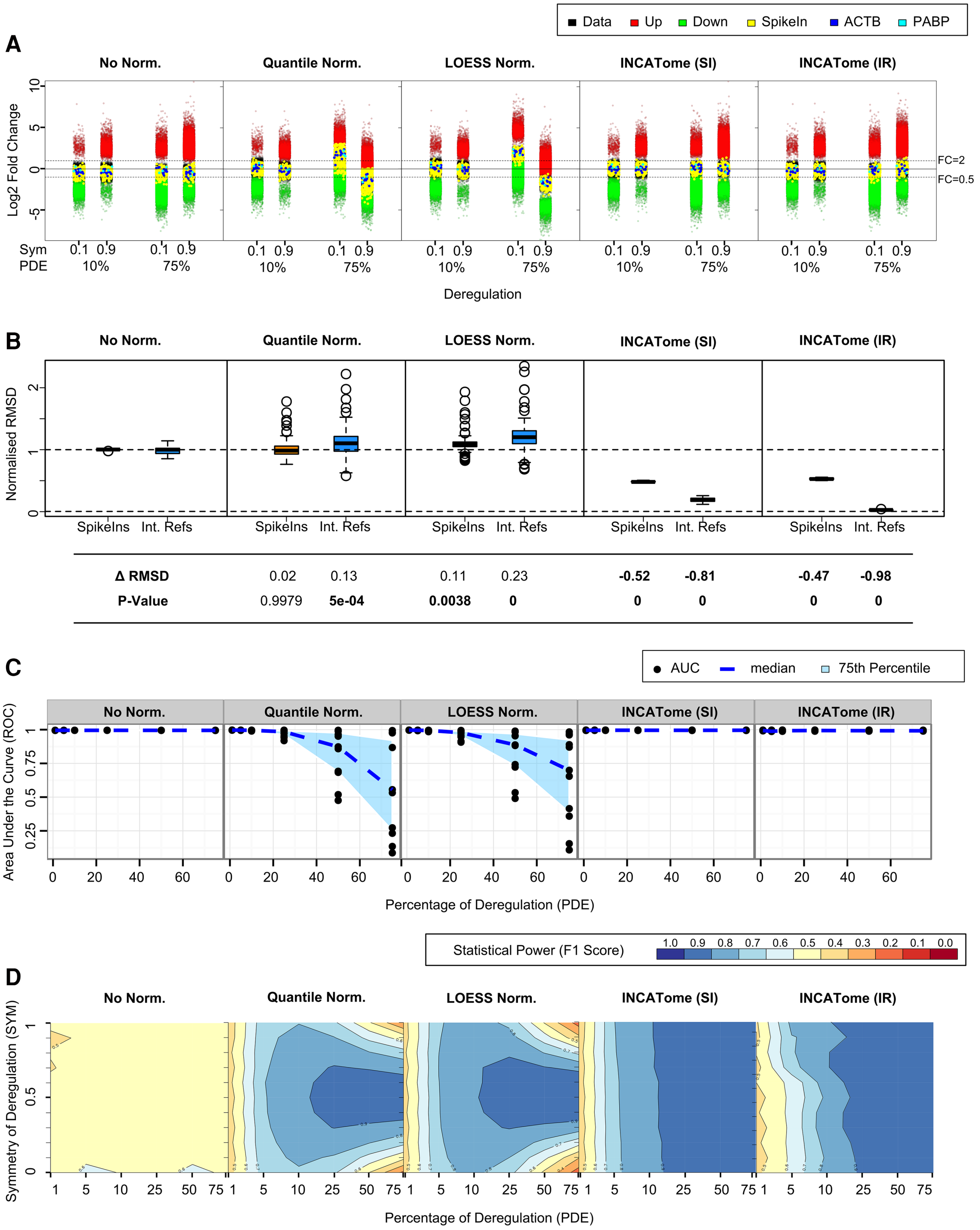
To further characterize the response of each normalization method in relation to known internal controls, we assessed the RMSD of either spike-in probes or internal reference probes (ACTB and PABP). Quantile and LOESS normalizations generated significantly more dispersion in RMSD than



	CTRL	CDT	Variation
Subs	30.03%	49.80%	20%
Polys	69.97%	50.20%	
Fold Change	0.43		

FIGURE 1. Translatome studies violate method assumptions. (*Ai*) Polysome profile gradient traces showing differences between siCTRL (red) and siPSF (blue). (*Aii*) RNA concentration of pooled subpolysomal and polysomal RNA showing content imbalance. Quantified distributions are indicated in the corresponding table. (*B*) Study design. (*C*) Measure of skewness of unprocessed simulations at different PDE and SYM, of standard distributions and several translatome studies deposited in public repositories.

in unnormalized data without any improvement between expected and observed values for these probes (Fig. 2B). Conversely, INCATome allowed for a significant reduction in RMSD, thus best reproducing the true biological state for these mRNAs. Performance was also computed with receiver operating characteristics (ROC) and statistical



Δ RMSD	0.02	0.13	0.11	0.23	**-0.52**	**-0.81**	**-0.47**	**-0.98**
P-Value	0.9979	**5e-04**	**0.0038**	**0**	**0**	**0**	**0**	**0**

FIGURE 2. Novel INCATome normalization outperforms existing methods. (*A*) Boxplot for different normalization methods on simulated data at PDE 10 and 75% and Symmetry 0.1 and 0.9. (*B*) RMSD measures per normalization method for a set of given internal probes (spike-in or internal references ACTB/PABP). (*C*) Area under the curve for the ROC curves for each normalization method at each PDE and SYM of the simulation. (*D*) Statistical power (F1 Score) for each normalization method at each PDE and SYM of the simulation.

descriptors (precision, accuracy, sensitivity, specificity). Poor performance is observed for Quantile and LOESS normalizations across the range of simulations as exemplified by the area under the ROC curves (AUC) (Fig. 2C) and the statistical descriptors (Supplemental Fig. 7). INCATome implementations delivered the highest performance, thus confirming the significant improvement over the existing methods.

