SCIENCE CHINA
Life Sciences

SPECIAL TOPIC
• RESEARCH PAPER •

February 2013 Vol.56 No.2: 134–142
doi: 10.1007/s11427-013-4437-9

mRNA enrichment protocols determine the quantification characteristics of external RNA spike-in controls in RNA-Seq studies

QING Tao†, YU Ying†, DU TingTing & SHI LeMing*

Center for Pharmacogenomics, School of Pharmacy, Fudan University, Shanghai 201203, China

Received December 23, 2012; accepted January 7, 2013

RNA-Seq promises to be used in clinical settings as a gene-expression profiling tool; however, questions about its variability and biases remain and need to be addressed. Thus, RNA controls with known concentrations and sequence identities originally developed by the External RNA Control Consortium (ERCC) for microarray and qPCR platforms have recently been proposed for RNA-Seq platforms, but only with a limited number of samples. In this study, we report our analysis of RNA-Seq data from 92 ERCC controls spiked in a diverse collection of 447 RNA samples from eight ongoing studies involving five species (human, rat, mouse, chicken, and *Schistosoma japonicum*) and two mRNA enrichment protocols, i.e., poly(A) and RiboZero. The entire collection of datasets consisted of 15650143175 short sequence reads, 131603796 (i.e., 0.84%) of which were mapped to the 92 ERCC references. The overall ERCC mapping ratio of 0.84% is close to the expected value of 1.0% when assuming a 2.0% mRNA fraction in total RNA, but showed a difference of 2.8-fold across studies and 4.3-fold among samples from the same study with one tissue type. This level of fluctuation may prevent the ERCC controls from being used for cross-sample normalization in RNA-Seq. Furthermore, we observed striking biases of quantification between poly(A) and RiboZero which are transcript-specific. For example, ERCC-00116 showed a 7.3-fold under-enrichment in poly(A) compared to RiboZero. Extra care is needed in integrative analysis of multiple datasets and technical artifacts of protocol differences should not be taken as true biological findings.

RNA-Seq, External RNA Control Consortium (ERCC), MAQC/SEQC, mRNA enrichment protocol, quality control, reproducibility, quantification bias, poly(A) versus RiboZero

Citation: Qing T, Yu Y, Du T T, et al. mRNA enrichment protocols determine the quantification characteristics of external RNA spike-in controls in RNA-Seq studies. Sci China Life Sci, 2013, 56: 134−142, doi: 10.1007/s11427-013-4437-9

RNA-Seq provides a new way at the genome-wide scale to measure gene expression profiles, distinguish isoforms, identify new transcripts, and make many other biomedical applications possible [1–3]. RNA-Seq has shown its accuracy for the quantification of gene expression levels, as determined by quantitative PCR. Its high level of technical reproducibility has also been demonstrated [4–6]. Despite these advantages, recent analyses have revealed larger technical variability for the quantification of genes expressed at lower levels [7] and biases introduced by RNA-Seq technologies and the normalization methods [8–10]. However, most of these studies lack the absolute "truth" to objectively assess the performance and biases of RNA-Seq technology and the impact of data analysis approaches.

The External RNA Control Consortium (ERCC), a group of about 70 scientists from private, public and academic organizations led by the National Institute of Standards (NIST), developed a series of spike-in RNA transcripts with

†Contributed equally to this work
*Corresponding author (email: lemingshi@fudan.edu.cn)

known concentrations and sequence identities and originally proposed for their use in quality assessment/control of gene expression platforms such as microarrays and qPCR [11−13]. Specifically, these external RNA standard controls can provide a powerful tool for comparing relationship between the measured signal response and the known input concentration of the controls. A good linear relationship would be an indication of reliable estimation of the expression measurements for endogenous RNA transcripts [11,12, 14].

Recently, the ERCC spike-in controls of known concentrations and sequences have been proposed for assessing the performance of RNA-Seq technology and data analysis approaches [15–17]. However, the type and the number of RNA samples spiked with ERCC controls are limited. Before the ERCC controls can be routinely and reliably used in RNA-Seq studies for quality monitoring, their behavior in more diverse and biologically relevant RNA matrices needs to be investigated.

In this study, we analyzed a collection of eight RNA-Seq datasets consisting of 447 unique total RNA samples isolated from five species (*Homo sapiens*, *Rattus norvegicus*, *Mus musculus*, *Gallus gallus*, and *Schistosoma japonicum*). Each of these 447 RNA samples was spiked with one of the two ERCC control mixes according to the manufacturer's guidelines (Life Technologies, see Materials and methods for more details). For most samples, one of the two mRNA enrichment protocols, poly(A) selection and RiboZero, was used before Illumina sequencing. For some RNA samples, the two mRNA enrichment protocols were applied and replicate libraries were made under each protocol. In addition, each library was sequenced on multiple lanes of a flow cell. The datasets allowed us to examine the behavior of the ERCC controls under diverse environments and should provide helpful information for the practical applicability of ERCC controls in RNA-Seq studies. On the other hand, the datasets also provided an opportunity for us to objectively assess the performance of the Illumina platform.

