

RESEARCH

Open Access

Reproducibility and reliability assays of the gene expression-measurements

Behrooz Darbani^{1,2*} and Charles Neal Stewart Jr³

Abstract

Background: Reliability and reproducibility are key metrics for gene expression assays. This report assesses the utility of the correlation coefficient in the analysis of reproducibility and reliability of gene expression data.

Results: The correlation coefficient alone is not sufficient to assess equality among sample replicates but when coupled with slope and scatter plots expression data equality can be better assessed. Narrow-intervals of scatter plots should be shown as a tool to inspect the actual level of noise within the data. Here we propose a method to examine expression data reproducibility, which is based on the ratios of both the means and the standard deviations for the inter-treatment expression ratios of genes. In addition, we introduce a fold-change threshold with an inter-replicate occurrence likelihood lower than 5% to perform analysis even when reproducibility is not acceptable. There is no possibility to find a perfect correlation between transcript and protein levels even when there is not any post-transcriptional regulatory mechanism. We therefore propose an adjustment for protein abundance with that of transcript abundance based on open reading frame length.

Conclusions: Here, we introduce a very efficient reproducibility approach. Our method detects very small changes in large datasets which was not possible through regular correlation analysis. We also introduce a correction on protein quantities which allows us to examine the post-transcriptional regulatory effects with a higher accuracy.

Keywords: Gene expression, Reliability, Reproducibility

Background

Reliability or accuracy of data-to-reality, and reproducibility or inter-replicate variance, are typically assessed using the correlation coefficient. As a reliability assay, the agreement between transcript abundance and their encoded protein abundance has long been investigated [1-4]. RNA-Seq [5] has made the global profiling of gene expression achievable by covering comprehensive transcriptome data. As the most recent approach, reliability of the output has been examined by correlation analysis using microarray and/or Real Time RT-PCR [5-10]. In addition, the correlation between replicates has been used to judge the reproducibility of data [7,10-15].

Correlation is the agreement level between two variables. The correlation coefficient varies between perfect

agreement ($r = \pm 1$) and no agreement ($r = 0$). A perfect agreement implies equal change in one of the variables for a given fixed change of another variable. A negative correlation indicates an opposite direction in the agreement, i.e. reduction in one of the variables in response to positive changes in another, regardless of the agreement level. For example, we can find a perfect agreement but in the opposite direction between the dose of insecticides and the number of survived insects.

The inter-replicate variation of gene expression-quantities is of the utmost importance to biologists because lower variance means higher reproducibility. The correlation coefficient has been used to inspect the reproducibility of gene expression data [7,10-15]. When the inter-replicate variance is low or zero, a higher or perfect correlation is expected but we can find the opposite, i.e. a low inter-replicate variance may be absent when there is a strong correlation. Thus, we cannot rely on the correlation digit as the sole criterion.

Since the correlation coefficient has been the main tool to assess agreement between replicates, technologies

* Correspondence: behroozdarbani@gmail.com

¹Department of Plant and Environmental Sciences, University of Copenhagen, Thorvaldsensvej 40, 1871 Frederiksberg, Denmark

²Department of Molecular Biology and Genetics, Research Centre Flakkebjerg, Aarhus University, Forsøgsvej 1, 4200 Slagelse, Denmark
Full list of author information is available at the end of the article

(microarray, Real Time RT-PCR, and RNA-Seq), and between protein and transcript abundance, this paper scrutinizes the power of correlation analysis for the biologist and introduces other alternatives. We introduce an alternative reproducibility assay approach. We advise that there should be an application of an inter-replicate based fold-change threshold to identify the most reliable variations among the significant inter-treatment changes regardless of what the inter-replicate reproducibility may be. In addition, we introduce a correction step based upon the lengths of open reading frames (ORFs) for protein quantity before correlation analysis with the respective transcript abundance in order to assess post-transcriptional regulation.

Results and discussion

The correlation coefficient is not sufficient as a reproducibility assay for expression data

By analyzing a publicly available RNA-Seq data we examined reproducibility of ≈ 25000 genes between two replicates. We found correlations of 0.986 and 0.985 by applying unigene reads and total exon reads, respectively (Figure 1A1 and B1). However, we also found 65% and 64.09% of the genes with a variation higher than 50% between the two replicates (Figure 1A1 and B1). Through further analyses, the insensitivity of the correlation coefficient to the large shifts of inter-replicate variation was confirmed after data correction. Comparisons of Figure 1A1 with A2 and Figure 1B1 with B2 indicate that the large declines in the levels of inter-replicate variances (\approx from 65% to 12% of the genes with an inter-replicate variation higher than 50%) were insufficient to influence the correlations. By comparing Figure 1A2 and B2 with 1C where the slopes are nearly the same, it is clear that the correlation coefficient depends on the noise level of the data. Alternatively, relying on the correlation and taking the slope of inter-replicate regression line into account could increase the efficiency of the reproducibility assay. In contrast to the correlation coefficient, considerable improvements of the slopes from 0.620 to 1.006 and 0.972 are good representatives of the correction efficiencies on unigene reads and total exon reads, respectively (Figure 1A1, A2, B1 and B2). However, the threshold level for the noise should be determined by the nature of the experiments. This is crucial when dealing with large sets of data as in microarray or RNA-Seq assays. In this situation, a slope with a small deviation from 1 and a significantly high correlation could coincide with a considerable inter-replicate variation. Although it is not discernible by the slope and the level of noise in the example chosen, 12% of the genes have greater than a 50% inter-replicate variation (Figure 1A2 and B2). Therefore, it is an easy pitfall to 'neglect' a few thousand genes

with considerable variance among 30,000 genes, which can result in false positive biological conclusions.