To summarize the global performance for each method, the statistical power (F1 score) was evaluated over the range of simulations (Fig. 2D). The data clearly illustrate the limit of global performance at PDE = 10% for both Quantile and LOESS methods. INCATome normalizations deliver good statistical power even at high PDE and, importantly, independently of the symmetry of deregulation. Overall, INCATome implementations provide an improved approach for translatome studies, achieving better performance than currently available methods and, critically, without any constraints on PDE and SYM of deregulation.

INCATome improved methodology for identification of DEGs reduces false discovery rate

On the basis of these data, we questioned the applicability of conventional statistical solutions for identification of DEGs since some rely on the same stringent assumptions. Simulated data, normalized by INCATome(IR), was subjected to four different classical statistical tests to assess deregulation: Welch's *T*-Test ("*T* Test"), parametric linear models for microarray ("LIMMA"), nonparametric rank-based approach ("RankProd"), and nonparametric variance-based significance analysis of microarrays ("SAM") (Supplemental Figs. 8, 9). Additionally, since these tests rely on nonredundant methods aiming to identify the same DEGs, we propose a new pipeline for identification of DEGs in translatomes, consisting of selecting significant candidates from the overlap of three out of four statistical tests ("INCATomeDEG").

In cases of asymmetric deregulation (down-regulation SYM = 0.1 and up-regulation SYM = 0.9), LIMMA conserves the fold-change distribution under low and high percentage of deregulation, whereas RankProd and SAM surprisingly induce a substantial deviation under high percentage of deregulation (Fig. 3A). Importantly, the INCATomeDEG approach restores the expected fold-change distribution. We next compared the expected PDE and SYM of deregulation during simulation to the ones obtained after statistical tests. Once again, all the tests except RankProd exhibit nonsignificant differences in RMSD for both PDE and SYM compared to the simulation. The RankProd approach significantly increases the RMSD for the PDE, thus confirming that this method does not respect the true PDE and forces erroneous identification of DEGs (Fig. 3B). Conversely, this deviation was not found in the INCATomeDEG implementation. These findings are reflected in performance differences delivered by RankProd and the rest of the tests, including INCATomeDEG, across the range of simulations as seen with the area under the ROC curve and the statistical descriptors (Fig. 3C; Supplemental Fig. 10).

More globally, the statistical power across the range of simulations shows good performance for most methods, independently of the symmetry (Fig. 3D). As seen previously, RankProd however fails to sustain this performance, especially at high percentage of deregulation or in asymmetrical conditions. Conversely, the new statistical solution proposed with INCATome delivers good statistical power, with an observable limit at the highest PDE 75% where implementation of the LIMMA approach is more preferable, as hinted previously (Jeanmougin et al. 2010). Overall, this new method should give confidence to the user that gathering significant genes from three different statistical tests by INCATomeDEG will give rise to a robust list of candidates DEGs.

INCATome implementation on a biological data set confirms validation improvement over existing methods

To provide biological evidence for INCATome identifying real deregulation in the translatome, we studied the response of our novel method when applied to a biological data set: a conditional silencing of splicing factor PSF in HeLa cells (Fig. 1A—$n = 3$ siCTRL and $n = 3$ siPSF) (Supplemental Fig. 11). Briefly, each sample was background corrected, normalized by either INCATome (IR ACTB and PABP), LOESS or Quantile, and dye-swapped. Each processed data set was then subjected to the INCATomeDEG statistical approach. Distributions of fold change from post-normalization data show distinct phenomenon with both Quantile and LOESS leading to an inference of 74 and 70% down-regulation, respectively, whereas INCATome suggests a 91% down-regulation (Fig. 4A). Furthermore, the global PDE inferred by the detection of DEGs after normalization show a small difference between Quantile or LOESS (4% and 3%) compared to INCATome (1%). Moreover, the RMSD estimations for ACTB and PABP highlight the significant improvement of INCATome over the other normalization methods in a biological context (Fig. 4B). Importantly, across the spectrum of simulations, both Quantile and LOESS normalizations once again try to shrink the inherent skewness and force additional tail heaviness, whereas INCATome has little effect on skewness and reduces tail heaviness (Supplemental Fig. 12).

Following statistical testing with the INCATomeDEG implementation, Quantile and LOESS once again predict a PDE of, respectively, 4% and 3% with 61% and 58% down-regulation (Fig. 4C). Interestingly, the INCATomeDEG method retained the same PDE and SYM from post-normalization, thus inducing no skew in the identified deregulation (1% with 84% down-regulation). Each method identified candidates among which, on average, 7.5% of all DEGs were found in all methods (Fig. 4C). As expected from previous data, Quantile and LOESS share the closest similarity with, on average, 28.8% of DEGs in common compared to only 1.8% shared with INCATome. Lastly, all methods were able to identify specific candidates, with INCATome being the most stringent and Quantile the most permissive (on average INCATome 3.5% < LOESS 14.1% < Quantile 44.3%).

To further ascertain whether these differences in identification of DEGs translate into real biological changes, a panel of candidates was selected in the top significant gene lists of each method independently and these underwent validation by quantitative PCR (qPCR) (Fig. 4D; Supplemental Fig. 13). The selected candidates represented each time the top