1 Materials and methods

1.1 ERCC Mix1 and Mix2

The two ERCC control mixtures (Mix1 and Mix2) were purchased from Life Technologies Corp. (Carlsbad, California, USA; http://www.invitrogen.com/). They are preformulated pools of 92 polyadenylated transcripts with length of 250–2000 nt and a span of approximately 10^6-fold difference in concentration. Each of the two mixtures contains the same set of 92 ERCC transcripts which are separated into four sub-pools (A, B, C, and D) each with 23 transcripts. The difference between Mix1 and Mix2 is the relative amounts of the four sub-pools, resulting in pre-defined Mix1:Mix2 concentration ratios of 4.0, 1.0, 0.67, and 0.50 for sub-pools A, B, C and D, respectively. Each RNA sample was spiked in with an appropriate amount of either Mix1 or Mix2 according to Life Technologies' guidelines which would lead to about 1% of the total number of RNA-Seq reads mapping to the 92 ERCC control sequences, assuming the mRNA fraction in the total RNA is 2%. Overall, about half of the 447 RNA samples were spiked with Mix1 and the other half with Mix2. The use of Mix1 and Mix2 in the same study also allows one to assess the accuracy of fold changes estimated by RNA-Seq.

1.2 Eight datasets with two mRNA enrichment protocols and five species

The eight RNA-Seq datasets with ERCC spike-ins are summarized in Table 1. They came from our ongoing RNA-Seq studies involving five species (*Homo sapiens*, *Rattus norvegicus*, *Mus musculus*, *Gallus gallus*, and *Schistosoma japonicum*) with 447 unique total RNA samples, each of which was pre-spiked with ERCC Mix1 or Mix2 before library preparation and sequencing in such a way so that about 1% of the sequenced reads is expected to map to the sequences of the 92 ERCC transcripts (assuming a 2% mRNA content in the total RNA [18]). Both poly(A) selection and RiboZero protocols were used for mRNA enrichment. Standard Illumina library preparation and sequencing protocols were used throughout these studies. Briefly, cDNA libraries were prepared for each sample and sequenced using Illumina's TruSeq Cluster v3 and TruSeq SBS kit v3. Usually, multiple indexed libraries from the same study were pooled together and loaded for sequencing on multiple lanes of an Illumina flow cell using either an Illumina HiScan or HiSeq 2000 sequencer. Library preparation and sequencing were conducted at the City of Hope Functional Genomics Core (Duarte, California, USA), the University of Texas Southwestern Medical Center's Microarray Core (Dallas, Texas, USA), or Expression Analysis Inc. (Durham, North Carolina, USA).

For most samples, only one library was built; however, for the purpose of estimating library reproducibility, two replicate libraries were constructed per sample for study Rn_RZ_2 and study Rn_PA. This allowed us to compare the repeatability and reproducibility between different studies and at different replication levels. The statistics of reads mapping to ERCC reference sequences in each study are listed in Table 1. To the best of our knowledge, this represents the largest and most diverse collection of RNA-Seq datasets with ERCC spike-in mixes so far.

1.3 Reads mapping and normalization

RNA-Seq reads were aligned to the ERCC reference sequences by using Bowtie2 [19], with a maximum of 2 mismatches allowed. The other default parameter settings were used. The alignment results were then processed using SAMtools [20] to filter the mapped reads, followed by

Table 1 Summary of eight RNA-Seq datasets with ERCC spike-in controls[a)]

Study	Species*	mRNA enrichment method	Number of RNA samples	Sequencing format***	Raw reads per RNA sample	Total raw reads per study	ERCC mapped reads per RNA sample	Total ERCC mapped reads per study	Mapping ratio (Min)	Mapping ratio (Max)	Mapping ratio (Max/Min)	Mapping ratio (Mean)	Mapping ratio (SD)
Hm_PA	Hs	Poly(A)	20	50 SE	21782082	435641643	196156	3923113	0.61%	1.52%	2.51	0.90%	0.25%
Rn_RZ_1	Rn	RiboZero	320	50 SE	41333962	13226867952	363409	116290946	0.35%****	3.46%	9.88	0.92%	0.35%
Rn_RZ_2	Rn	RiboZero	2**	100 PE	94308066	188616132	660639	1321278	0.69%	0.71%	1.03	0.70%	0.01%
Rn_PA	Rn	Poly(A)	2**	100 PE	113520882	227041764	540195	1080389	0.42%	0.55%	1.32	0.48%	0.09%
Ms_PA_1	Mm	Poly(A)	16	50 PE	20353995	325663925	79468	1271490	0.20%	0.79%	3.84	0.40%	0.16%
Ms_PA_2	Mm	Poly(A)	27	50 PE	12480732	336979765	137508	3712728	0.64%	2.73%	4.30	1.14%	0.47%
Gg_PA	Gg	Poly(A)	36	101 SE	17775776	639927940	70299	2530776	0.19%	0.80%	4.27	0.40%	0.18%
Sj_PA	Sj	Poly(A)	24	50 SE	11225169	269404054	61378	1473076	0.25%	1.04%	4.22	0.56%	0.25%
	Total		447			15650143175		131603796					

a) *, Hs: *Homo sapiens*; Rn: *Rattus norvegicus*; Mm: *Mus musculus*; Gg: *Gallus gallus*; Sj: *Schistosoma japonicum*; **, these RNA samples were obtained from the same biological tissues; therefore, they are aliquots of the same RNA isolation; ***, SE: single-end; PE: paired-end; ****, two samples showed ERCC mapping ratios of <0.001%; it was suspected that no ERCC mix was spiked in due to experimental errors.