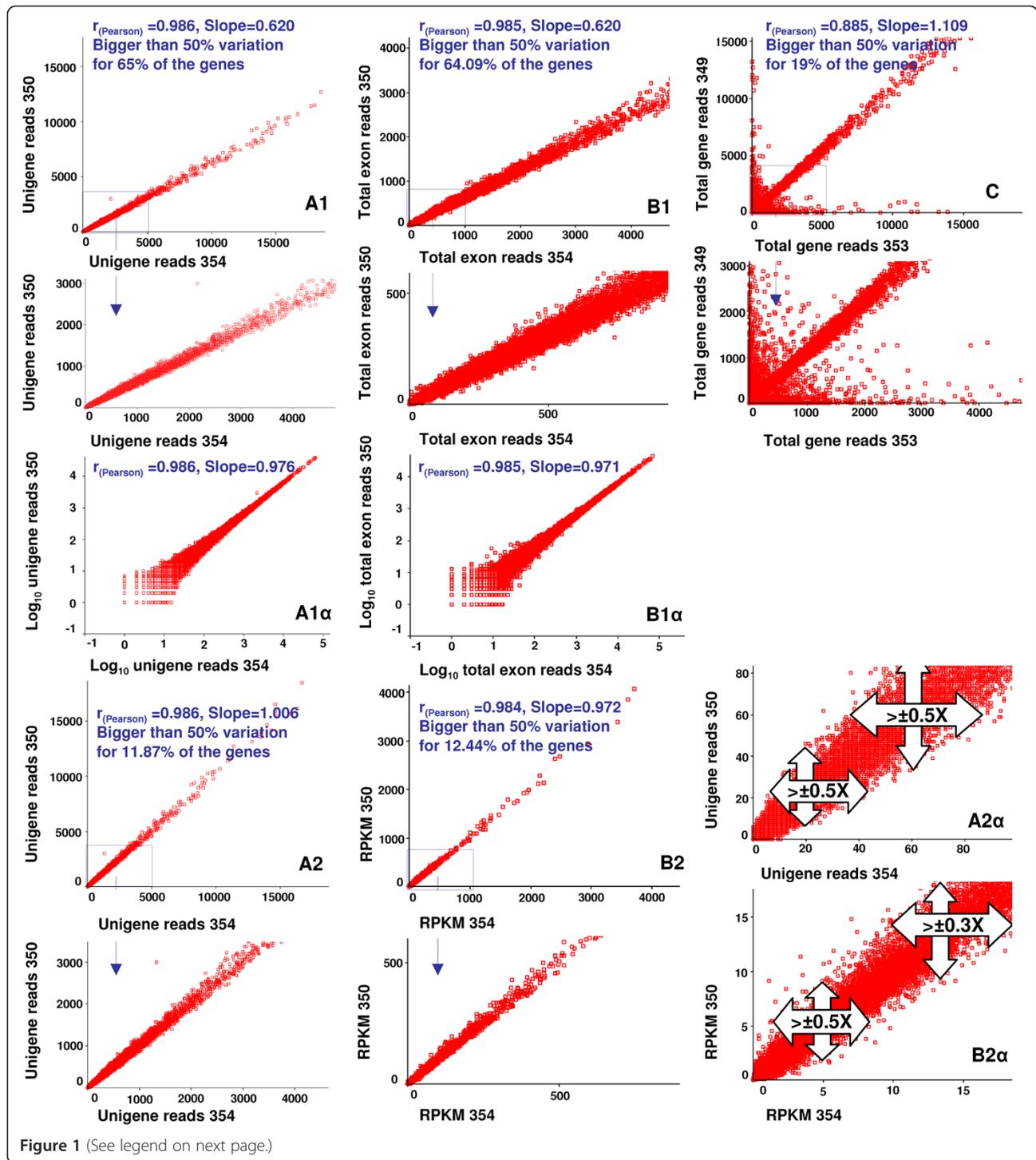
Although the actual level of noise on the plots is not visible owing to the large dataset size, the inspection of very narrow intervals of the data range can be helpful (e.g., compare Figure 1A2 and B2 with A2 α and B2 α , respectively). By following this procedure we found $> \pm 0.5\times$ noise-level in the presented examples, which were not detectable when showing the whole dynamic range of expression. We advise that it would be helpful to include figures and supplements showing narrow-intervals of representations in the reports to provide readers with a useful overall picture.

Scatter plots using the logarithmic scale have been reported previously to represent expression data reproducibility [2,4,5,11,12,15-17]. However, logarithmic transformation of data evokes changes in the variation between replicates, which affects slope and diminishes its usefulness. The similarity level between replicates and the percent of data which are 1 and lower than 1 have an impact on the direction, i.e. whether slope-deviation will be increased or decreased, and the rate of change of slope. We found considerable decreases in the deviation of the slope of inter-replicate regression line using uncorrected data where the inter-replicate similarity levels were low (compare Figure 1A1 and B1 with A1 α and B1 α , respectively). This artificially low slope-deviation could be incorrectly interpreted as high reproducibility. In conclusion, logarithmic scale slope is of little utility to reproducibility assays.