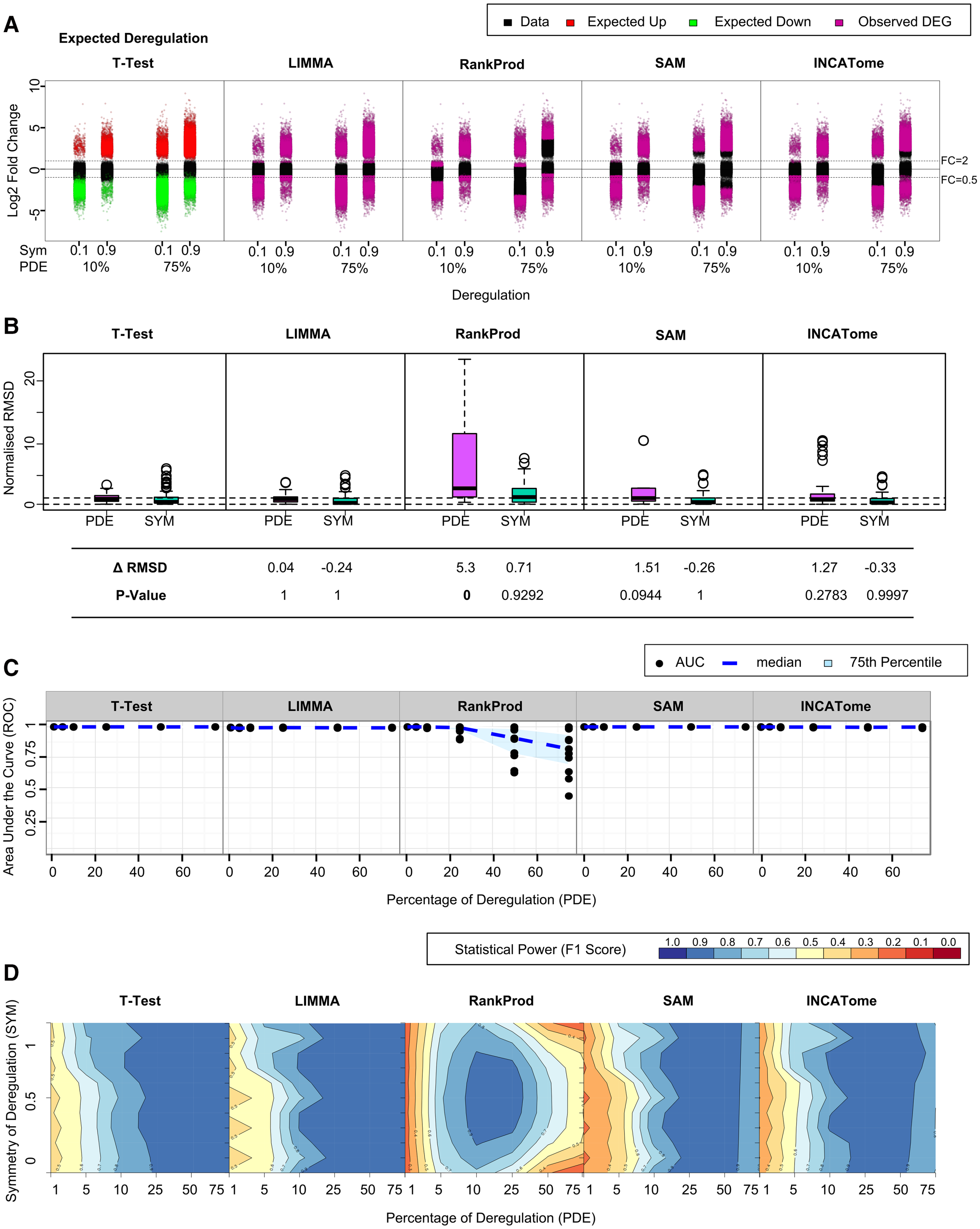


FIGURE 3. Identification of DEGs is hindered in cases of extreme deregulation. (*A*) Boxplot for different statistical methods on simulated data normalized with the internal reference-based INCATome (ACTB/PABP) approach at PDE 10 and 75% and SYM 0.1 and 0.9. (*B*) RMSD measures per statistical test for discovered PDE and SYM. (*C*) Area under the curve for the ROC curves for each statistical method at each PDE and SYM of the simulation. (*D*) Statistical power (F1 Score) for each statistical method at each PDE and SYM of the simulation.

changing candidates as well as mRNAs whose fold change ranged down to the limit cut off (FC > 2 and FC < 0.5). First, a set of 20 targets identified in the Quantile method were selected. Twelve out of 20 validated with a similar fold change trend with qPCR and microarray data (60% validation concordance). In a second step, 21 targets were identified from the LOESS method. Only ten out of 21 mRNAs were confirmed as being deregulated (48% validation

]