BEDTools [21] for counting reads mapped to each ERCC transcript. For RNA samples with multiple replicating libraries and/or sequencing replicates (i.e., loaded on multiple lanes), the total number of reads for each ERCC transcript was calculated and normalized using the following formula:

$$X_{i,j} = \log_2\left(\Sigma_1^m R_{i,j} \times \frac{(\Sigma_1^n \Sigma_1^m \Sigma_1^{92} R_{i,j})/n}{\Sigma_1^m \Sigma_1^{92} R_{i,j}} + 1\right),$$

where i represents the ith ERCC transcript (from 1 to 92); j is the jth RNA sample in a study (from 1 to 447); m is the number of sequencing replicates per RNA sample; n is the number of total RNA samples, and $R_{i,j}$ is the total number of reads mapped to ERCC transcript i in RNA sample j. $X_{i,j}$ is the normalized expression value for ERCC transcript i in sample j.

1.4 Data analysis methods

The ERCC mapping ratio for a given RNA sample was calculated as the number of reads mapped to the 92 ERCC transcripts divided by the total number of reads sequenced for that RNA sample. Principal component analysis (PCA), principal variance component analysis (PVCA), ERCC concentration versus RNA-Seq quantification curves, and Pearson correlation coefficient were performed in the R (www.r-project.org/) environment with its "base" functions and "stat" packages. The normalized reads counts in $\log_2$ scale defined above in Section 1.3 were used for PCA, PVCA and Pearson correlation by "prcomp" and "cor" function in the "stat" and "lme4" packages. The linear model fitting plots were generated with the "ggplots2" packages (http://had.co.nz/ggplot2/ book).

2 Results

2.1 The percentage of reads mapped to ERCC spike-in controls varies among studies and samples

We first examined the percentage of RNA-Seq reads from an RNA sample mapped to the 92 ERCC control sequences. Figure 1A shows the individual percentages for the 447 samples and Table 1 lists the summary information of the mapping ratio by study. As expected, for most samples the ERCC reads accounted for around 1% of the total sequenced reads. However, the mean mapping ratio did show appreciable differences among various studies, ranging from 0.40% for study Gg_PA with *Gallus gallus* to 1.14% for study Ms_PA_2 with *Mus musculus*, representing a 2.8-fold difference across studies. We also observed large variations in ERCC mapping ratio for some samples in the same study. For example, in the Ms_PA_2 study, the highest mapping ratio is 2.73%, whereas the lowest mapping ratio is only 0.64%, representing a 4.3-fold difference among samples. In addition, in the Rn_RZ_1 study one sample showed a very high mapping ratio of 3.46%, whereas the lowest mapping ratio in the same study is only 0.35% (excluding the two samples for which ERCC mixtures apparently were not spiked in), representing a 9.9-fold difference.

We further summarized the Rn_RZ_1 study data by separating the RNA samples according to tissue type (Table 2). The mean ERCC mapping ratio per tissue type varied from 0.67% for adrenal gland to 1.49% for liver, or a 2.2-fold difference across tissue types. Furthermore, the ERCC mapping ratio varied up to 5.0-fold across the 32 liver samples.

2.2 Principal component analysis shows clear grouping of samples by the type of spike-in mixtures and by mRNA enrichment protocols

PCA was performed on the 92 by 447 matrix of ERCC

Table 2 Summary of the Rn_RZ_1 study consisting of 11 tissue types with ERCC spike-in controls

Study	Tissue	Number of RNA samples	Raw reads per RNA sample	Total raw reads per tissue	ERCC mapped reads per RNA sample	Total ERCC mapped reads per tissue	Mapping ratio (Min)	Mapping ratio (Max)	Mapping ratio (Max/Min)	Mapping ratio (Mean)	Mapping ratio (SD)
Rn_RZ_1	Adrenal gland	32	43786644	1401172623	288987	9247569	0.35%	0.92%	2.61	0.67%	0.11%
	Testes	16	45205366	723285852	308164	4930625	0.41%	1.24%	2.99	0.69%	0.22%
	Kidney	32	47282344	1513035020	352091	11266904	0.55%	1.20%	2.18	0.76%	0.20%
	Lung	32	43764138	1400452416	334477	10703250	0.53%	1.08%	2.03	0.77%	0.13%
	Brain	32	44040794	1409305414	337998	10815923	0.50%	1.17%	2.34	0.79%	0.17%
	Heart	32	48174059	1541569887	409255	13096164	0.55%	1.70%	3.11	0.85%	0.24%
	Thymus	32	38897728	1244727299	354675	11349614	0.50%	2.31%	4.66	0.91%	0.34%
	Uterus	16	43672502	698760031	404171	6466729	0.58%	2.35%	4.09	0.96%	0.46%
	Muscle	32	34291628	1097332090	358838	11482824	0.75%	1.58%	2.11	1.05%	0.20%
	Spleen	32	37445301	1198249628	392967	12574932	0.61%	1.40%	2.30	1.06%	0.28%
	Liver	32	31218053	998977692	448638	14356412	0.69%	3.46%	5.02	1.49%	0.47%
Total				13226867952		116290946					

