SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
TRANSCRIPTOMICS
GENE EXPRESSION PROFILING
RNA SEQUENCING

Received
12 September 2013

Accepted
13 December 2013

Published
14 January 2014

Correspondence and requests for materials should be addressed to S.S. (shankar@ucsd.edu)

Technical Variations in Low-Input RNA-seq Methodologies

Vipul Bhargava[1], Steven R. Head[2], Phillip Ordoukhanian[2], Mark Mercola[3,4] & Shankar Subramaniam[1,3,5]

[1]Bioinformatics and Systems Biology Graduate Program, University of California at San Diego, La Jolla, California, USA, [2]Next Generation Sequencing Core Facility, The Scripps Research Institute, La Jolla, California, USA, [3]Department of Bioengineering, University of California at San Diego, La Jolla, California, USA, [4]Sanford-Burnham Medical Research Institute, La Jolla, California, USA, [5]Departments of Cellular and Molecular Medicine and Chemistry and Biochemistry, University of California at San Diego, La Jolla, California, USA.

Recent advances in RNA-seq methodologies from limiting amounts of mRNA have facilitated the characterization of rare cell-types in various biological systems. So far, however, technical variations in these methods have not been adequately characterized, vis-à-vis sensitivity, starting with reduced levels of mRNA. Here, we generated sequencing libraries from limiting amounts of mRNA using three amplification-based methods, viz. Smart-seq, DP-seq and CEL-seq, and demonstrated significant technical variations in these libraries. Reduction in mRNA levels led to inefficient amplification of the majority of low to moderately expressed transcripts. Furthermore, noise in primer hybridization and/or enzyme incorporation was magnified during the amplification step resulting in significant distortions in fold changes of the transcripts. Consequently, the majority of the differentially expressed transcripts identified were either high-expressed and/or exhibited high fold changes. High technical variations ultimately masked subtle biological differences mandating the development of improved amplification-based strategies for quantitative transcriptomics from limiting amounts of mRNA.

Mammalian transcriptomes display a power-law distribution in transcript abundance with transcript expression ranging over six orders of magnitude in RNA concentrations[1,2]. RNA-seq with its large dynamic range and high sensitivity has facilitated accurate quantification of a vast majority of these transcripts[3–5]. One widely used RNA-seq protocol relies on fragmentation of mRNA into short 100–200 bp fragments which are later converted to double stranded cDNA and processed to prepare a sequencing library (Std. RNA-seq)[4]. Since there is no pre-amplification step involved this method requires at least 1–10 ng of mRNA, restricting its usefulness in applications where obtaining large amounts of mRNA is impossible such as in developmental biology, stem cell and cancer biology.

To address this issue of sequencing from limiting amounts of mRNA, a number of amplification-based methodologies[6–13] have been proposed. These methodologies generated large amounts of amplified cDNA as required for successful production of sequencing libraries, by performing either exponential or linear amplification of mRNA. In Smart-seq[8], exponential amplification of the mRNA is achieved by associating universal primer sequences to either ends of the cDNA library followed by global PCR amplification of all the transcripts using complementary sequences of the universal primers. In another instance of exponential amplification, DP-seq[11], the hybridization and extension potential of heptamer primers are utilized to amplify a majority of the transcripts. Exponential amplification based strategies generate large amounts of amplified DNA within a few hours although with high proportions of primer dimerization and/or PCR spurious products[14]. Linear amplification of the mRNA, as in the CEL-seq[6] method, requires incorporation of a T7 promoter sequence to the cDNA template followed by *in vitro* transcription (IVT) by T7 RNA polymerase that performs over 1000-fold amplification of the DNA. Owing to stringent binding of the T7 RNA polymerase to its promoter region, the IVT strategy results in reduced accumulation of spurious products. However, it requires at least 400 pg of total RNA for successful linear amplification, which is obtained by attaching unique barcodes to individual RNA samples and pooling them together before the IVT step.

Here, we assessed technical variations in the sequencing libraries prepared from limiting amounts of mRNA and their impact on data interpretation. Three amplification-based methods, Smart-seq, DP-seq and CEL-seq, were used to generate technical replicate libraries from serial dilutions of mRNA ranging from 1 ng to 25 pg. Each method involved multiple steps that were susceptible to technical variations. During the amplification step, these variations were non-linearly amplified, resulting in an increased noise in the quantification of low expressed

transcripts[8,15]. Additionally, the inefficient amplification of the majority of low to moderately expressed transcripts shifted their representation to noisy low read counts. Upon comparison with Std. RNA-seq and quantitative real time PCR (qPCR), we further noted significant distortions in the relative abundance of the transcripts, as the amount of mRNA was reduced. Consequently, differential expression analysis exclusively identified transcripts that were either highly expressed and/or exhibited high fold changes, thus masking small biological differences.

Results

Experimental design. For each amplification-based method, viz., Smart-seq, DP-seq and CEL-seq, we constructed sequencing libraries using the same mRNA source (Figure 1). The mRNA was derived from an *in vitro* cell culture based model of primitive streak (PS) induction in mouse embryonic stem cells (mESCs)[16,17]. Activation of Activin A/TGFβ pathway by high dosage of Activin A (100 ng/mL) induced mes-endoderm tissue[18–22]. Absence of Activin A, however, resulted in negligible activation of Activin A/TGFβ pathway leading to neuro-ectoderm induction[23]. Mouse ESCs were differentiated in serum-free conditions and the mRNA was collected at day 4 (equivalent to 6.5 – 7.5 days per coitum) from embroid bodies maintained in control serum free media (SFM) and those subjected to Activin A treatment (AA100). Next, serial dilutions of mRNA ranging from 50 ng – 25 pg were prepared. Std. RNA-seq libraries were prepared from 50 ng of mRNA while sequencing libraries from the amplification-based methods were prepared for rest of the dilutions (1 ng, 100 pg, 50 pg and 25 pg). For all methods, technical replicates were prepared for each dilution to access technical variations in the library preparation protocol.

Libraries obtained from Std. RNA-seq, Smart-seq and DP-seq were subjected to single-end 100 bp sequencing using the Illumina platform. Paired-end sequencing was performed for CEL-seq libraries where the first read was used to determine the barcodes of the pooled samples while the second read was mapped to the mouse transcriptome (see Supplementary Table S1 online).