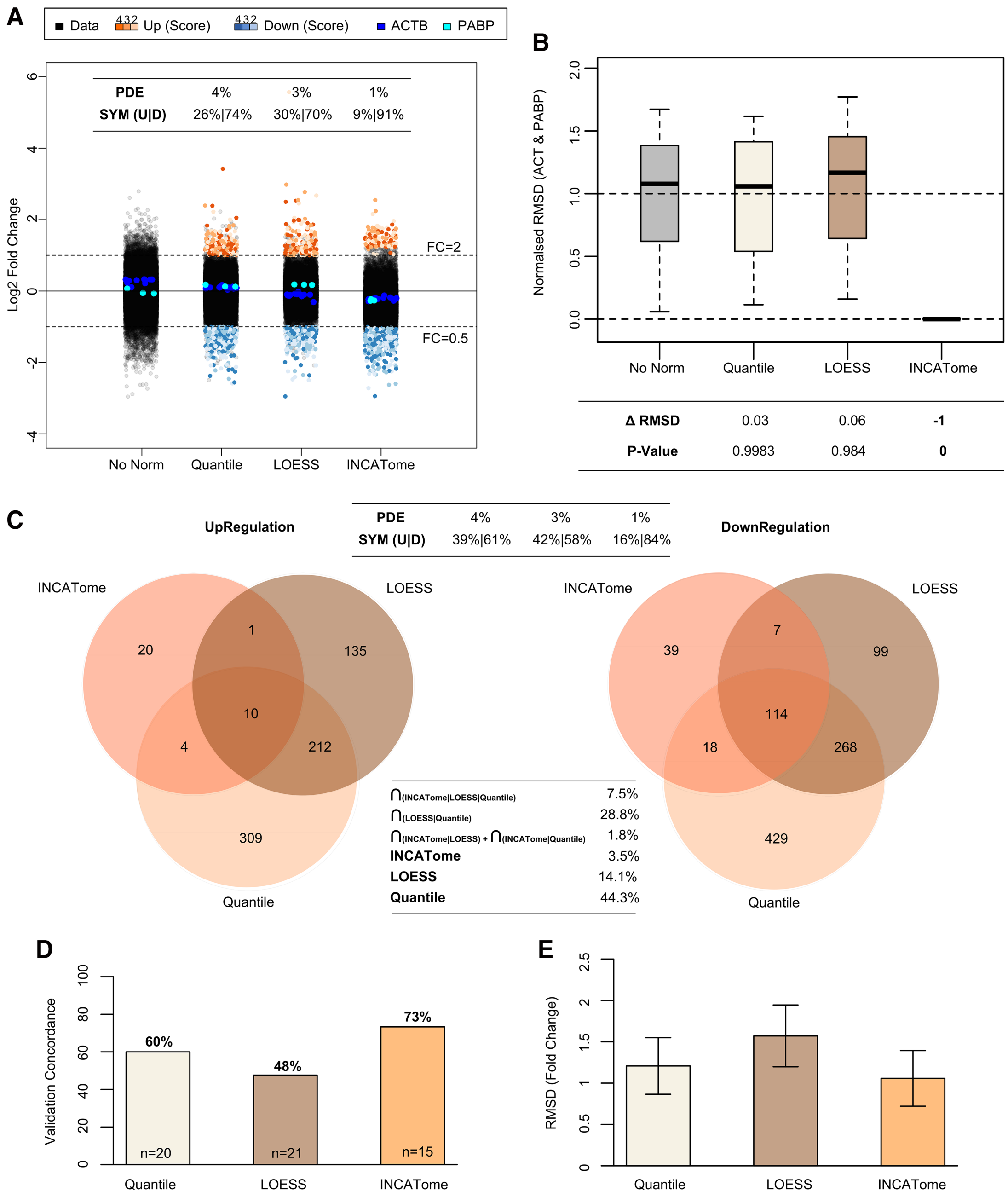


FIGURE 4. INCATome implementation on PSF silenced data set provides the best validation concordance. (*A*) Boxplot for different normalization methods on PSF biological data set. Estimated PDE and SYM are indicated in the corresponding table. (*B*) RMSD measures per normalization method for internal references ACTB/PABP. (*C*) Venn Diagrams representing up-regulation and down-regulation identified in each method. Estimated PDE and SYM and averaged percentage of overlap are indicated in the corresponding tables. (*D*) Barplot of validation concordance between microarray and qPCR validation data. (*E*) Barplot of RMSD applied to fold change for both microarray data and qPCR validation data.

concordance). Finally, we selected 15 top ranked mRNAs identified only by INCATome. Among these, 11 were correctly validated as deregulated translationally (73% validation concordance). Thus, the use of INCATome in the study of the translatome of a biological data set has reinforced the evidence that INCATome outperforms the existing methods by correctly processing the data, identifying DEGs in a stringent manner and, more importantly, providing the best validation concordance on its own, even compared to candidates identified by all three approaches.

New biological inferences discovered with INCATome

Since Quantile and LOESS normalizations induced a skewed identification of DEGs, we sought to provide evidence that the INCATome implementation was also able to discover new biological pathways enriched upon translational deregulation in siPSF cells. To this end, we selected the first top 100 significant DEGs from each method for both up-regulation and down-regulation. Gene functional analysis allowed the identification of significantly enriched functional categories as determined by the modified Fisher exact *P*-value (EASE Score) for gene enrichment analysis. Following gene ontology (GO) reduction, we cross-referenced the identified nonredundant GO categories to associated publications. The relevance of the results from the different methods was assessed by their relative contribution to established concepts (GOPubmed hits of more than two publications), emerging concepts (GOPubmed hits of one publication), and novel concepts (no GOPubmed hits).

As expected from the polysome profile (Fig. 1Ai) and as confirmed by the identification of DEGs by INCATome (Fig. 4), translatome down-regulation seems to prevail in siPSF cells. Thus, functional analysis of the top 100 up-regulated genes yielded few significant nonredundant GO categories (Fig. 5A). Among the established concepts, the category "Phosphorus metabolic process," cross-referenced to the highest number of publications, was identified specifically from the INCATome methodology. PSF has been shown to regulate gene expression of mitochondrial phosphate carrier (PiC) as well as interact with oxidatively modified GAPDH (Iacobazzi et al. 2005; Hwang et al. 2009). Furthermore, two out of three emerging concepts were revealed by INCATome and all hinted toward a role in adhesion. For instance, through its interaction with PSF, E3 ligase Hakai is involved in the adhesion of epithelial and fibroblast cells (Figueroa et al. 2009; Rodriguez-Rigueiro et al. 2011). Similarly, INCATome contributed with other methods in the discovery of novel concepts ("Epithelium development" and "Regulation of MAPK cascade").

As opposed to up-regulation, all methods generally identified more nonredundant GO categories for down-regulation (Fig. 5B). INCATome stringently deduced two functional categories (most significantly enriched and linked to publications) along with other methods ("Neuron projection development" and "Regulation of apoptotic process"). In zebrafish, PSF was reported as essential for neuronal development whereas it also participates in the pathogenesis of spinal muscular atrophy (Lowery et al. 2007; Cho et al. 2014). The role of PSF in apoptosis is well established, as a main regulator of cell death and as playing a critical role in subcellular relocalization of the protein during apoptosis (Shav-Tal et al. 2001; King et al. 2013; Tsukahara et al. 2013). Additionally, the most significantly enriched functional category in the emerging concepts was uniquely identified by INCATome ("Viral life cycle"). Several lines of evidence have been reported linking PSF and viral replication, gene expression, and reproductivity of HIV, influenza, and hepatitis delta viruses (Zolotukhin et al. 2003; Greco-Stewart et al. 2006; Landeras-Bueno et al. 2011; Kula et al. 2013). Finally, albeit INCATome yielded only four novel concepts, these have the specificity of not representing either responses to stimulus or developmental processes unlike all other novel concepts identified. Thus, INCATome once again, via its stringency and novelty, allows focusing of biological interest onto relevant pathways.