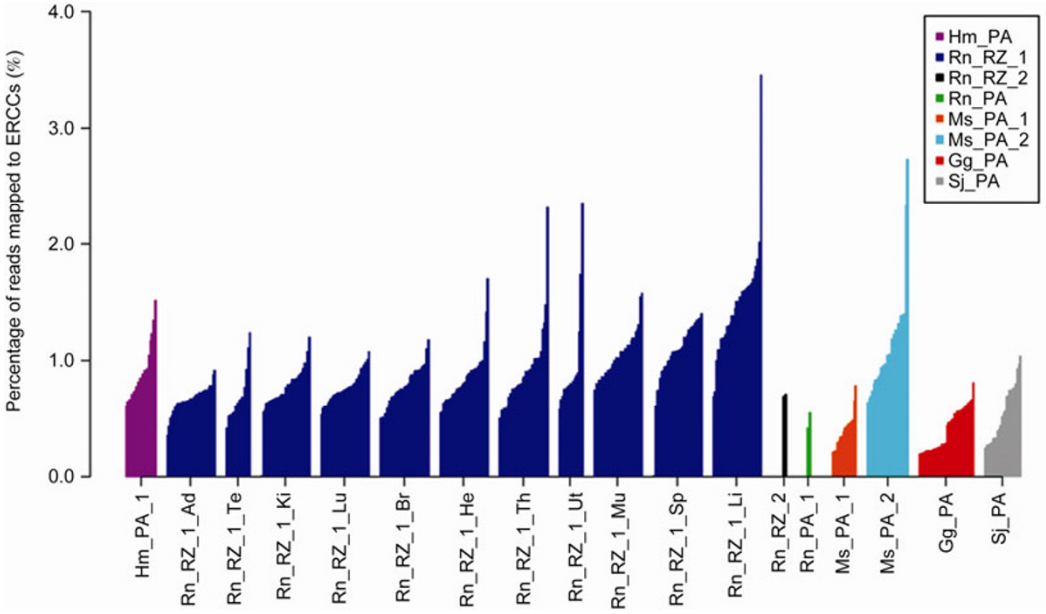

Figure 1 Percentage of RNA-Seq reads mapped to the 92 ERCC controls. It was calculated by dividing the number of reads mapped to the 92 ERCC reference sequences by the total number of sequence reads collected for an RNA sample. The 447 RNA samples were ordered first by study, then by tissue type (for study Rn_RZ_1 only), and lastly by the mapping ratio (percentage). Samples in study Rn_RZ_1 (colored in blue) were separated into 11 tissue types that were sorted by the mean mapping ratio per tissue type.

quantification data; the proportion of the total variance explained by the first and second principal components (PC1 and PC2) is 31.8% and 19.6%, respectively (Figure 2A). Samples were firstly grouped according to the type of ERCC mixtures (PC1), as expected due to the different compositions of the Mix1 and Mix2 mixtures, then by different RNA enrichment protocols (PC2). For the Rn_PA and Rn_RZ_2 studies, the same RNA samples were enriched for mRNA by the poly(A) selection and RiboZero protocols, and the resulting ERCC expression profiles (samples from Rn_RZ_2 were colored in black and green for Rn_PA) clustered together with samples enriched using the same mRNA enrichment protocol. This shows that the choice of mRNA enrichment protocols plays an important role on the quantification characteristics of the ERCC controls as determined by RNA-Seq.

2.3 Principal variance component analysis

To further quantitatively understand the factors that contribute to the variation in ERCC quantification profiles, we performed PVCA based on all 447 samples. The results were shown in Figure 2B. The ERCC "Mixture Type" explained 49.2% of the total variance, followed by mRNA enrichment protocols that accounted for 19.5% of the total