To examine the reliability of gene expression measurements, the agreement levels of expression data between technologies, i.e. RNA-Seq, microarray, and Real Time RT-PCR, have also been investigated [5-10]. These assays might have general agreement with each other but equality is not observed. Therefore, we should require a strong positive correlation but not equality among platforms, which could be represented by the slope of inter-replicate regression line. Aside from the evidence-free acceptance of one technology as the most reliable one, reliability assays suffer from the inefficiency of correlation coefficient when analyzing large sets of data. The acceptable threshold of correlation should be evaluated. Like for the reproducibility assay, we advise that the real level of noise should be examined on narrow-intervals of data ranges in scatter plots.

The mean and standard deviation of the inter-treatment gene expression ratios shed light on reproducibility

As an alternative to the correlation coefficient, we advise reporting the mean and standard deviation of the inter-treatment gene expression ratios to examine reproducibility. By calculating the ratios of each of the descriptive statistics for single replicates when compared to different replicates of another treatment, we should check the deviation of ratios from 1. The average of deviations and their standard deviation could be utilized as reproducibility



(See figure on previous page.)

Figure 1 Correlation-based reproducibility and reliability assays on real gene expression data. A publicly available RNA-Seq dataset from rice including two root-sample replicates and two shoot-sample replicates is used as an example to illustrate the power of correlation coefficient in the reproducibility assay of read counts as gene expression data. Briefly, the reads were mapped on the rice genome to apply the mapped total gene, unigene, and total exon reads (**A1**, **A2** and **B1**, **B2**). Large bias shifts from ~65% in **A1** and **B1** to ~12% in **A2** and **B2** after both reference gene-based correction (**A**) and read count-based correction (**B**) were not able to increase the correlation coefficient. (**C**) As the level of noise increases the correlation becomes weaker. The slopes are nearly the same and approach a slope of 1, but the lower correlation observed results from the higher level of noise in **C** compared to **A2** and **B2**. Considering the level of noise in the data and the slope of inter-replicate regression line, it is possible that, while helpful, they might not be precise for reproducibility assay. After data correction (in **A2** and **B2**), the correlation was not changed but the slope was improved, i.e. the slope-deviation from 1 was decreased. However, 12% of the genes still show 50% or greater than 50% inter-replicate variation in expression, which would be indistinguishable by taking both the correlation and the slope into account. (**A1a**, **B1a**) Logarithmic transformation of data could change the difference between sample replicates which is represented by the changed slope-deviation on scatter plots. (**A2a**, **B2a**) Scatter plot narrow-intervals allow us to observe noise in the dataset. There is almost a clear $\pm 0.5x$ variation shown by scatter plots. 350, 354: rice untreated root sample replicates; 349, 353: rice untreated shoot sample replicates. We evaluated the expression of 24122 to 24701 genes by applying the different types of reads and samples. The log10 transformed data were used to calculate the Pearson correlation. All correlations were significant ($P = 0.000$) by a two-tailed test.

coefficients. The higher deviation and higher standard deviation for the deviations indicate lower reproducibility.

As an example, we performed the analysis on the total exon mapped reads for a rice experiment and its corrected “reads per kilobase of exon model per million mapped reads” (RPKM) data by applying different inter-treatment comparisons of untreated root and salinity treated shoot, untreated shoot and salinity treated shoot, untreated shoot and untreated root, untreated root and salinity treated root, salinity treated root and salinity treated shoot, and salinity treated root and untreated shoot (see Additional file 1). Analysis of two replicates of each treatment resulted in 2×48 ratios for each of the datasets, i.e. corrected and uncorrected reads. The average of the ratios’ deviations were 0.338 and 0.123 and the standard deviation of ratios’ deviations were 0.318 and 0.133 on total exon reads and on RPKM data, respectively. As expected, these coefficients represent higher reproducibility when using the corrected RPKM data. Compared to the correlation-based reproducibility assay, where there was no difference between the correlations of uncorrected and corrected data, it is also more effective to examine reproducibility by applying the new approach. As shown in Figure 1, the correlation coefficient was not effective to represent the significant reduction in the inter-replicate variance (\approx from 65% to 12% of the genes with an inter-replicate variation higher than 50%) after data correction. In contrast, the reduction from 33.8% to 12.3% in the reproducibility level proves the effectiveness of the new method compared to the correlation coefficient. Therefore, the new approach provides a higher level of efficiency to represent different variability levels compared with the correlation-based approach, at least using the applied experimental data. In addition, the decreased standard deviation of ratios’ deviations from 0.318 to 0.133 indicates an increased reproducibility among the treatments after data correction. A similar approach is also applicable between replicates just by applying the deviations of averages and standard

deviations of gene expression ratios. The average and standard deviation of inter-replicate expression ratios are expected to be 1 and 0 when there is perfect reproducibility. However, we need a higher number of replicates to obtain reliable measures. This should not be a concern when the main aim is to evaluate the reproducibility of a method or technology. As shown by the example analyzed here, four samples each with two replicates works well. However, it is worth to evaluate the impact of the number of replicates on the proposed method. Our analysis indicates that the method works at an acceptable level on a dataset with three replicates. By analyzing an online-available yeast RNA-Seq dataset [18] we found a Pearson correlation of 0.994 among technical replicates of samples 1 and 2 while 4.3% and 4.2% of the 6387 genes had 50% or higher inter-replicate variations, respectively. In contrast, our proposed approach found 14.2% and 13.7% deviations for samples 1 and 2, respectively. This indicates that the proposed method can inform small changes, e.g., 4.3% to 4.2% of the genes with 50% or higher inter-replicate variations. In agreement with this, we also found a 15.3% deviation when comparing biological replicates, i.e. sample 1 with sample 2, with 4.4% of the genes showing 50% or higher inter-replicate variations (see Additional file 1). This indicates that our method has high sensitivity and can even detect very small changes in large datasets. Therefore, we advise our method to be utilized as an efficient alternative to the correlation coefficient. To examine the efficiency of our method on various gene expression datasets, we analyzed microarray data from *Arabidopsis*, barley, yeast, and *E. coli* (Table 1; Additional file 2). Similar to the RNA-Seq data, correction of the microarray data resulted in strong reduction in the number of genes showing over 50% inter-replicate variations without considerable improvement in the correlation coefficients between replicates (Table 1). The average of slopes of inter-replicate regression lines showed improvements after data correction (Table 1) and as discussed before, it can complement the correlation coefficient. As seen in RNA-Seq data, very high correlations