Overall, we have shown that INCATome is highly efficient in identifying DEGs enriched in established concepts, either uniquely or in concordance with other methods. Secondly, INCATome has proved useful to confirm emerging concepts by validating nonredundant GO categories as highly significant or multirepresented by different processes. Last, but not least, INCATome is decisive in discovering significantly enriched novel concepts in a stringent manner.

DISCUSSION

Analysis of microarray data by existing methods can be severely compromised by significant skew inherent to biological data sets following treatment, and in systems where skew develops rapidly, this can be very difficult to mitigate. The translational state of a cell can be measured by "polysome profiling" followed by genome-wide technologies such as microarray and deep-sequencing (for a discussion, see King and Gerber 2014). However, levels of translation can show very large directional change with treatment, as cells respond rapidly to changes in global protein synthesis demand: The resulting skews make data difficult to analyze but present an ideal system in which to (i) assess the relative effectiveness of dealing with skew of existing analytical methods and (ii) develop and validate a novel method which shows improved ability to accurately analyze change against a background of substantial skew.

While RNA deep-sequencing (RiboSeq) is increasingly being used in place of microarray analysis, both present considerable analytical challenges and the latter remains a popular choice where time and/or financial resources are limited and the level of detail available via RiboSeq may not be essential. An ongoing challenge for microarray analysis of translatomes is that no tailored methodologies have been available. Instead, analysis workflows designed for transcriptome (global mRNA) studies have been adapted, but these cannot cope well with the significant and rapid deregulation that is so often associated with translational response.

Conventionally implemented normalization methods include Quantile or LOESS, followed by statistical testing (modified *T*-test, LIMMA, RankProd, SAM, etc.). The former consist of nonparametric approaches based on standardizing distributions or local regression, respectively, while the latter offer a variety of different approaches (parametric or nonparametric). Most importantly, all these procedures assume that deregulation must occur in low proportion (PDE below 10%) and in equal symmetry (SYM of 0.5). Here, we have shown that translatome studies in general violate these assumptions. Our comprehensive simulations of microarray translatome studies (which vary PDE and SYM to levels of

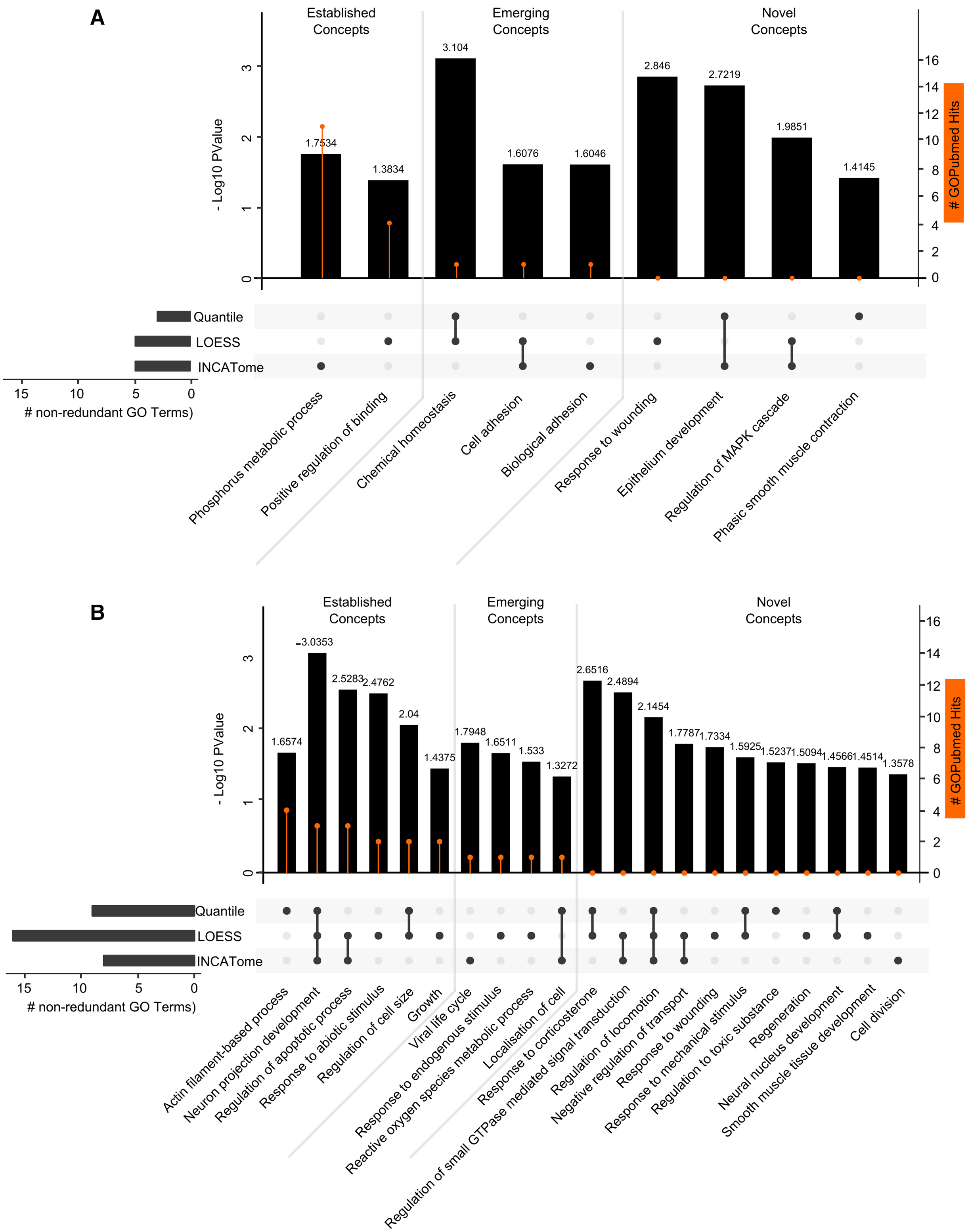


FIGURE 5. Biological inference enhanced by the implementation of INCATome. Barplots representing the logged *P*-value for the nonredundant gene ontology (GO) categories associated with the top 100 genes identified as up-regulated (*A*) or down-regulated (*B*) with each normalization method (in at least three statistical tests). Overlap in identification is represented by joined dots in the central table. Orange bars correspond to manually curated GOPubmed hits linking the *SFPQ* gene to the given GO category. Side barplot summarizes the number of nonredundant GO terms associated with each method.

intermediate skewness relative to published translatome data sets) confirm that general assumptions of standard methods can be violated.