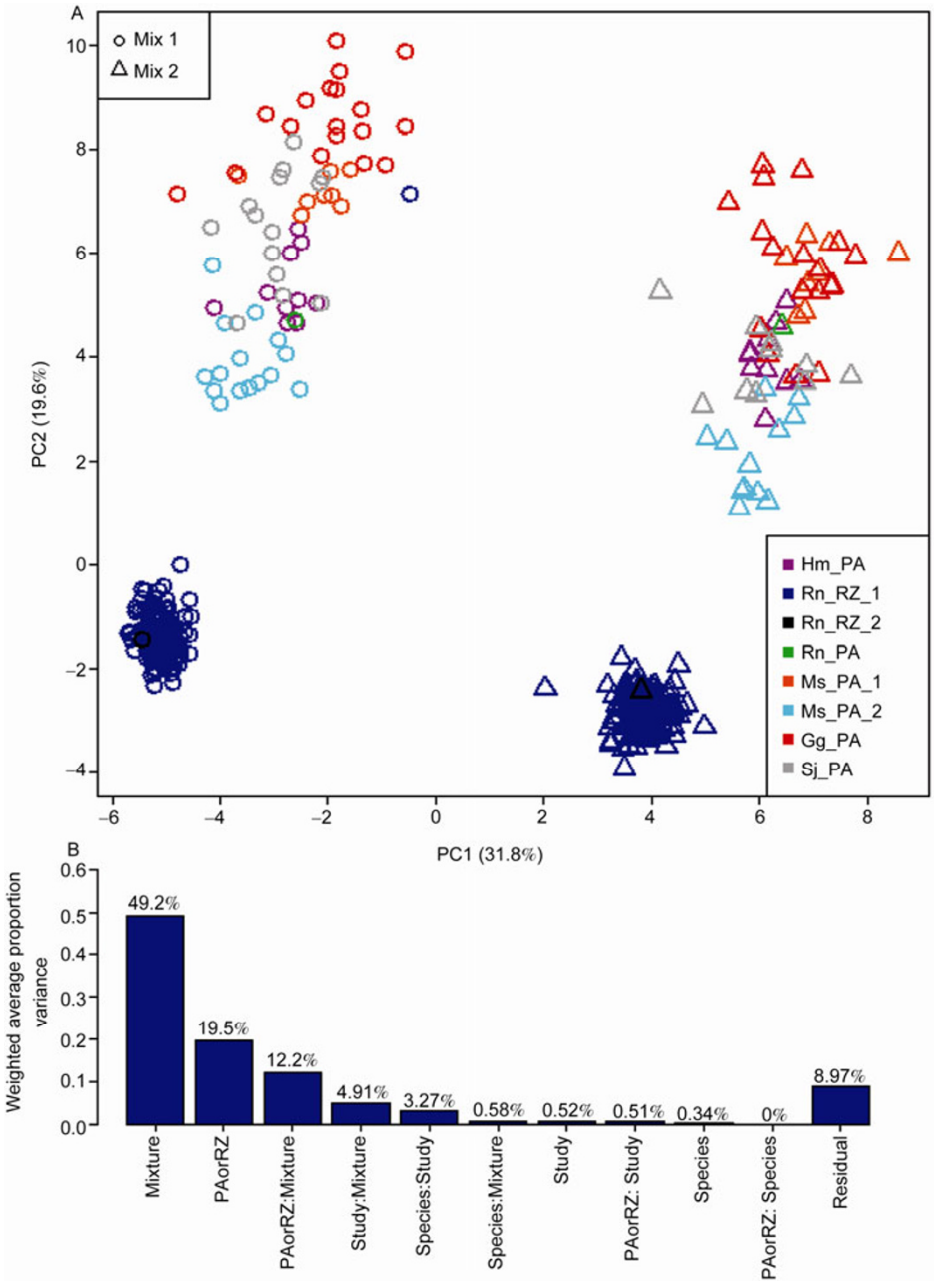

Figure 2 Factors impacting the variation in the quantification of ERCC spike-in controls. A, Principal component analysis shows that the 447 samples were firstly clustered by mixture type (Mix1 or Mix2), then by mRNA enrichment protocol (RiboZero or poly(A)). Studies Rn_PA and Rn_RZ_2 (colored in green and black, respectively) represents data from applying the two mRNA enrichment protocols to the same RNA. B, Principal variance component analysis quantifies the proportion of variance explained by each experimental factor.

variance. Intriguingly, the variance explained by study or species alone is quite low and accounted for only 0.52% and 0.34%, respectively, much smaller than the residuals (8.97%). We also observed that the interactions of mixture type with mRNA selection protocol and study accounted for 12.2% and 4.91% of the total variance, respectively. These results further demonstrated that the choice of mRNA enrichment protocols dramatically impact the behavior (quantification) of ERCC control transcripts by RNA-Seq.

2.4 Repeatability and reproducibility of ERCC spike-in abundance estimated by RNA-Seq

To further understand the degree of variation resulting from different experimental factors, we assessed the consistency of expression values of the 92 ERCC controls under different experimental scenarios. The scatterplots in Figure 3 show the repeatability of sequencing data collected from the same library but on two different lanes of the same flow cell (A, sequencing replicates), the repeatability of two libraries