Table 1 Reproducibility of the microarray data

Experiment ^a	Uncorrected data				After quantile correction			
	Average number of genes with >50% inter-replicate variation	$r_{(p)}$ ^b	Slope ^c	A.D ^d %	Average number of genes with >50% inter-replicate variation	$r_{(p)}$ ^b	Slope ^c	A.D ^d %
E-GEOD-53990: <i>Arabidopsis</i> data with 6 treatments and 3 replicates	8488/22810 (37%)	0.975	0.804	91%	2000/22810 (8.8%)	0.980	0.982	26%
E-GEOD-27822: Barley data with 2 treatments and 3 replicates	6433/22840 (28%)	0.992	1.22	34%	519/22840 (2.3%)	0.994	0.992	8.2%
E-GEOD-39950: Yeast data with 2 treatments and 3 replicates	1645/10928 (15%)	0.997	1.257	26%	396/10928 (3.6%)	0.997	0.980	12%
E-GEOD-49918: <i>E. coli</i> data with 4 treatments and 3 replicates	1059/10208 (10%)	0.996	1.032	28%	299/10208 (2.9%)	0.997	0.981	10%
E-GEOD-24524: <i>E. coli</i> data with 4 treatments and 2 replicates	837/10208 (8.2%)	0.995	0.999	22%	300/1028 (2.9%)	0.996	0.990	15%

Different publicly available microarray data were analyzed using the Affymetrix Expression Console build 1.3.1.187 following the 3 Expression Arrays-RMA protocol. Both the corrected and uncorrected data were extracted. The Pearson correlations between the replicates were calculated on the \log_2 transformed data. Following our method, the average of the ratios' deviations was also calculated in order to evaluate its usefulness compared to the correlation coefficient.

^aThe experiments are available at <http://www.ebi.ac.uk/arrayexpress/experiments/browse.html>, ^bThe average of Pearson correlations between the replicates in each experiment, ^cThe average of slopes of inter-replicate regression lines in each experiment, ^dThe average of the ratios' deviations.

(>0.98) and small slope deviations (<0.02) are possible while up to 10% of the genes show over 50% inter-replicate variations in microarray data (Table 1). In addition, not only do the slopes not show any considerable improvement after correction of the *E. coli* microarray data, but they also decreased while the number of genes with higher than 50% inter-replicate variation was decreased from 8-10% to 2.9% (Table 1). In contrast, our new measurements followed the increased reproducibility and showed considerable reduction in deviations after data correction (Table 1; Additional file 2). However, the deviations were still large enough to explain the existent variance after data correction in all the experiments.

It should be emphasized that the levels of reproducibility should not be compared among different experimental data. This is due to the data-dependency of the measured statistics. We found higher deviation values (15.3%, 14.2%, and 13.7%) for the yeast data when there were fewer genes with greater than 50% inter-replicate variance (4.4%, 4.3%, and 4.2%) compared to the rice data.

As indicated, we have included X/Y and Y/X as separate measurements of the genes between the treatments in order to fix the asymmetric aspect of the ratios around 1. We have not applied the well-known logarithmic transformation, i.e. $\log(X/Y) = \log(X) - \log(Y)$. The latter, can easily reduce the difference level between comparisons as we have already explained. Therefore, we will not see the actual difference at the reproducibility levels by applying the logarithmic transformation.

What can be done when reproducibility is low or not known?

The correlation coefficient or the alternative method introduced here shed light on general data reproducibility based on predefined thresholds. The main application of these approaches is in comparing different analytical methods such as correction measures with each other in order to use the most appropriate method. However, they do not provide detailed information as to which part of data has an acceptable reproducibility or what stringency level we should apply in our analysis to meet the low reproducibility levels in order to avoid making the data useless. Under the assumption of a similar natural variation of expression for all genes, i.e. since a gene with extreme inter-replicate variation for expression is assumed to be random, we advise applying a fold-change threshold to solve the problem. To be acceptable, the level of the statistically significant expression changes between the treatments should be not smaller than the predetermined threshold. This threshold should be determined by keeping the frequency of genes showing same or higher levels of inter-replicate changes lower than an acceptable likelihood of being false positive, i.e. 5% or 1%. Therefore, a significant inter-treatment fold-change of a gene should have an inter-replicate likelihood lower than either 1% or 5%. Applying this criterion keeps the usability gate open even when the reproducibility is low just by considering the reliable fold-changes. When analyzing the rice untreated root [GenBank ID: DRR000350 and DRR000354] and untreated shoot

[GenBank ID: DRR000349 and DRR000353] samples, we found the fold-change thresholds of >3.7 and >2.0 at 1% and 5% levels, respectively. By comparing untreated shoot with untreated root, we found 5212 (48% of the significant differentially expressed genes) and 783 (7.3% of the significant differentially expressed genes) genes showing same or lower levels of fold-changes according to the 1% and 5% cut-offs, respectively. We can also apply a genegroup specific fold-change threshold when there is a bias towards specific genes, e.g., by gene-length or expression-quantity based clustering of the genes due to the higher variation of shortlength or low-abundant genes compared to the long or high abundant genes, respectively, in RNA-Seq.