Comparative transcriptomics analysis of the three amplification-based methods. For data sets obtained from each of the three library preparation methods, we randomly selected 16 million reads to perform comparative analysis. Transcriptome coverage obtained from all three amplification-based methods was high for libraries prepared from 1 ng of mRNA. However, the coverage dropped as

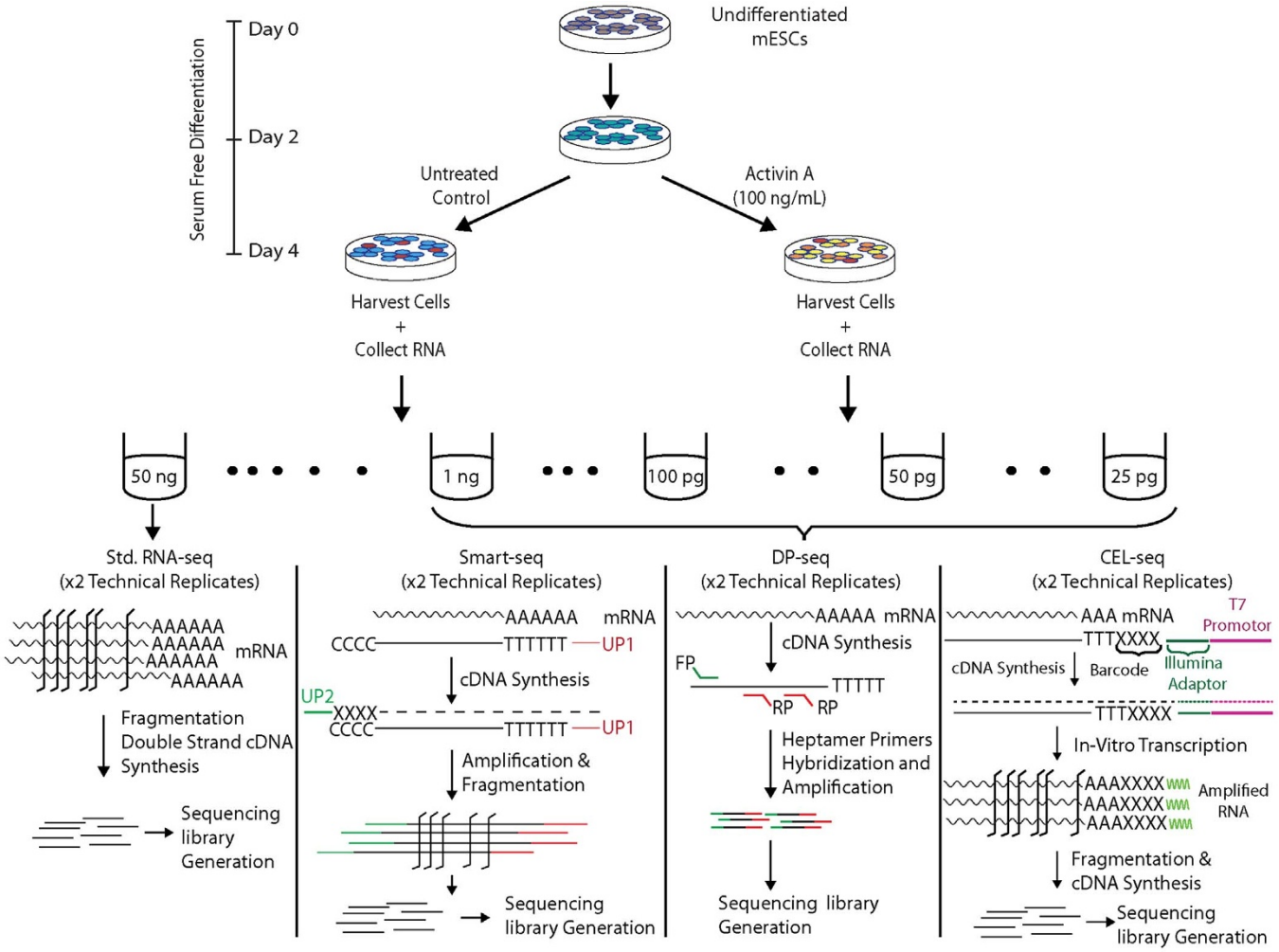

Figure 1 | Schematic representation of the experimental design. Mouse ESCs were differentiated in serum free conditions for four days. At day 2 of differentiation, embroid bodies were dispersed and Activin A was added to the culture media to stimulate Activin A/TGFβ signaling pathway. Cells were harvested at day 4 from serum free media control (SFM) and Activin A containing well (AA100) and mRNA was isolated. The mRNA was later subjected to serial dilutions ranging from 50 ng– 25 pg. Std. RNA-seq libraries were prepared from 50 ng of mRNA derived from SFM and AA100 samples. Sequencing libraries were prepared from serial dilutions (1 ng, 100 pg, 50 pg and 25 pg) of mRNA using Smart-seq, DP-seq and CEL-seq. All sequencing libraries were prepared with two technical replicates where same mRNA source was used and the library preparation steps were replicated. Salient details of all the methods are shown.

the amount of mRNA was progressively reduced (Figure 2a). Smart-seq libraries exhibited the highest transcriptome coverage at all amounts of mRNA explored. DP-seq was designed to amplify > 80% of the transcripts using 44 heptamer primers and as such it exhibited marginally less transcriptome coverage as compared to Smart-seq. CEL-seq's transcriptome coverage showed the greatest reduction in coverage as the amount of mRNA was reduced. We further determined that low expressed transcripts were the most affected with decreasing amounts of mRNA (see Supplementary Fig. S1a online).

Exponential amplification of mRNA has previously been shown to accumulate primer-dimers and PCR spurious products as the number of amplification cycles are increased[14]. Despite good coverage, mapping statistics of the libraries revealed high proportions of spurious PCR products in DP-seq libraries specifically at low amounts of mRNA (Figure 2b). On the other hand, Smart-seq libraries possessed the smallest proportions of unmapped reads. CEL-seq libraries exhibited high (~80%) mappability for all dilutions of mRNA although a slightly higher proportion of the reads mapped to intergenic/intronic locations (excluded in the NCBI RefSeq mRNA database) in comparison to the other methods.

In our previous study[11], we demonstrated the limitation of Smart-seq to efficiently amplify long transcripts (>4 Kb). DP-seq performs targeted amplification of selected regions of the transcripts; as a consequence, it did not exhibit a transcript length bias. Expectedly, the long transcripts in Smart-seq libraries exhibited lower read counts in comparison to DP-seq and Std. RNA-seq (Figure 2c). Interestingly, CEL-seq also showed low read counts for long transcripts. Next, we investigated the distribution of mapped reads across the length of the mRNA. Smart-seq and Std. RNA-seq libraries displayed overlapping distribution of the reads across the length of the transcripts (Figure 2d). DP-seq libraries showed a bias towards the 3′ end of the transcripts presumably because of the inability of reverse transcriptase to generate full-length cDNA libraries. CEL-seq libraries, on the other hand, preferentially amplified last exons of the transcripts with the vast majority of the reads mapping exclusively to the 3′ end of the transcripts.