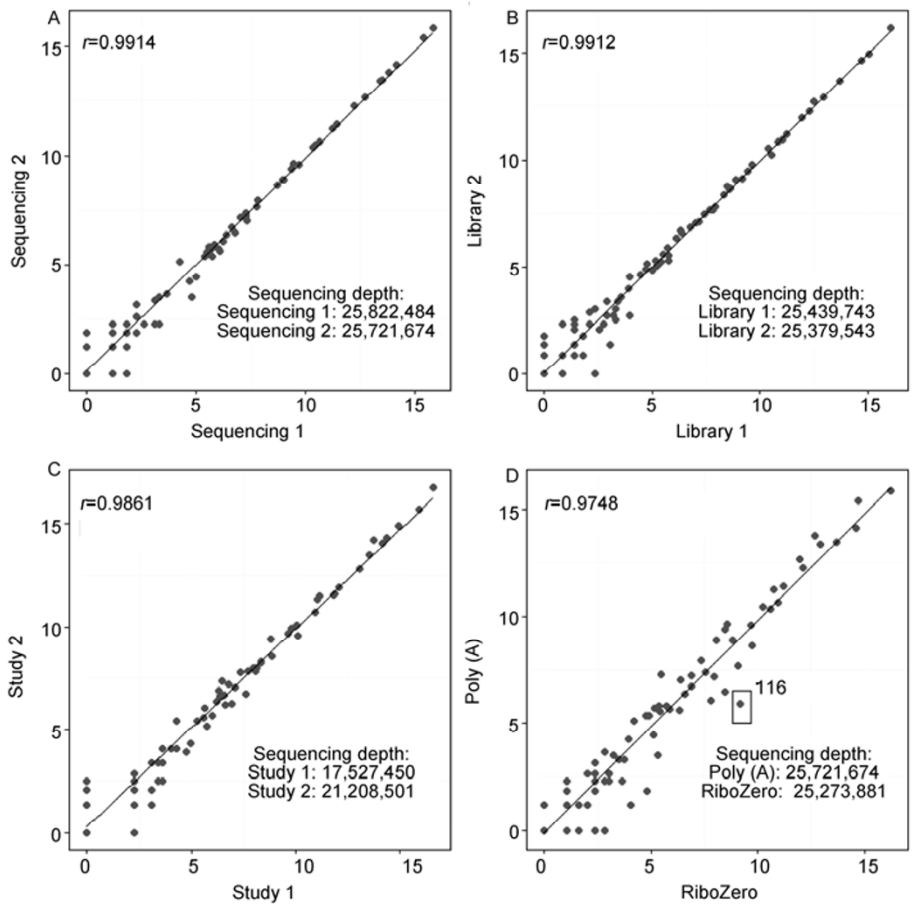

Figure 3 Consistency of RNA-Seq data at different replication levels including sequencing, library, study, and mRNA enrichment protocol. A, Two different sequencing replicates (two lanes on the same flow cell) of the same library. B, Two libraries constructed from the same RNA sample. C, Two different studies using the same poly(A) protocol. D, Two different mRNA enrichment protocols. Several ERCC-transcripts (e.g., ERCC-00116) show significantly different RNA-Seq signals between the two protocols. Each dot represents one of the 92 ERCC spike-in controls. Mix2 data were plotted.

constructed from the same RNA sample (B, library replicates), the reproducibility of quantification of ERCC controls measured in two unrelated studies using the same poly(A) selection protocol (C, different studies), and the consistency between poly(A) and RiboZero based libraries from the same RNA sample (D, different mRNA enrichment protocols). A high correlation coefficient was obtained between sequencing technical replicates (Pearson's correlation coefficient r=0.9914). The correlation between two different libraries from same sample is slight lower than that between sequencing technical replicates (r=0.9912). This indicates a good repeatability of sequencing technology and library preparation. In addition, a reasonably good correlation (r=0.9861) between two different studies suggests that the quantification of ERCC controls used in different studies are comparable.

2.5 Poly(A) versus RiboZero enrichment protocols

We also compared the replicate samples with different mRNA selection protocols (Figure 3D). Although there is a good overall correlation (r=0.9748), several ERCC controls such as ERCC-00116 behaved dramatically different between the two mRNA enrichment protocols. For example, the RNA-Seq estimated abundance for ERCC-00116 in the poly(A) protocol was 7.3-fold (p=1.4×10^{-18}) lower compared to that in the RiboZero protocol. An important difference between Figure 3D and the other three panels in Figure 3 is that the inconsistency between the two mRNA enrichment protocols occurs not only for ERCC controls at low concentrations due to low abundance (lower-left) but also for controls with high concentrations (upper-right). This indicates that the deviation between the two protocols is not simply because of the sequencing depth but something inherently different between the two protocols.

2.6 Relationship between RNA-Seq counts and ERCC concentrations

In Figure 4, we plotted the normalized RNA-Seq counts versus the concentrations for the 92 ERCC spike-in controls for RiboZero (Figure 4A) and poly(A) (Figure 4B) data.

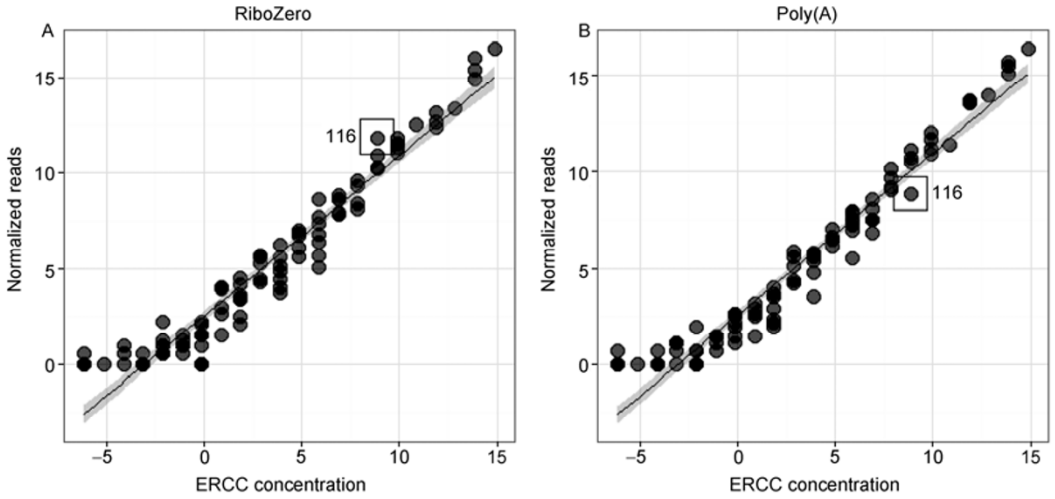

Figure 4 Relationship between ERCC control concentration and RNA-Seq measured signal in $\log_2$ scale. A, RiboZero protocol; the gray area represents the range of standard error and black line is the liner model curve. B, poly(A) protocol. Each dot represents one of the 92 ERCC spike-in controls. Mix1 data were plotted.