Our novel translatome analysis methodology, "INCATome," consists of (i) a normalization approach based on a RMSD algorithm tied to internal controls (SpikeIn or internal references measured experimentally in parallel) and (ii) a statistical pipeline, combining four different statistical tests to reduce the false-positive hits. INCATome outperforms both Quantile and LOESS at the normalization step by providing efficient and powerful normalization without inducing unwanted skew. INCATome also delivers better performance during statistical testing compared to the individual tests alone and so allows users to more accurately identify DEGs, even in extreme deregulation conditions.

Application of INCATome and other methods to a real biological data set (control siRNA vs. knockdown of a protein, PSF, with known, critical roles in splicing and DNA repair; for review, see Yarosh et al. 2015) confirms that INCATome outperforms with respect to normalization and stringent DEG identification, with notable improvements in rates of validation, as determined by qPCR of high- and low-translated candidate mRNAs. The DEGs from the PSF knockdowns identified by INCATome categorize into several significantly enriched functional pathways: In some cases these are linked to established concepts, further supporting the suitability of the INCATome method, but in others they identify interesting novel pathways that may benefit from further research.

Overall, the novel methodology INCATome improves the statistical approach of difficult systems characterized by systematic bias and skewness. One example of such a system is the study design, which can in some instances introduce by nature a systematic bias or skewness. We have shown that INCATome is highly performing in translatome studies and predict that it could be equally powerful when studying miRNA-mediated control. Thus, the improvements in data quality wrought by INCATome compared to preexisting methods are sufficiently beneficial that reanalysis of existing data sets in which skew is significant may be a quick and valuable means for researchers to generate novel avenues for research, as well as applying it to future studies.

MATERIALS AND METHODS

Cell culture

HeLa cells were plated on 15-cm plates and grown in DMEM with 15% FCS. Dharmacon on-target predesign siRNAs (four individuals) for PSF or a control siRNA (C3; UGGUUUACAUGUUU UCUGA, Dharmacon) were transfected using RNAiMax (Invitrogen). Each individual siRNA was used as 0.4 nM final concentration. Fresh media was changed after 6 h, cells were split 24 h after the start of transfection and harvested after 48 h.

Sucrose density-gradient centrifugation and RNA detection

Sucrose density-gradient (10%–60%) centrifugation was used to separate ribosomes into polysomal and subpolysomal fractions. Gradients were then fractionated with continuous monitoring at 254 nm and RNA was isolated from each fraction as described previously (Sbarrato et al. 2016).

RNA analysis

Northern analysis of RNA isolated from sucrose density gradients was performed as described previously (Sbarrato et al. 2016). Radiolabeled DNA hybridization probes were generated using the RadPrime Kit according to the manufacturer's instructions (Invitrogen).

Preparation of fluorescently labeled cDNA for microarray hybridization and data analysis

Microarrays used were Agilent 8×60k Human Gene Expression arrays (Agilent Technologies LDA UK Ltd.). Equal proportions of RNA from pooled subpolysomal fractions (fractions 1–5) and pooled polysomal fractions (fractions 6–11) were fluorescently labeled, using the Agilent Low Input Quick Amp Labeling Kit, two-color (Agilent Technologies LDA UK Ltd.). Following hybridization, the arrays were scanned using an Agilent SureScan High Resolution scanner, and Agilent Feature Extraction software was applied to the resulting images.

Use of spike-in mix for microarrays

RNA spike-in mix, containing 10 in vitro synthesized, polyadenylated transcripts prepared in predetermined ratios, targeted Agilent microarray control probes. Spike-ins provided in the labeling kit were diluted (according to manufacturer's protocol), then added to pooled RNA fractions prior to the labeling reaction (Agilent Technologies).

Development of a new normalization procedure

A novel methodology for the normalization of polysome profiling data was developed (Supplemental Fig. 1). It relies on the use of either Spike-In controls or internal references. Each species presents the advantage of having expected values (known concentration for Spike-In and measured polysomal distribution for the internal references such as ACTB and PABP). Thus, expected $\log_2$ ratios are computed and a root-mean-square deviation (RMSD) algorithm is applied to each sample individually in order to converge toward the smallest RMSD residual. This assured (i) an internal normalization for each sample, (ii) the possibility to compare samples across microarrays, and (iii) the conservation of the inherently skewed nature of each sample. The R code for implementing the INCATome normalization is available as a package of the open-resource CRAN project.

Simulation of microarray data

Polysome profiling of siCTRL HeLa cells was performed and the resulting subpolysomal and polysomal fractions were hybridized together onto two microarrays (dye-swap design). Preprocessing steps included background correction (LIMMA package) and dye-swap correction (averaging). Values were then sampled from the resulting data set and simulated based on a previously reported method (Dembélé 2013) with several modifications (Supplemental Figs. 2, 3). Briefly, simulation was performed in order to create three replicates for two different conditions. Gene variation (λ_1 and λ_2), technical variability (sd_{TV}), and deregulation parameters (μ_{DE}, sd_{DE} and λ_2) were constant across the simulation. Conversely, the percentage of deregulation (PDE), the symmetry of deregulation (SYM), and the data noise (sd_N) ranged between (1%, 5%, 10%, 25%, 50%, 75%), (0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0), and (0.1, 0.4), respectively. Each simulation was then subjected to several normalization procedures: Quantile, LOESS, INCATome (Spike-In), and INCATome (ACTB/PABP) (Bolstad et al. 2003; Smyth and Speed 2003). Finally, each simulation was subjected to INCATome (ACTB/PABP) normalization and carried over for comparisons across different statistical tests: Welch's *T*-Test, model fitting LIMMA, RankProd, and SAM (Tusher et al. 2001; Hong et al. 2006; Ritchie et al. 2015). The R code for simulating translatome data is available on request by emailing the corresponding author.