Amplification-based methods possess a variety of PCR biases. Consequently, a subset of transcripts was preferentially amplified resulting in reduced representation of the remaining transcripts. We examined the percentage of unique reads occupied by top 100 highly expressed/amplified transcripts in the sequencing libraries prepared from all the methods. Std. RNA-seq, with no pre-amplification step, occupied only 20% of the mapped reads. CEL-seq and Smart-seq libraries showed high occupation, 51% and 39% respectively, of the top 100 amplified transcripts (see Supplementary Fig. S1b online). DP-seq used a defined set of 44 heptamer primers to amplify the majority of the expressed transcripts, and showed less PCR bias, with top 100 highly amplified transcripts occupying only 29% of the mapped reads.

We further investigated the robustness in measurements of transcript expression for all methods as a function of sequencing depth. To measure robustness, random sets of reads were selected at varying sequencing depths and the transcripts displaying similar normalized expression to the original set were determined. Std. RNA-seq libraries demonstrated robust quantification for the highest number of

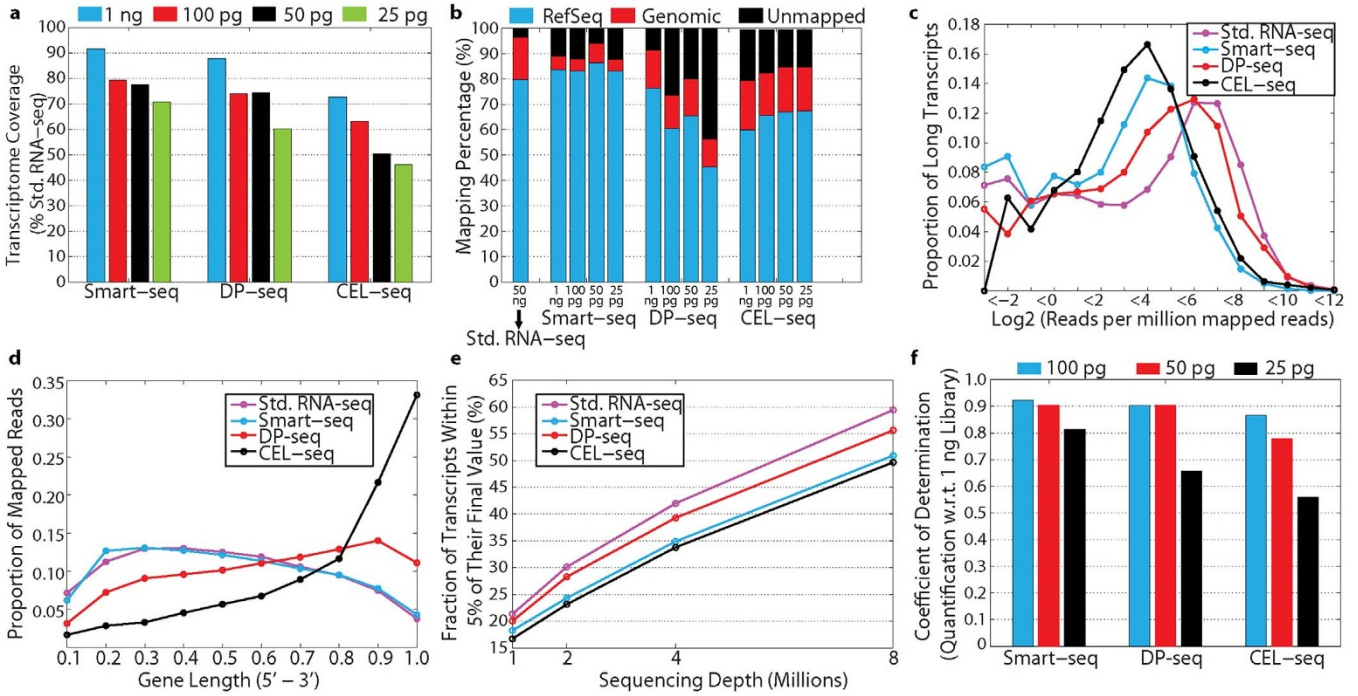

Figure 2 | Comparative transcriptomics analysis between all methods. (a) Transcriptome Coverage. Transcriptome coverage obtained by amplification-based methods was normalized to the coverage obtained in Std. RNA-seq libraries. (b) Mapping Statistics. "RefSeq" refers to proportion of reads that mapped to the NCBI RefSeq database; "Genomic" refers to reads that mapped to intergenic and intronic locations on the mouse genome; "Unmapped" refers to reads that did not map to the mouse genome. DP-seq exhibited higher proportions of primer dimerization and spurious PCR products at low amounts of mRNA. (c) Length Bias. Smart-seq failed to efficiently amplify transcripts with length > 4 Kb. (d) Distribution of mapped reads along the transcript length. Majority of the CEL-seq reads mapped to the last exon of the transcripts. (e) Robustness of unique reads measurements as a function of transcript expression levels and depth of sequencing. 16 million reads were taken from AA100 sequencing libraries to ascertain the expression of the transcripts. These reads were then successively reduced by factor of two and the expression of all the transcripts were ascertained at each depth. These measurements were then normalized to the reduction factor and the number of transcripts displaying expressions within ± 5% of the original expression was determined. (f) Coefficient of determination (R^2) was estimated in global expression measurements in sequencing libraries constructed from lower dilutions of mRNA (100 pg, 50 pg, 25 pg) with the libraries made from 1 ng of mRNA.

these transcripts followed by DP-seq (Figure 2e). These observations remained unchanged for sequencing libraries prepared from varying amounts of mRNA. Finally, global transcript measurements of libraries constructed from at least 50 pg of mRNA showed high correlation with the libraries constructed from 1 ng of mRNA for all methods (Figure 2f). However, the coefficient of determination (R^2) dropped significantly as the amount of mRNA was further reduced to 25 pg, with CEL-seq libraries showing the highest distortions in global transcript expression measurements.

We next sought to determine the read duplicates in the sequencing libraries prepared by all the methods. We followed the approach adopted by Alexa-seq[24] where 1 million uniquely mapped reads (NCBI RefSeq mRNA database) were randomly selected three times and the amount of read duplicates and the transcriptome coverage were assessed for all methods (see Supplementary Fig. S2a online). As expected, Std. RNA-seq represented the most number of unique coordinates (~74%), highlighting the small percentage of duplicates. Smart-seq consistently exhibited low proportions of duplicates in the sequencing libraries prepared from varying amounts of mRNA. This demonstrated uniform and full-length cDNA amplification of the majority of the expressed transcripts by Smart-seq. DP-seq, on the other hand, performed targeted amplification of regions of interest in the mouse transcriptome by using a defined set of 44 heptamer primers. This resulted in a high proportion of duplicates (>75%) in the DP-seq libraries. The presence of these duplicates, however, did not affect the measurements of relative abundance of the transcripts[11]. Similarly, CEL-seq targeted the last exons of the expressed transcripts for PCR amplification and as such the CEL-seq libraries also displayed high proportions of duplicates. The read duplicates increased significantly for CEL-seq as the amount of mRNA was reduced. This can be attributed to a reduction in overall transcriptome coverage and high biases observed in the CEL-seq libraries.