ORF length-based correction on protein quantities is necessary before correlating them with their transcript abundance

No strong correlation has been found between protein and the RNA abundance [1-4]. One reason could be the multiple levels of regulatory mechanisms, from transcription to post-translation, that govern the expression of genes. By excluding all the technical and biological noise as well as experimental biases, different determinants including ribosomal occupancy (the fraction of a given gene's transcripts engaged in translation at least by one ribosome) and ribosomal density (number of ribosomes per length-unit of transcripts) as translation activating factors, sequence features of translation initiation, elongation, and termination as ORF-specific translation-efficiency related factors, and amino acid availability as well as mRNA/protein half-life have been proposed to influence the mRNA-protein correlation [19-22]. Among these, ORF-specific translation-efficiency related factors explain 15-30% of the variation of mRNA-protein correlations [22-24]. The mRNA half-life does not show a considerable impact in contrast to the protein half-life, which explains $\approx 17\%$ of the mRNA-protein correlation variations [24].

For many reasons, however, biologists most typically rely on transcript abundance as a gauge for gene expression. But, how reasonable is it to expect a high correlation between the transcript and the protein levels? To answer, let us assume the two expressed genes, A and B, are governed by post-transcriptional-free regulation mechanisms. This means that they have the same stability, half-life, and translation efficiency for the transcripts and same stability and half-life for the proteins of genes A and B. It is still possible to have different rates of induction or suppression between the protein levels of genes when there is a similar induction or repression rate at the transcript levels of the genes. Similar rates of change at the transcript levels of the genes A and B could result in different rates of

change at the protein levels when the lengths of the coding sequences differ between the genes. It might be expected to find twice as much induction for the protein A compared to the protein B if the coding sequence of gene B is twice as long as that of the gene A. The longer protein B will take more time to be translated than the protein A. This can simply decrease the correlation between the protein and the RNA levels by uneven variation of the variables (the transcript level and the protein level). As an example, the induced protein levels for A and B were four and two times, respectively, if the transcript levels of the genes were induced two times. In this situation, there are uneven changes of two and four times at the protein levels (one of the variables) compared to a fixed two times change in transcript levels (another variable). Therefore, it is not reasonable to look at the agreement level between the RNA and their encoded protein quantities. Instead, we should consider the effects of treatments/tissues on the gene expression quantities regardless of gene expression assay. In conclusion, different levels of post-transcriptional regulations should not be interpreted as the sole reason for the medium-weak correlation between the protein and transcript levels due to the above-mentioned technical problem which could easily have a large impact on the issue, either positively or negatively.

To inspect the pure effects of post-transcriptional regulation we advise that protein quantities should be corrected for the length of coding sequences before correlating the protein and the transcript quantities with each other. We performed a length-based correction on the previously reported protein quantities before examining the correlation between protein and transcript quantities. For the 10 proteins of *Plasmodium falciparum* and their corresponding transcript abundance measured during six growth stages published in [2], the correlation showed a shift from -0.029 to 0.135 after the ORF length-based correction was applied for protein quantities (Additional file 1). We also examined two different expression datasets (Additional file 1): cell cycle with 173 genes and cell rescue with 115 genes in yeast where the expression quantities exist both at the protein and at transcript levels, reported by Greenbaum *et al.* [1]. After the length-based correction on protein quantities the Pearson correlations were changed from 0.71 to 0.41 and from 0.45 to 0.29 in cell cycle and cell rescue datasets, respectively.

To further analyze the impact of ORF length-based bias in quantitative analysis of transcript and protein expressions, we looked at the correlation of mRNA changes with protein changes between human and chimpanzee reported by Fu *et al.* [25]. We do not expect a negative correlation between changes of transcripts

and of proteins. This is due to the fact that cells try to consume as little energy as possible. In addition, an overall trend of decreased mRNAs and increased proteins is not reasonable. However, we found a correlation of -0.37 between mRNAs (chimpanzee/human ratios) and proteins (chimpanzee/human ratios) of 143 genes. The averages of five and three biological replicates were applied to calculate the mRNAs changes and the proteins changes, respectively (Additional file 1). In contrast, a more reasonable correlation of -0.11 was obtained by using the ORF length-based corrected protein changes (Additional file 1).

Increased ribosomal occupancy means increased usage of minimum physical space on transcripts for ribosomal functionality. This can alleviate the effect of different ORF lengths on the deviations of fold-changes of protein quantities from the fold-changes of transcript quantities. In agreement with this, a higher protein transcript correlation was reported for the transcripts with high ribosomal occupancy [1]. This indicates that the ORF length differences can easily affect the agreement between the transcripts and their coded protein quantities.