There is an overall linear relationship between RNA-Seq detected signal and the true concentration of the ERCC controls, especially for controls with higher concentrations. At the lower concentration end (lower-left), RNA-Seq was unable to detect concentration gradient. Interestingly, for ERCC-00116 the RNA-Seq detected signal with poly(A) protocol was almost 4-fold lower than expected, whereas the RNA-Seq signal with RiboZero was almost 2-fold higher than expected. These two opposite biases led to a 7.3-fold difference between the two mRNA enrichment protocols for ERCC-00116 shown in Figure 3D.

3 Discussion

The external RNA controls originally developed for monitoring the performance of microarray and qPCR platforms of gene expression have recently been proposed for the same purpose for RNA-Seq, a rapidly-evolving technology with the potential for more accurate and reliable gene expression measurement. Our study included a collection of eight datasets consisting of 447 unique RNA samples from five species pre-spiked with ERCC control mixes. To the best of our knowledge, this collection of datasets represents the largest and most diverse RNA-Seq analysis to date with external RNA standards and the observations from these datasets on the quantification characteristics of the ERCC controls should become useful for future applications and interpretation of RNA-Seq data.

We observed dramatic impact of mRNA enrichment methods on the behavior of ERCC controls in the RNA-Seq experiments. Specifically, the poly(A) enrichment protocol resulted in under selection of certain ERCC controls, most notably for ERCC-00116 that showed a 7.3-fold of under enrichment in the poly(A) protocol compared to the RiboZero protocol. It should be pointed out that the poly(A) selection biases observed in this study is not unique to RNA-Seq, because the poly(A) selection procedure is also widely used in microarray-based gene expression studies. Our observation on poly(A) selection biases has several implications in the applications of gene expression data in biomedical research. First, extra care must be exercised in the integrative analysis and interpretation of multiple gene-expression datasets resulting from different mRNA enrichment protocols. Otherwise, research findings may be merely artifacts of protocol differences rather than true biological signal that we set to discover. Secondly, although there is in general a good agreement between the concentration of spike-in controls and the RNA-Seq reads count, it is difficult to accurately estimate the absolute expression level of RNA transcripts, because the level of RNA-Seq measurement biases is clearly protocol dependent and transcript specific. Future attempts for correcting protocol dependent and transcript-specific biases are welcome but will be challenging. Thirdly, a decision to switch the mRNA enrichment protocol in the middle of a large study should not be made lightly and its implications should be adequately assessed.

The entire collection of datasets used in this analysis consists of 15650143175 RNA-Seq reads, 131603796 (i.e., 0.84%) of which were successfully mapped to the 92 ERCC reference sequences. This ERCC mapping ratio (0.84%) is close to what is expected from the manufacturer's guidelines on the use of the ERCC controls. However, the actual ERCC mapping ratio showed significant level of variation among studies, ranging from 0.40% to 1.14%, or a 2.8-fold difference across studies. In addition, the ERCC mapping ratio among samples from the same study involving only

one tissue type showed up to a 4.3-fold difference between the maximum and minimum ratios (Table 1, excluding study Rn_RZ_1). The difference in ERCC mapping ratio among various studies may reflect the inherent difference in mRNA fraction in the type of RNA samples profiled. Indeed, in the Rn_RZ_1 study where 11 different rat tissue types were separately analyzed (Table 2), we observed a clear tissue-dependence of ERCC mapping ratio: the ratio is the highest for liver (1.49%) and the lowest for adrenal gland (0.67%). On the other hand, the fluctuation of ERCC mapping ratio among samples in the same study involving only one tissue type may reflect the true accumulated technical variability in measuring RNA concentrations, pipetting RNA samples and ERCC controls, the completeness of mRNA enrichment, and preparing and sequencing the libraries. It seems to us that an up to 4.3-fold difference in ERCC mapping ratio may prevent the effective use of ERCC reads count as a basis for cross-sample normalization in RNA-Seq.

Another interesting observation in our study is that two RNA samples in the Rn_RZ_1 study were clearly not spiked with any ERCC mixes (Table 1), because their mapping ratio was <0.001%, which is at the level of mapping background (data not shown). Mislabeling of sample identity, which may be resulting from mis-pipetting, presents a common problem in the four large RNA-Seq studies in the US Food and Drug Administration (FDA) initiated SEquencing Quality Control (SEQC) project (http://www.fda.gov/MicroArrayQC/), which is the third phase of the FDA initiated, community-wide MicroArray Quality Control (MAQC) project [22,23]. In the SEQC project, it was estimated that 2%–3% of the ~2000 samples were mis-labeled or mis-pipetted (Shi L, personal communications) based on the availability of technical and biological replicates and cross-check of the sample annotation information. This raises serious concerns on the need of minimizing and preventing the occurrence of sample mislabeling in clinical settings where RNA-Seq and other diagnostics platforms are expected to be used as routine equipment and no technical replicates are usually available for assisting the identification of mislabeling.

It should be pointed out that relative, not absolute gene-expression analysis is the main use of microarray and RNA-Seq gene profiling. In this study we did not examine the behavior of ERCC controls in terms of differential expression. In addition, we did not investigate the molecular reasons leading to the transcript-specific biases under different mRNA enrichment protocols. Furthermore, another potentially useful utility of the ERCC controls is to help identify outlier samples in an RNA-Seq study, but we failed to identify any obvious outlying samples in our studies. Therefore, our current study cannot conclude on the effectiveness of ERCC controls for outlier identification purposes. Nevertheless, the amount of new information that can be gained from the ERCC spike-in controls justifies their continued uses in future RNA-Seq studies so that the behavior of the spike-in controls themselves and the endogenous RNA transcripts under more diverse experimental conditions can be better evaluated. The ERCC spike-in data presented in this study should also prove useful for evaluating the performance of different sequencing platforms and various data analysis approaches.

The ERCC spike-in data reported in this study were extracted from ongoing RNA-Seq projects in collaboration with our collaborators. We are grateful to them for allowing us to use the ERCC data in this study.

1 Xuan J, Yu Y, Qing T, et al. Next-generation sequencing in the clinic: promises and challenges. Cancer Lett, 2012, doi: 10.1016/j.canlet.2012.11.025
2 Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet, 2009, 10: 57–63
3 Mutz K, Heilkenbrinker A, Lönne M, et al. Transcriptome analysis using next-generation sequencing. Curr Opin Biotechnol, 2012, 24: 1–9
4 Nagalakshmi U, Wang Z, Waern K, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science, 2008, 320: 1344–1349
5 Cloonan N, Forrest AR, Kolle G, et al. Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nat Methods, 2008, 5: 613–619
6 Marioni J C, Mason C E, Mane S M, et al. RNA-Seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res, 2008, 18: 1509–1517
7 McIntyre L M, Lopiano K K, Morse A M, et al. RNA-Seq: technical variability and sampling. BMC Genomics, 2011, 12: 293
8 Schwartz S, Oren R, Ast G. Detection and removal of biases in the analysis of next-generation sequencing reads. PLoS ONE, 2011, 6: e16685
9 Zheng W, Chung L M, Zhao H. Bias detection and correction in RNA-sequencing data. BMC Bioinformatics, 2011, 12: 290
10 Zhang J X, Coombes K R. Sources of variation in false discovery rate estimation include sample size, correlation, and inherent differences between groups. BMC Bioinformatics, 2012, 13: S1
11 Tong W, Lucas AB, Shippy R, et al. Evaluation of external RNA controls for the assessment of microarray performance. Nat Biotechnol, 2006, 24: 1132–1139
12 Kralj J G, Salit M L. Characterization of *in vitro* transcription amplification linearity and variability in the low copy number regime using External RNA Control Consortium (ERCC) Spike-ins. Anal Bioanal Chem, 2013, 405: 315–320
13 Baker S C, Bauer S R, Beyer R P, et al. The External RNA Controls Consortium: a progress report. Nat Methods, 2005, 2: 731–734
14 Devonshire A S, Elaswarapu R, Foy C A. Evaluation of external RNA controls for the standardisation of gene expression biomarker measurements. BMC Genomics, 2010, 11: 662
15 Jiang L, Schlesinger F, Davis C A, et al. Synthetic spike-in standards for RNA-Seq experiments. Genome Res, 2011, 21: 1543–1551
16 Loven J, Orlando D A, Sigova A A, et al. Revisiting global gene expression analysis. Cell, 2012, 151: 476–482
17 Zook J M, Samarov D, McDaniel J, et al. Synthetic spike-in standards improve run-specific systematic error analysis for DNA and RNA sequencing. PLoS ONE, 2012, 7: e41356
18 Warrington J A, Corbisier P, Feilotter H, et al. Use of external RNA controls in gene expression assays: approved guideline. CLSI

document MM16-A (ISBN 1-56238-617-4), Wayne, Peennsylvania, USA, 2006
19 Langmead B, Salzberg S L. Fast gapped-read alignment with bowtie 2. Nat Methods, 2012, 9: 357–359
20 Ramirez-Gonzalez R H, Bonnal R, Caccamo M, et al. Bio-samtools: ruby bindings for samtools, a library for accessing bam files containing high-throughput sequence alignments. Source Code Biol Med, 2012, 7: 6
21 Quinlan A R, Hall I M. Bedtools: a flexible suite of utilities for comparing genomic features. Bioinformatics, 2010, 26: 841–842
22 Shi L, Reid L H, Jones W D, et al. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat Biotechnol, 2006, 24: 1151–1161
23 Shi L, Campbell G, Jones W D, et al. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. Nat Biotechnol, 2010, 28: 827–838