We also assessed the transcriptome coverage obtained from one million uniquely mapped reads for all the methods (see Supplementary Fig. S2b online). Smart-seq and DP-seq showed similar transcriptome coverage. The high proportions of duplicates in DP-seq libraries did not affect the overall transcriptome coverage. The transcriptome coverage for CEL-seq was severely affected as the amount of mRNA was reduced with majority of lowly expressed transcripts losing their representation in the sequencing libraries. Overall, the decrease in the amount of mRNA resulted in a decrease in transcriptome coverage and increase in read duplicates in the sequencing libraries prepared from the amplification-based methods.

Technical variations. Std. RNA-seq libraries prepared from 50 ng of mRNA were very highly reproducible. For amplification-based methods, the technical variations arising out of library preparation protocol increased substantially as the amount of mRNA was reduced (Figure 3a, see Supplementary Fig. S3 online). DP-seq libraries prepared from 25 pg mRNA exhibited high technical variations presumably because of accumulation of spurious PCR products (see Supplementary Fig. S4 online). CEL-seq libraries displayed the largest technical variations in the libraries prepared from 50 pg or less amounts of mRNA (see Supplementary Fig. S5 online).

The reduction in mRNA resulted in highly inefficient amplification of low expressed transcripts (RPKM < 10, in Std. RNA-seq library). Expectedly, the distributions of reads coming from these transcripts were progressively shifted towards low read counts with the majority of these transcripts losing their representation in the sequencing library (Figure 3b). We also observed a similar trend even for moderately expressed transcripts (200 > RPKM > 10, in Std. RNA-seq library) with the majority of these transcripts failing to amplify efficiently (see Supplementary Fig. S6 online).

Next, we estimated the technical variations in the replicate libraries by measuring the standard deviations in fold changes of the transcripts as a function of average read counts. Within the replicates, transcripts were not expected to be differentially regulated implying that the fold changes should be close to zero. All amplification-based methods showed characteristic profiles of variations as a function of average read counts with high variations reported for transcripts with low expression. Regardless of the method used, we noticed significant increase in technical variations in the libraries prepared from low amounts of mRNA (Figure 3d,e and f). This resulted in a poor quantification of the vast majority of moderate to low expressed transcripts including the transcription factor family of genes (Figure 3c).

Differential gene expression analysis. The biological system considered in our study was highly divergent with thousands of transcripts differentially regulated. We used the R package, DESeq[25], to identify differentially expressed genes (DEG) in order to compare the performance of the different library preparation methods at varying input mRNA amounts. In the Std. RNA-seq libraries, we identified more than 8400 differentially expressed genes. The pathway and GO term (Biological Processes) enrichments for genes up-regulated in AA100 samples contained terms specific to mesoderm/endoderm formation (see Supplementary Table S2 and S3 online). On the contrary, down-regulated genes were enriched for terms specific to ectoderm lineage. The amplification-based methods, with the exception of CEL-seq, identified large sets of DEGs in libraries prepared from 1 ng of mRNA with the majority of them shared with those identified by Std. RNA-seq. High technical variations and inefficient amplification of the transcripts resulted in drastic reduction of the transcripts identified as differentially regulated in all three methods as the amount of mRNA was reduced (Figure 4a). CEL-seq libraries consistently identified low numbers of DEGs with only 26 differentially regulated genes identified in libraries prepared from 25 pg of mRNA.

Using the expression profiles obtained from Std. RNA-seq as a control, DEGs identified in the amplification–based methods were designated as false positives if they were not represented among the differentially regulated transcripts in the Std. RNA-seq libraries. Similarly, the DEGs in Std. RNA-seq that were not represented in the amplification-based methods were designated as false negatives (see Supplementary Table S4 online). We anticipated that the transcripts expressed at low levels were prone to noise in their amplification and were likely to be over-represented among the false positives in the amplification-based methods. Indeed, the average expression of the false positives was shifted towards low expression for all amplification-based methods (see Supplementary Fig. S7 online). Moreover, the majority of the false positives exhibited a P-value distribution close to the threshold (0.01) of statistical significance implying low confidence in calling them as differentially regulated (see Supplementary Fig. S8 online). A large proportion of DEGs identified in Std. RNA-seq were not identified in the amplification-based methods as the amount of mRNA was reduced. Many of these transcripts were low expressed that lost their representation in the amplified libraries. The majority of the false negatives, including the transcripts exhibiting moderate to high expression, displayed low fold changes in the Std. RNA-seq libraries (see Supplementary Fig. S9 online). Owing to high technical variations in the amplification-based methods, these transcripts were not identified as differentially regulated.

PCR biases associated with Smart-seq led to preferential amplification of transcripts with high expression and short lengths. Hence, DEGs identified in Smart-seq libraries were over-represented by these transcripts (see Supplementary Fig. S10 online). The majority of the DEGs identified in libraries prepared from high amounts of mRNA exhibited low to moderate fold changes (see Supplementary Fig. S11 online). However, as the amount of mRNA was reduced, the technical variations increased significantly and transcripts with low