It is conceivable that the longer transcripts may be occupied by a higher number of ribosomes, which might alleviate the aforementioned effect of the ORF length on the correlation coefficient between protein and transcript abundance. However, the very weak correlation ($r_P = 0.22$, $r_S = 0.34$) between the ORF length and number of ribosomes on the transcripts and negativeweak correlation ($r_P = -0.39$, $r_S = -0.56$) between the ORF length and ribosomal occupancy, which we found (see Additional file 1) in the data published by Greenbaum *et al.* [1], introduce the number of ribosomes or the ribosomal occupancy of transcripts not as a random length-based event, but rather, as a post-transcriptional regulation mechanism. Not only there is no strong correlation between the transcript length and ribosome numbers but also longer transcripts show lower ribosomal density compared to shorter transcripts reported in [19]. This could be explained by the fact that the first 40 codons of transcripts have over three-fold higher ribosomal density compared with any given 40 codons long fragment from the +40th codon downstream region of transcripts [26]. In addition, trans-regulatory divergence and not ribosomal occupancy has recently been introduced as the force behind the differences in the yeast translational efficiency [27]. Therefore, there should not be any concern about the alleviated effect of ORF length on the correlation coefficient. Correction for ORF length on protein quantity seems to be useful to extract much pure post-transcriptional regulation effects, at least based on the available data.

Although we introduce ORF length as a factor affecting the mRNA-protein correlation, it should be highlighted

that our ORF length-based correction of protein quantities is too simplistic and cannot guarantee the purity of the variation of mRNA-protein correlation. Among the ORF-specific translation-efficiency factors, peptide elongation has been found as the major determinant to explain the variation of mRNA-protein correlation [22-24]. One conceivable reason could be the biological importance of the elongation process itself compared to the initiation and termination processes. However, the other key player could be the transcript/ORF length, which can easily influence the elongation process. Therefore, our introduced ORF length-based correction of protein quantities can be affected by the elongation-related regulatory variance and needs further development in order to extract the pure length-dependent variance from the variation of mRNA-protein correlation.

Methods

RNA-Seq data analyses

To evaluate the efficiency of correlation coefficient on real data, a publicly available rice RNA-Seq dataset was used in the analysis. The dataset included two replicates of root and shoot samples from plants before and after a one hour salinity stress treatment [GenBank ID: DRX000191, DRX000192, DRX000193, and DRX000194]. The software CLC Genomics Workbench (<http://www.clcbio.com>; Aarhus, Denmark) was used to map the reads on the rice Nipponbare reference genome to isolate different type of reads, i.e. total and uniexon and gene reads. The average of inter-replicate variations of selected reference genes were applied for the reference gene based data correction (data not shown). The differentially expressed genes were found as described by Robinson *et al.* [28] with a *p*-value cut-off of 0.01 and using the Benjamini & Hochberg method [29] for multiple testing correction.

Microarray data analysis

We also analysed four different microarray datasets which are publicly available in EMBL-EBI, ArrayExpress (<http://www.ebi.ac.uk/arrayexpress/>). The datasets studied included *Arabidopsis* data [E-GEOD-53990] with six treatments and three replicates, barley data [E-GEOD-27822] with two treatments and three replicates, yeast data [E-GEOD-39950] with two treatments and three replicates, and two *Escherichia coli* datasets [E-GEOD-49918 and E-GEOD-24524]. One *E. coli* dataset had four treatments and three replicates and another had four treatments and two replicates. We used the Affymetrix Expression Console build 1.3.1.187 to analyze the data; the 3 Expression Arrays-RMA was applied to extract the quantities both without any correction and together with quantile-based correction.

An alternative reproducibility test by measuring the mean and standard deviation statistics of intertreatment ratios of gene expression quantities

After measuring the descriptive statistics of mean and standard deviation for intertreatment expression ratios, we calculated the ratios of each statistic for a single replicate when compared to different replicates of another treatment. Perfect agreement between replicates would result in a ratio of 1. Therefore, any deviation could be an indicator of decreased reproducibility. The analytical procedure was as follows:

1. Measurement of the genes' expression ratios between the treatments

$T_{1iR(T1)}/T_{2iR(T2)}, T_{1iR(T1)}/T_{3iR(T3)}, T_{2iR(T2)}/T_{3iR(T3)}, \dots, T_{n-1iR(Tn-1)}/T_{niR(Tn)}$, and $T_{2iR(T2)}/T_{1iR(T1)}, T_{3iR(T3)}/T_{1iR(T1)}, T_{3iR(T3)}/T_{2iR(T2)}, \dots, T_{niR(Tn)}/T_{n-1iR(Tn-1)}$ where T represents the treatments (1, ..., n) and i represents the genes. $R_{(T1)}, \dots,$ and $R_{(Tn)}$ represent the replicates of treatments.

2. Computing the mean (M) and standard deviation (S) for each population of ratios

$M_{T_{1iR(T1)}/T_{2iR(T2)}}, \dots, M_{T_{n-1iR(Tn-1)}/T_{niR(Tn)}}$, and $T_{2iR(T2)}/T_{1iR(T1)}, \dots, M_{T_{niR(Tn)}/T_{n-1iR(Tn-1)}}$
 $S_{T_{1iR(T1)}/T_{2iR(T2)}}, \dots, S_{T_{n-1iR(Tn-1)}/T_{niR(Tn)}}$, and $S_{T_{2iR(T2)}/T_{1iR(T1)}}, \dots, S_{T_{niR(Tn)}/T_{n-1iR(Tn-1)}}$

3. Computing the ratios of the statistics between paired populations among which one of the replicates of one sample is compared to all the replicates in another sample.