Assessment of distribution characteristics and performance

Normalization procedures were assessed by determining the specificity, sensitivity, accuracy, precision, and F1 score following a Welch's *T*-Test between the two conditions. Skewness, excess kurtosis, and tail heaviness (Hogg) were also determined and tested by the Kolmogorov–Smirnov test. Receiver operating characteristic (ROC) curves were generated and AUCs calculated for each condition.

Quantitative PCR validation

qPCR validation was carried out using Qiagen Quantitect primer assays and Quantitect SYBR Green PCR mix, according to the manufacturer's instructions. (Qiagen GmbH).

Gene functional analysis

Lists of the top 100 deregulated genes were subjected to gene functional analysis using the DAVID v6.7 webtool (Huang da et al. 2009). All GO categories significantly enriched (as determined by the modified Fisher exact *P*-value "EASE Score" for gene enrichment analysis) were then reduced using the REVIGO webtool (Supek et al. 2011) to yield nonredundant significant GO categories. Functional analysis plots were produced in R based on modifications of the UpsetR package (Lex et al. 2014). Publication records were manually curated on GOPubmed based on a search of *SFPQ* gene followed by filtering by GO categories (Doms and Schroeder 2005).

Image analysis

Image quantification was performed using ImageJ software (Schneider et al. 2012).

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGMENTS

We thank Nick Burgoyne for his contribution to the initial method development. This work was funded by grants from the Medical Research Council programme (T.S., L.A.W., R.S., A.B., T.P., and A.E.W.), the Wellcome Trust (X.P.), and the Core Medical Research Council Toxicology Unit (C.J. and K.D.)

Received December 22, 2016; accepted July 27, 2017.

REFERENCES

Bala R, Agrawal RK, Sardana M. 2010. *Relevant gene selection using normalized cut clustering with maximal compression similarity measure*, Lecture Notes in Computer Science book series (LNCS, Vol. 6119). Springer-Verlag, Berlin.

Blagden SP, Willis AE. 2011. The biological and therapeutic relevance of mRNA translation in cancer. *Nat Rev Clin Oncol* **8:** 280–291.

Bolstad BM, Irizarry RA, Astrand M, Speed TP. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19:** 185–193.

Cho S, Moon H, Loh TJ, Oh HK, Williams DR, Liao DJ, Zhou J, Green MR, Zheng X, Shen H. 2014. PSF contacts exon 7 of SMN2 pre-mRNA to promote exon 7 inclusion. *Biochim Biophys Acta* **1839:** 517–525.

Dabney AR, Storey JD. 2007. A new approach to intensity-dependent normalization of two-channel microarrays. *Biostatistics* **8:** 128–139.

Dembélé D. 2013. A flexible microarray data simulation model. *Microarrays (Basel)* **2:** 115–130.

Doms A, Schroeder M. 2005. GoPubMed: exploring PubMed with the Gene Ontology. *Nucleic Acids Res* **33:** W783–W786.

Dudoit S, Yang YH, Callow MJ, Speed T. 2002. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Stat Sin* **12:** 111–139.

Edgar R, Domrachev M, Lash AE. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* **30:** 207–210.

Figueroa A, Kotani H, Toda Y, Mazan-Mamczarz K, Mueller EC, Otto A, Disch L, Norman M, Ramdasi RM, Keshtgar M, et al. 2009. Novel roles of Hakai in cell proliferation and oncogenesis. *Mol Biol Cell* **20:** 3533–3542.

Galfalvy HC, Erraji-Benchekroun L, Smyrniotopoulos P, Pavlidis P, Ellis SP, Mann JJ, Sibille E, Arango V. 2003. Sex genes for genomic analysis in human brain: internal controls for comparison of probe level data extraction. *BMC Bioinformatics* **4:** 37.

Greco-Stewart VS, Thibault CS, Pelchat M. 2006. Binding of the polypyrimidine tract-binding protein-associated splicing factor (PSF) to the hepatitis *delta* virus RNA. *Virology* **356:** 35–44.

Harr B, Schlotterer C. 2006. Comparison of algorithms for the analysis of Affymetrix microarray data as evaluated by co-expression of genes in known operons. *Nucleic Acids Res* **34:** e8.

Hong F, Breitling R, McEntee CW, Wittner BS, Nemhauser JL, Chory J. 2006. RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics* **22:** 2825–2827.

Huang da W, Sherman BT, Lempicki RA. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4:** 44–57.

Huber W, von Heydebreck A, Sultmann H, Poustka A, Vingron M. 2002. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18:** S96–S104.

Hwang NR, Yim SH, Kim YM, Jeong J, Song EJ, Lee Y, Lee JH, Choi S, Lee KJ. 2009. Oxidative modifications of glyceraldehyde-3-phosphate dehydrogenase play a key role in its multiple cellular functions. *Biochem J* **423:** 253–264.

Iacobazzi V, Infantino V, Costanzo P, Izzo P, Palmieri F. 2005. Functional analysis of the promoter of the mitochondrial phosphate carrier human gene: identification of activator and repressor elements and their transcription factors. *Biochem J* **391:** 613–621.

Ingolia NT. 2010. Genome-wide translational profiling by ribosome footprinting. *Methods Enzymol* **470:** 119–142.

Ingolia NT. 2016. Ribosome footprint profiling of translation throughout the genome. *Cell* **165:** 22–33.

Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. 2003a. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res* **31:** e15.

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. 2003b. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4:** 249–264.

Jeanmougin M, de Reynies A, Marisa L, Paccard C, Nuel G, Guedj M. 2010. Should we abandon the *t*-test in the analysis of gene expression microarray data: a comparison of variance modeling strategies. *PLoS One* **5:** e12336.