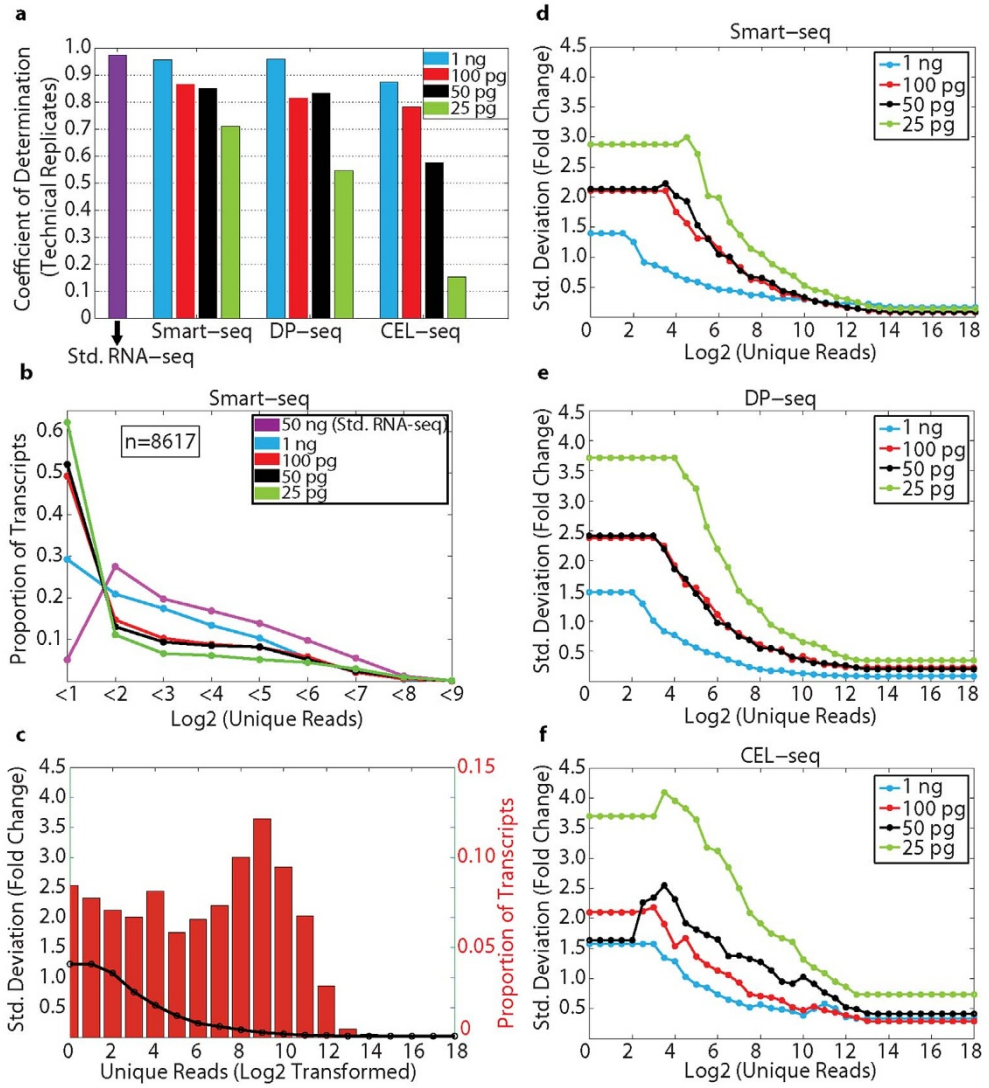

Figure 3 | Technical Variations as a function of the amount of starting material (mRNA). (a) Coefficient of determination (R^2) observed between the technical replicates in global transcriptome measurements. (b) Distribution of unique reads obtained for low expressed transcripts in Smart-seq libraries generated from different amounts of mRNA (average RPKM < 10 in Std. RNA-seq libraries prepared from control and AA100 samples). Similar distributions were observed for libraries prepared from DP-seq and CEL-seq. (c) Distribution of unique reads mapping to known mouse transcription factors (n = 1596) for AA100 sample. The black curve represents standard deviation in fold changes (log 2 transformed) observed in technical replicates of Std. RNA-seq libraries as a function of average reads. Standard deviations in fold changes (log 2 transformed) were also estimated in technical replicates as a function of average reads in libraries prepared from different amounts of mRNA using (d) Smart-seq (e) DP-seq (f) CEL-seq.

fold changes were not detected with statistical significance. We next sought to compare the fold changes of the DEGs identified for each amplification-based method to their fold changes obtained in Std. RNA-seq libraries. DP-seq demonstrated higher correlations in the fold changes in comparison to Smart-seq (Figure 4b). CEL-seq libraries showed large distortions in the fold changes. More importantly, these correlations dropped substantially for all methods as the amount of mRNA was reduced (see Supplementary Fig. S12, S13 and S14 online).

We next investigated which characteristics are necessary for a transcript to be identified as differentially regulated by the amplification-based methods as the mRNA amount is reduced. DEGs identified by the Std. RNA-seq method were classified into different categories based on their fold changes. We noted that the category consisting of high fold changes (>16 fold change) was consistently identified as the differentially expressed by all three methods. However, the identification for moderate (16 > fold change > 4) and low (fold change < 4) fold change DEGs was poor. Importantly, all three categories of DEGs suffered heavy loss as the amount of mRNA was reduced, irrespective of the method used (Figure 4c). A similar analysis was performed where DEGs identified in Std. RNA-seq were classified into different categories based on their average expression. Smart-seq identified larger proportions of highly expressed (RPKM > 200) DEGs as compared to moderate (200 > RPKM > 10) and low (RPKM < 10) expressed genes (Figure 4d). Since DP-seq distorts the relative order of gene expression, it did not discriminate based on the gene expression and identified similar proportions of DEGs for all categories of expression. CEL-seq, because of high technical noise even at high expression, failed to identify the majority of the highly expressed DEGs. We again noticed that the proportions of DEGs identified by all methods dropped significantly as the amount of mRNA was reduced.

Distortion in fold changes. Our transcriptome data showed differential regulation of the majority of the TGFβ target genes[11]. Overall, Smart-seq and DP-seq showed similar profiles for both up