For example, in case of $M_{T_{1iR(T1)}/T_{2iR(T2)}}$ we do as follows:

$[M_{T_{1i1}/T_{2i1}}]/[M_{T_{1i1}/T_{2i2}}], [M_{T_{1i1}/T_{2i1}}]/[M_{T_{1i1}/T_{2i3}}], \dots, [M_{T_{1i2}/T_{2i1}}]/[M_{T_{1i2}/T_{2i2}}], [M_{T_{1i2}/T_{2i1}}]/[M_{T_{1i2}/T_{2i3}}],$

The same calculation should be performed for the ratios of standard deviations. Perfect agreement between replicates would result in a ratio of 1 and deviations can be an indication of decreased reproducibility.

4. Measuring the variation of the ratios

Since we deal with ratios, the variations should be applied in order to calculate the deviations from 1. For that, we must use the ratios of X/Y when $X > Y$. If $X < Y$, then we must use $1/(X/Y)$.

5. Measuring the average and standard deviation of the deviations of ratios

These parameters could be utilized as reproducibility coefficients. There will be no deviation when there is a perfect reproducibility. In addition, the standard deviation of the deviations is an indication of how constant the reproducibility is when considering all treatments. To facilitate all the steps of the analyses, an Excel file (Additional file 3) is prepared for users.

It includes different sheets based on the number of treatments and replicates and it just need the data to be pasted in in order to calculate the reproducibility measurements. There is a manual sheet included as well. We analyzed the mentioned rice data including four treatments each with two replicates. Therefore, there were 48 populations of expression ratios for uncorrected data and the same for corrected data (see Additional file 1). To check the robustness of the method, we also applied the online-available yeast RNA-Seq dataset explained by Lipson *et al.* [18]. This includes expression data of 6387 genes reported as transcript per million (t.p.m) and were found expressed in all the three technical replicates of the samples (see Additional file 1). Furthermore, we analyzed the mentioned five different microarray datasets from *Arabidopsis*, barley, yeast, and *E. coli* (see Additional file 2).

Protein quantity correction in order to assess the correlation between the protein and transcript abundance

Translation time can vary among transcripts with different lengths. This could be a potential source of variation of mRNA-protein correlation. Here, we introduce a correction approach to expel this source of variation out of data. Table 2 represents an example of protein quantity correction. The ORF lengths could be corrected by dividing the ORF-of interest-length by the shortest ORF length in the data. The corrected lengths are the correction factors that act as multipliers on protein quantities prior to correlating them with respective transcript abundance (Table 2).

Conclusions

Here, we conclude to enhance the power of reproducibility assay of gene expression data by considering the

Table 2 Pure post-transcriptional regulation effects can be examined after correction of the protein quantities

Gene	TF ^a	ORF ^b (bp)	PF ^c	CF ^d	CPF ^e
1	2	1000	10	1	10
2	2	2000	9	2	18
3	2	3000	8	3	24
4	2	4000	7	4	28
5	2	5000	6	5	30
6	2	6000	5	6	30
7	2	7000	4	7	28
8	2	8000	3	8	24

The table represents an artificial example of similar transcript fold-change for all the genes with different lengths and different protein fold-changes.
^aObserved transcript fold-change, ^bCoding sequence length, ^cObserved protein fold-change, ^dCalculated correction factor (length/1000), ^eCorrected PF (PF.CF).

slopes of inter-replicate linear regression lines and narrow-intervals of scatter plots as well as the correlation coefficient. We also introduce an efficient alternative approach which can detect small changes among large datasets. This works based on the means and standard deviations of the expression ratios of genes. As a reliability assay, it is advised not to rely on the correlation between the transcript and protein levels. This is due to the differences among the ORF-lengths of genes which can result in a weak correlation without post-transcriptional regulatory mechanisms. Finally, we advise an ORF-length based correction of protein quantities to examine a much pure effect of post-transcriptional regulatory mechanisms.

Additional files

Additional file 1: It is an Excel file which contains all RNA-Seq data analyses plus RNA-protein datasets.

Additional file 2: It is an Excel file which contains all the microarray data analyses.

Additional file 3: This Excel file can be used as an easy tool in order to get the reproducibility measurements.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

BD conceived the research, designed and implemented the algorithm, performed data analyses, and drafted the manuscript. CNS supervised the study and drafted the manuscript. Both authors read and approved the final manuscript.

Acknowledgements

We thank Jennifer Hinds for formatting and line-editing work on the manuscript. We also thank Dr. Ilias Kappas from the editorial board of the journal and the anonymous reviewers for their useful comments. BD is supported by Aarhus University and University of Copenhagen. CNS is supported by University of Tennessee Ivan Racheff Endowment and AgResearch.

Author details

¹Department of Plant and Environmental Sciences, University of Copenhagen, Thorvaldsensvej 40, 1871 Frederiksberg, Denmark. ²Department of Molecular Biology and Genetics, Research Centre Flakkebjerg, Aarhus University, Forsøgsvej 1, 4200 Slagelse, Denmark. ³Department of Plant Sciences, University of Tennessee–Knoxville, 252 Ellington Plant Sciences, 37996-4561 Knoxville, TN, USA.