Kerr MK, Martin M, Churchill GA. 2000. Analysis of variance for gene expression microarray data. *J Comput Biol* **7:** 819–837.

King HA, Gerber AP. 2014. Translatome profiling: methods for genome-scale analysis of mRNA translation. *Brief Funct Genomics* **15:** 22–31.

King HA, Cobbold LC, Pichon X, Poyry T, Wilson LA, Booden H, Jukes-Jones R, Cain K, Lilley KS, Bushell M, et al. 2013. Remodelling of a polypyrimidine tract-binding protein complex during apoptosis activates cellular IRESs. *Cell Death Differ* **21:** 161–171.

Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, Williams E, Dylag M, Kurbatova N, Brandizi M, Burdett T, et al. 2015. ArrayExpress update—simplifying data submissions. *Nucleic Acids Res* **43:** D1113–D1116.

Kula A, Gharu L, Marcello A. 2013. HIV-1 pre-mRNA commitment to Rev mediated export through PSF and Matrin 3. *Virology* **435:** 329–340.

Landeras-Bueno S, Jorba N, Perez-Cidoncha M, Ortin J. 2011. The splicing factor proline-glutamine rich (SFPQ/PSF) is involved in influenza virus transcription. *PLoS Pathog* **7:** e1002397.

Landfors M, Philip P, Ryden P, Stenberg P. 2011. Normalization of high dimensional genomics data where the distribution of the altered variables is skewed. *PLoS One* **6:** e27942.

Larsson O, Sonenberg N, Nadon R. 2010. Identification of differential translation in genome wide studies. *Proc Natl Acad Sci* **107:** 21487–21492.

Le Quesne JP, Spriggs KA, Bushell M, Willis AE. 2010. Dysregulation of protein synthesis and disease. *J Pathol* **220:** 140–151.

Lex A, Gehlenborg N, Strobelt H, Vuillemot R, Pfister H. 2014. UpSet: visualization of intersecting sets. *IEEE Trans Vis Comput Graph* **20:** 1983–1992.

Li C, Wong WH. 2001. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc Natl Acad Sci* **98:** 31–36.

Lin DY, Zou F. 2004. Assessing genomewide statistical significance in linkage studies. *Genet Epidemiol* **27:** 202–214.

Lowery LA, Rubin J, Sive H. 2007. *whitesnake/sfpq* is required for cell survival and neuronal development in the zebrafish. *Dev Dyn* **236:** 1347–1357.

Pelz CR, Kulesz-Martin M, Bagby G, Sears RC. 2008. Global rank-invariant set normalization (GRSN) to reduce systematic distortions in microarray data. *BMC Bioinformatics* **9:** 520.

Polunovsky VA, Bitterman PB. 2006. The cap-dependent translation apparatus integrates and amplifies cancer pathways. *RNA Biol* **3:** 10–17.

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. 2015. *limma* powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43:** e47.

Rodriguez-Rigueiro T, Valladares-Ayerbes M, Haz-Conde M, Aparicio LA, Figueroa A. 2011. Hakai reduces cell-substratum adhesion and increases epithelial cell invasion. *BMC Cancer* **11:** 474.

Sbarrato T, Horvilleur E, Poyry T, Hill K, Chaplin LC, Spriggs RV, Stoneley M, Wilson L, Jayne S, Vulliamy T, et al. 2016. A ribosome-related signature in peripheral blood CLL B cells is linked to reduced survival following treatment. *Cell Death Dis* **7:** e2249.

Schneider CA, Rasband WS, Eliceiri KW. 2012. NIH Image to ImageJ: 25 years of image analysis. *Nat Methods* **9:** 671–675.

Shav-Tal Y, Cohen M, Lapter S, Dye B, Patton JG, Vandekerckhove J, Zipori D. 2001. Nuclear relocalization of the pre-mRNA splicing factor PSF during apoptosis involves hyperphosphorylation, masking of antigenic epitopes, and changes in protein interactions. *Mol Biol Cell* **12:** 2328–2340.

Smyth GK. 2004. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* **3:** Article3.

Smyth GK, Speed T. 2003. Normalization of cDNA microarray data. *Methods* **31:** 265–273.

Smyth GK, Yang YH, Speed T. 2002. Statistical issues in cDNA microarray data analysis. *Methods Mol Biol* **224:** 111–136.

Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, Deng S, Johnsen H, Pesich R, Geisler S, et al. 2003. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci* **100:** 8418–8423.

Spriggs KA, Bushell M, Willis AE. 2010. Translational regulation of gene expression during conditions of cell stress. *Mol Cell* **40:** 228–237.

Supek F, Bosnjak M, Skunca N, Smuc T. 2011. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* **6:** e21800.

Thomas JG, Olson JM, Tapscott SJ, Zhao LP. 2001. An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genome Res* **11:** 1227–1236.

Tsukahara T, Haniu H, Matsuda Y. 2013. PTB-associated splicing factor (PSF) is a PPARγ-binding protein and growth regulator of colon cancer cells. *PLoS One* **8:** e58749.

Tusher VG, Tibshirani R, Chu G. 2001. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci* **98:** 5116–5121.

van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, et al. 2002. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415:** 530–536.

Yang YH, Speed T. 2002. Design issues for cDNA microarray experiments. *Nat Rev Genet* **3:** 579–588.

Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP. 2002. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res* **30:** e15.

Yarosh CA, Iacona JR, Lutz CS, Lynch KW. 2015. PSF: nuclear busybody or nuclear facilitator? *Wiley Interdiscip Rev RNA* **6:** 351–367.

Zhao Y, Pan W. 2003. Modified nonparametric approaches to detecting differentially expressed genes in replicated microarray experiments. *Bioinformatics* **19:** 1046–1054.

Zolotukhin AS, Michalowski D, Bear J, Smulevitch SV, Traish AM, Peng R, Patton J, Shatsky IN, Felber BK. 2003. PSF acts through the human immunodeficiency virus type 1 mRNA instability elements to regulate virus expression. *Mol Cell Biol* **23:** 6618–6630.