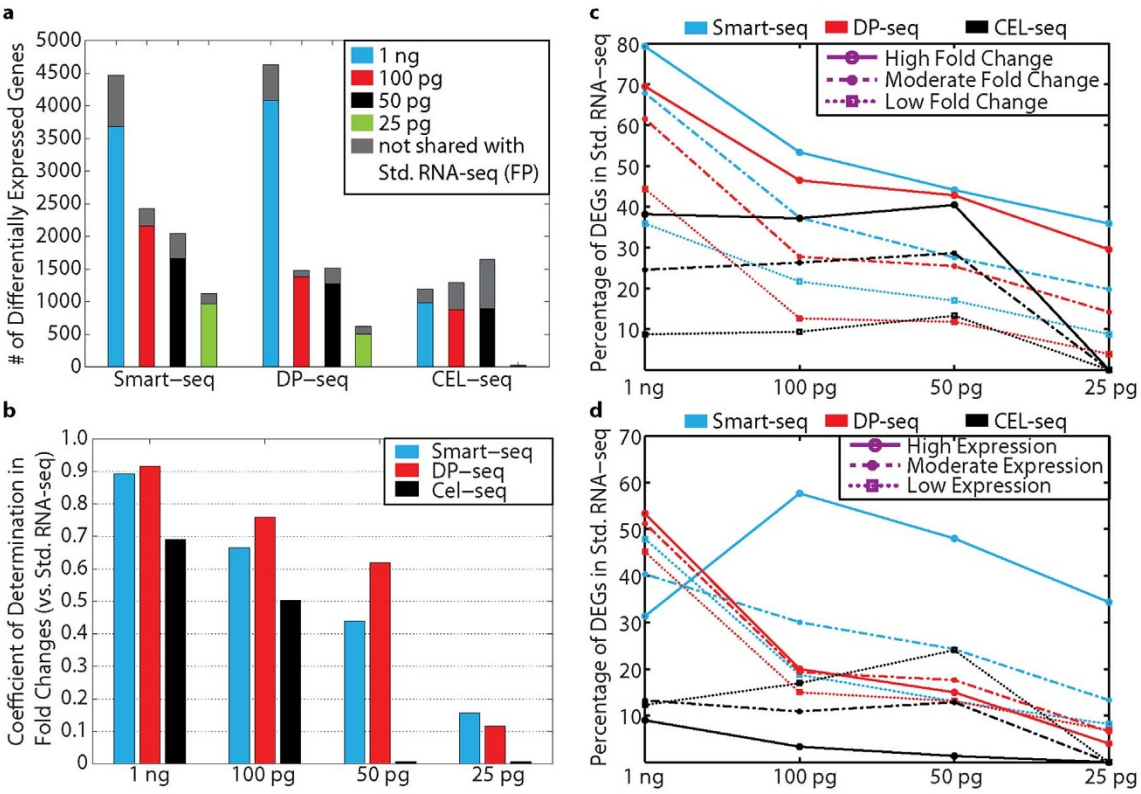

Figure 4 | Differential gene expression analysis. (a) Differentially expressed genes identified from the sequencing libraries prepared from different amounts of mRNA. FP represents false positives identified in all amplification-based methods using Std. RNA-seq as a control. (b) R^2 between the fold changes of differentially expressed genes observed between amplification-based method and Std. RNA-seq. (c) Differentially expressed genes identified from Std. RNA-seq libraries were classified into three categories of differential expression: High (fold change $>$ 4, log2 scale), Moderate (4 $>$ fold change $>$ 2, log2 scale) and Low (fold change $<$ 2, log2 scale). Proportions of these genes identified by amplification-based methods as a function of the amount of mRNA used for library preparation, are plotted. (d) Differentially expressed genes identified from Std. RNA-seq libraries were classified into three categories of transcript expression: High (RPKM $>$ 200), Moderate (200 $>$ RPKM $>$ 10) and Low (RPKM $<$ 10). Proportions of these genes identified by amplification-based methods as a function of the amount of mRNA used for library preparation, are plotted.

and down-regulated TGFβ target genes (Figure 5a). CEL-seq displayed similar trends of expression although with suppressed fold changes. The fold change distortions of the TGFβ target genes were apparent for the libraries prepared from low amounts of mRNA. To access these distortions, we compared fold changes of the transcripts in our sequencing libraries to the gold standard measurements of fold changes obtained from qPCR. For this analysis, we selected 40 transcripts, representing TGFβ target genes and known lineage markers, exhibiting moderate to low expression in the Std. RNA-seq libraries. Std. RNA-seq method conserved the relative abundance of these transcripts (R^2 = 0.91). Smart-seq libraries displayed considerably lower R^2 as the amount of mRNA was reduced. Interestingly, DP-seq showed strong correlations in the fold changes of the transcripts for all amounts of mRNA used (Figure 5b). Fold changes obtained from CEL-seq libraries showed poor correlation with the qPCR fold changes.

Out of the 181 Activin A/TGFβ pathway associated genes, 74 genes were identified as differentially regulated in differentiating mESCs treated with a high dosage of Activin A in Std. RNA-seq libraries. Regardless of the method used, the number of identified DEGs associated with the Activin A/TGFβ pathway reduced significantly as the amount of mRNA was reduced (Figure 5c). This underscored the observation that increased technical variations in low-input sequencing libraries affect biological interpretation of the datasets.

Discussion

Current sequencing technologies require nanogram quantities of RNA before being processed and made compatible for high-throughput sequencing. This motivated the development of amplification-based strategies to generate libraries for whole transcriptome profiling from low amounts of mRNA. The transcriptomics data obtained from these strategies have shown expression of thousands of transcripts even at single cell resolution, albeit with considerable noise. Notably, previous studies have implicated biological variations as the dominant source of noise in sequencing libraries prepared from either large[26–28] or ultra-low amounts of mRNA[8,15,29]. However, a comprehensive characterization of the origin of the noise observed in libraries prepared from limiting amounts of mRNA, especially from technical variations arising out of the library preparation protocols, was not performed. Here, we generated sequencing libraries from limiting amounts of mRNA using three amplification-based methods. Two of these methods, Smart-seq[8] and CEL-seq[6] have previously been used to generate libraries from mRNA derived from a single cell. Surprisingly, the libraries prepared from these methods demonstrated overwhelming technical variations as the mRNA was reduced to 25 pg (equivalent to RNA derived from tens of mammalian cell) More importantly, these technical variations were large enough to confound the biological interpretation of the datasets, thus undermining the applications of these methods at ultra-low inputs of mRNA.

To access technical variations intrinsic to these methods, technical replicates were generated with serial dilutions of mRNA. Our biological system, comprising of different germ-line lineages, exhibited diverse transcriptional changes thereby facilitating a detailed analysis of the impact of technical variations on fold change estimations and biological interpretation of the datasets. For a transcriptome wide