Received: 9 November 2013 Accepted: 7 March 2014

Published: 13 May 2014

References

- Greenbaum D, Colangelo C, Williams K, Gerstein M: Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biol* 2003, **4**:117.
- Le Roch KG, Johnson JR, Florens L, Zhou Y, Santrosyan A, Grainger M, Yan SF, Williamson KC, Holder AA, Carucci DJ, Yates JR, Winzeler EA: Global analysis of transcript and protein levels across the *Plasmodium falciparum* life cycle. *Genome Res* 2004, **14**:2308–2318.
- Ghazalpour A, Bennett B, Petyuk VA, Orozco L, Hagopian R, Mungrue IN, Farber CR, Sinsheimer J, Kang HM, Furlotte N, Park CC, Wen PZ, Brewer H, Weitz K, Camp DG, Pan C, Yordanova R, Neuhaus I, Tilford C, Siemers N, Gargalovic P, Eskin E, Kirchgessner T, Smith DJ, Smith RD, Lusic AJ: Comparative analysis of proteome and transcriptome variation in mouse. *PLoS Genet* 2011, **7**:e1001393.
- Ning K, Fermin D, Nesvizhskii AI: Comparative analysis of different label-free mass spectrometry based protein abundance estimates and their correlation with RNA-Seq gene expression data. *J Proteome Res* 2012, **11**:2261–2271.
- Wang Z, Gerstein M, Snyder M: RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009, **10**:57–63.
- Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y: RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* 2008, **18**:1509–1517.
- Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 2008, **5**:621–628.
- Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, Penkett CJ, Rogers J, Bähler J: Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* 2008, **453**:1239–1243.
- Tuch BB, Mitrovich QM, Homann OR, Hernday AD, Monighetti CK, De La Vega FM, Johnson AD: The transcriptomes of two heritable cell types illuminate the circuit governing their differentiation. *PLoS Genet* 2010, **6**:e1001070.
- Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L: Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol* 2011, **12**:R22.
- Quackenbush J: Computational analysis of microarray data. *Nat Rev Genet* 2001, **2**:418–427.
- Yang IV, Chen E, Hasseman JP, Liang W, Frank BC, Wang S, Sharov V, Saeed AI, White J, Li J, Lee NH, Yeatman TJ, Quackenbush J: Within the fold: assessing differential expression measures and reproducibility in microarray assays. *Genome Biol* 2002, **3**: research0062.1-0062.12.
- Hoen PA T, Ariyurek Y, Thygesen HH, Vreugdenhil E, Vossen RH, de Menezes RX, Boer JM, van Ommen GJ, den Dunnen JT: Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. *Nucleic Acids Res* 2008, **36**:e141.
- Ozsolak F, Goren A, Gymrek M, Guttman M, Regev A, Bernstein BE, Milos PM: Digital transcriptome profiling from attomole-level RNA samples. *Genome Res* 2010, **20**:519–525.
- Tang F, Barbacioru C, Nordman E, Li B, Xu N, Bashkurov VI, Lao K, Surani MA: RNA-Seq analysis to capture the transcriptome landscape of a single cell. *Nat Protoc* 2010, **5**:516–535.
- Quackenbush J: Microarray data normalization and transformation. *Nat Genet* 2002, **32**(suppl):496–501.
- Khan Z, Amini S, Bloom JS, Ruse C, Caudy AA, Kruglyak L, Singh M, Perlman DH, Tavazoie S: Accurate proteome-wide protein quantification from high-resolution 15 N mass spectra. *Genome Biol* 2011, **12**:R122.
- Lipson D, Raz T, Kieu A, Jones DR, Giladi E, Thayer E, Thompson JF, Letovsky S, Milos P, Causey M: Quantification of the yeast transcriptome by single-molecule sequencing. *Nat Biotechnol* 2009, **27**:652–658.
- Arava Y, Wang Y, Storey JD, Liu CL, Brown PO, Herschlag D: Genome-wide analysis of mRNA translation profiles in *Saccharomyces cerevisiae*. *Proc Natl Acad Sci U S A* 2003, **100**:3889–3894.
- Beyer A, Hollunder J, Nasheuer HP, Wilhelm T: Post-transcriptional expression regulation in the yeast *Saccharomyces cerevisiae* on a genomic scale. *Mol Cell Proteomics* 2004, **3**:1083–1092.
- Brockmann R, Beyer A, Heinisch JJ, Wilhelm T: Posttranscriptional expression regulation: what determines translation rates? *PLoS Comput Biol* 2007, **3**:e57.
- Nie L, Wu G, Zhang W: Correlation of mRNA expression and protein abundance affected by multiple sequence features related to translational efficiency in *Desulfovibrio vulgaris*: a quantitative analysis. *Genetics* 2006, **174**:2229–2243.
- Guimaraes JC, Miguel Rocha M, Arkin AP: Transcript level and sequence determinants of protein abundance and noise in *Escherichia coli*. *Nucleic Acids Res* 2014, doi:10.1093/nar/gku126.
- Wu G, Nie L, Zhang W: Integrative analyses of posttranscriptional regulation in the yeast *Saccharomyces cerevisiae* using transcriptomic and proteomic data. *Curr Microbiol* 2008, **57**:18–22.
- Fu N, Drinnenberg I, Kelso J, Wu JR, Pääbo S, Zeng R, Khaitovich P: Comparison of protein and mRNA expression evolution in humans and chimpanzees. *PLoS One* 2007, **2**:e216.

26. Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS: **Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling.** *Science* 2009, **324**:218–223.
27. McManus J, May GE, Spealman P, Shteyman A: **Ribosome profiling reveals post-transcriptional buffering of divergent gene expression in yeast.** *Genome Res* 2013, doi:10.1101/gr.164996.113.
28. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.** *Bioinformatics* 2010, **26**:139–140.
29. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *J R Stat Soc: Series B* 1995, **57**:289–300.

doi:10.1186/2241-5793-21-3

Cite this article as: Darbani and Stewart: **Reproducibility and reliability assays of the gene expression-measurements.** *Journal of Biological Research-Thessaloniki* 2014 **21**:3.

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit