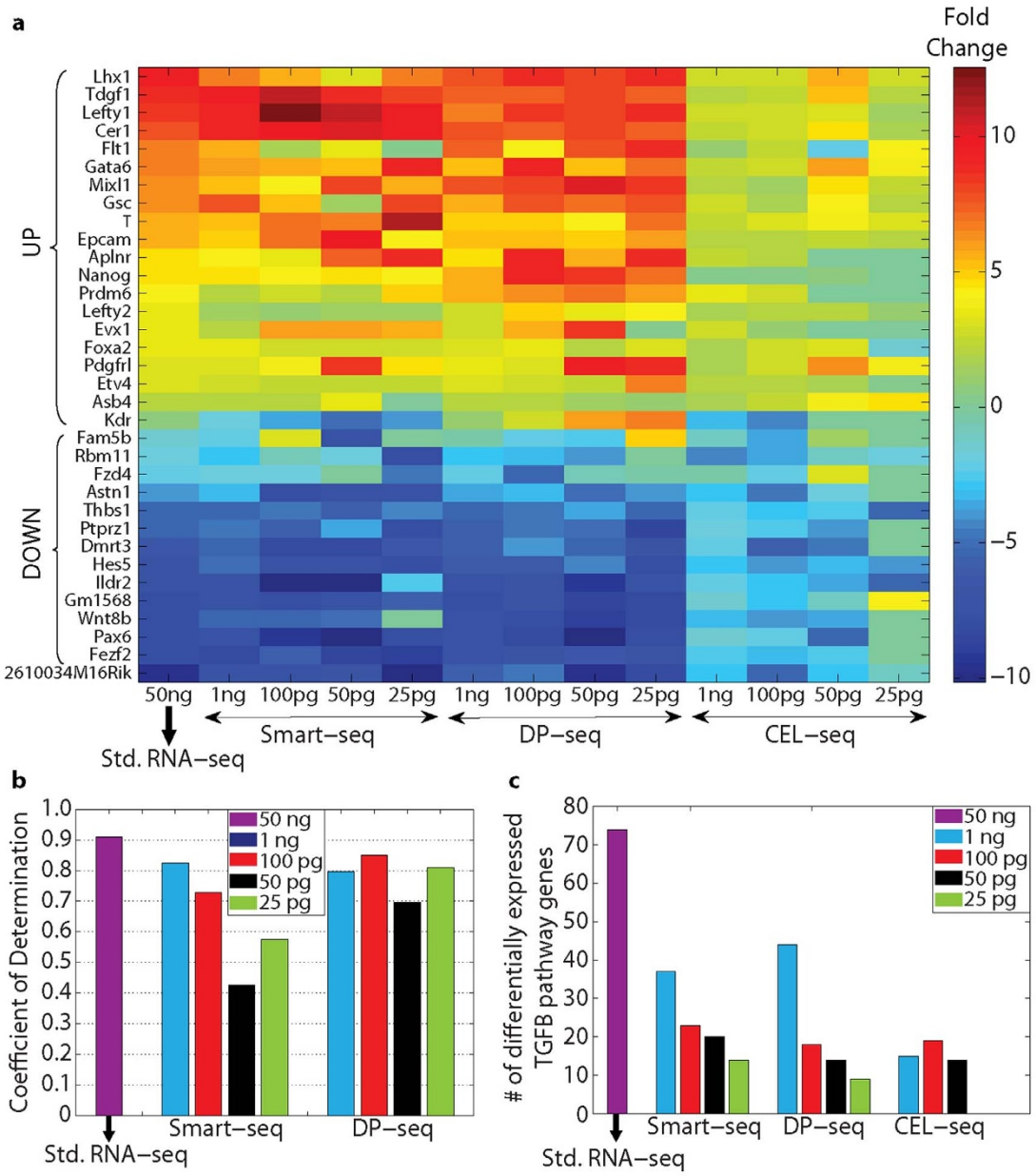

Figure 5 | Expression of Activin A/TGFβ pathway target genes in day 4 mouse embryoid bodies. (a) Heatmap displaying up/down regulation of Activin A/TGFβ pathway target genes upon introduction of Activin A in the culture media in comparison to control. All fold changes were reported in log 2 scale. (b) R^2 between the fold changes observed in the sequencing libraries and quantitative real time PCR fold changes for 40 transcripts that included TGFβ target genes and lineage markers. CEL-seq libraries exhibited poor correlations with R^2 staying close to zero. (c) Number of Activin A/TGFβ pathway associated genes identified as differentially regulated in all methods at varying amounts of mRNA.

analysis of the fold changes reported by these methods, Std. RNA-seq libraries were generated as a standard. We further estimated relative abundance of 40 transcripts, representing TGFβ target genes and lineage makers, by performing qPCR. This analysis covered the entire dynamic range of transcript expression, obviating the need for spike-in controls, e.g. ERCC libraries[30].

Smart-seq libraries exhibited high transcriptome coverage for varying amounts of mRNA and gave uniform coverage along the length of the transcripts. However, the transcript length bias in these libraries resulted in a higher representation of short transcripts in the differentially regulated transcripts. DP-seq, in comparison to Smart-seq, exhibited similar transcriptome coverage and overlapping technical noise for libraries prepared from at least 50 pg of mRNA. DP-seq exhibited less PCR biases resulting in efficient amplification and hence better quantification of more transcripts. DP-seq was also consistent in maintaining the relative abundance of the transcripts at varying amounts of mRNA. Furthermore, it was the most cost effective of the three methods in generating sequencing libraries.

At the lowest amounts of mRNA tested, DP-seq libraries showed accumulation of spurious PCR products. The 44 heptamer primers used for amplification in DP-seq were split into three tubes, which implied that only 8.33 pg of mRNA (at the lowest dilution) was amplified by each tube. A better primer design where more primers are accommodated in a single tube while ensuring high transcriptome coverage and minimizing the primer-primer interactions ($\Delta G < -4$ Kcal/mol), is expected to reduce technical noise and spurious PCR products.

In our experiments, the CEL-seq libraries performed worst in all metrics. This method exhibited the highest technical variations and fold change distortions in comparison to the other methods. Even

though the CEL-seq libraries showed expression of thousands of transcripts, the transcriptome coverage was considerably low. CEL-seq requires at least 400 pg of total RNA for successful IVT reaction[6]. In our library preparation, we satisfied this criterion by associating different barcodes to cDNA libraries prepared from the same dilution of mRNA and pooling them for the IVT reaction. We suspect that the incorporation of T7 RNA polymerase to its promoter region is subjected to high noise that is exacerbated during the final amplification step. Cost-wise, CEL-seq required paired-end sequencing where the first reads were used only for barcode identification. Moreover, CEL-seq required more steps to construct sequencing libraries and a considerable amount of time was spent handling less stable RNA.

Regardless of the method used, increased technical variations in low-input sequencing libraries prevented accurate quantification of the majority of the low to moderately expressed transcripts. As a consequence, subtle biological differences between the different cellular states, represented by the presence and absence of Activin A treatment, were lost as the amount of mRNA was reduced. We expect biological interpretation of the transcriptome data to suffer further as the amounts of mRNA are reduced to single cell levels and biological variations[31–33] are incorporated.

Sequencing library generation from few cells requires a number of RNA processing and enzymatic steps that are susceptible to technical noise. The majority of these steps are followed by either bead (solid phase reversible immobilization method) or column purification that results in loss of the starting material. Quartz-seq[9] has shown the potential to generate robust sequencing libraries from low amounts of mRNA by eliminating spurious PCR products and reducing the loss of material by performing multiple enzymatic reactions in the same reaction tube. Another potential source of variation comes from the inability of DNA polymerases to efficiently amplify lowly expressed transcripts. While optimizing DP-seq, we assessed the ability of different polymerases to amplify these transcripts and noticed that polymerases with low dissociation constant for DNA (Deep Vent R and Vent R DNA polymerase)[34] were able to efficiently amplify these transcripts. However, these polymerases also showed high proportions of primer dimerization and spurious PCR products in our sequencing libraries. Based on our experiments, we deduced that a combination of Klenow and Taq polymerase were best suited for our methodology. Smart-seq[8] uses a variant of Taq polymerase (Titanium Taq DNA polymerase) for long PCR amplifications. QUARTZ-seq, employs a mutant of Taq polymerase (MightAmp DNA polymerase) to reproducibly amplify the transcripts in the presence of PCR inhibitors. CEL-seq[6], on the other hand, uses T7 RNA polymerase to perform linear amplification of the transcripts flanked by the T7 promoter sequence. The high K_m for DNA of the T7 RNA polymerase[35] ($\sim$10 nM) can partly explain the loss of lowly expressed transcripts in the CEL-seq libraries prepared from limiting amounts of mRNA. Finally, improvements in designing new enzymes that operate at low temperatures with a high fidelity, reducing the volume of the reactions and minimizing the loss of mRNA, will substantially reduce the technical variations in the low-input libraries.

Methods

Mouse embryonic stem cell culture and differentiation. Mouse R1 embryonic stem cells were cultured on mouse embryonic fibroblast (MEF) on gelatin-coated dishes in high glucose DMEM (Hyclone, Logan, UT) supplemented with 10% fetal calf serum (FCS) (Hyclone, Logan, UT), 0.1 mM b-mercaptoethanol (GIBCO), 1% non-essential amino acids (GIBCO), 2 mM L-glutamine (Sigma, St. Louis, MO), sodium pyruvate (Sigma), antibiotics (Sigma), and 1,000 U/ml of LIF (Sigma) and passaged with 0.25% Trypsin (GIBCO).

For embryoid body (EB) differentiation, MEF were stripped from the cultures by 15 minutes incubations on gelatin-coated dishes. mESCs were collected and washed in PBS to remove traces of serum. mESCs were differentiated in serum free media containing N2 and B27 supplements as described elsewhere[16,17]. mESCs were aggregated at 50,000 cells/ml in non-coated polystyrene plates. After 2 days, EBs were dissociated by trypsin treatment and re-aggregated in fresh media in presence of Activin A at a dosage of 100 ng/mL. Activin A was obtained from R&D. EBs were harvested at day 4 for RNA extraction and processing.

mRNA purification and dilution series. Total RNA was extracted from harvested cells using Trizol (Invitrogen). Total RNA was later subjected to Oligo(dT) selection using Dynabeads mRNA Purification Kit (Invitrogen) according to the manufacturer's protocol. The enriched mRNA was later quantified using Nanodrop 2000 and serial dilutions were made ranging from 50 ng – 25 pg of mRNA.

To assess the quality of the mRNA obtained from day four EBs maintained in the serum free media control and in Activin A dosage of 100 ng/mL, the mRNA samples were analyzed on the Agilent 2100 Bioanalyzer using the Eukaryote total RNA pico chip. The traces showed characteristic size distribution of the mRNA (see Supplementary Fig. S15 online). Since the mRNA was depleted of the ribosomal RNA (18s and 28s), we could not determine the RNA integrity number[36] for these samples.

Library generation using Std. RNA-seq protocol. Std. RNA-seq[4] libraries were constructed in replicates from about 50 ng of mRNA derived from serum free media control and Activin A (100 ng/mL) samples using Illumina's TruSeq RNA Sample Prep Kit v2.

Library generation using Smart-seq. Smart-Seq cDNA library generation and amplification was performed on mRNA dilutions (1 ng, 100 pg, 50 pg and 25 pg) derived from serum free media control and Activin A (100 ng/mL) using SMARTer Ultra Low RNA Kit for Illumina sequencing (Clontech). For each mRNA dilution, libraries were generated in replicates. Following PCR cycles were used for the cDNA amplification:

1 ng – 12 cycles
100 pg – 14 cycles
50 pg – 14 cycles
25 pg – 15 cycles

These libraries were later sheared using Covaris system to obtain 200–500 bp fragments. Later, the standard Illumina library preparation protocol was followed to prepare the sequencing libraries using Illumina Paired-End DNA Sample Prep kit.

Library generation using CEL-seq. CEL-seq libraries were constructed using the protocol described earlier[6]. We used CEL-seq primers # 37, 38, 39 and 40 to generate double stranded cDNA libraries from same dilution of mRNA (including the technical replicates). The libraries were later pooled together for an *in vitro* transcription reaction. For instance in the case of lowest dilution of 25 pg of mRNA, the cDNA prepared from 100 pg of mRNA (25 pg $\times$ 2 biological samples $\times$ 2 technical replicates) was pooled and subjected to IVT reaction. This ensured that we met the minimum requirement of 400 pg of total RNA for a successful IVT reaction. Similar strategy was implemented for all mRNA dilutions. The PCR cycles used for final amplification are as follows:

1 ng – 13 cycles
100 pg – 15 cycles
50 pg – 15 cycles
25 pg – 16 cycles

To minimize the loss of the material, we performed all of the purification steps involved in the protocol with Agencourt RNAClean XP purification system according to the kit's instructions. The bead to cDNA ratio was kept at 1.5 times the reaction volume to get rid off unutilized CEL-seq primers. The sequencing libraries prepared by CEL-seq were run on the Agilent Bioanalyzer using high sensitivity DNA chip to assess the size distribution and the quality of the sequencing library.

Library generation using DP-seq. mRNA dilutions (1 ng, 100 pg, 50 pg and 25 pg) prepared from serum free media control and Activin A (100 ng/mL) were subjected to DP-seq library preparation as described[11]. The DP-seq libraries were also prepared in replicates for all mRNA dilutions. The first stand cDNA synthesis was performed using oligo dT primers (20 bp) and QuantiTect reverse transcription kit (Qiagen) according to the manufacturer's protocol. Later, the purified cDNA was split into three reaction tubes to perform amplification using our defined set of 44 heptamer primers. The PCR cycles were increased for lower dilutions of mRNA to obtain appropriate amounts of DNA for the library construction. The numbers of PCR cycles used are as follows:

1 ng – 14 cycles
100 pg – 17 cycles
50 pg – 17 cycles
25 pg – 18 cycles

The amplicon libraries thus constructed, were phosphorylated and ligated with Illumina's Y-adaptors and amplified using adaptor specific primers consisting of a different Illumina's Truseq barcode sequence for each library. The amplified libraries were run through the 2% agarose gel and size selected (150 – 500 bp) for sequencing. Custom sequencing primer was used for DP-seq sequencing libraries in the Illumina's HiSeq 2000 instrument: 5′ - ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT CCG AAT A – 3′.

Quantification of the sequencing library. Quantitative real time PCR was used to determine the concentration of the sequencing libraries prepared from DP-seq method. The standard curve for various dilutions of phiX control library was generated using the adapter specific primers recommended by Illumina. We later used the standard curve to determine the molarity of our sequencing libraries. Libraries prepared from Std. RNA-seq, Smart-seq and CEL-seq were quantified using Qubit Flurometer (Invitrogen) according to the manufacturer's protocol.

The concentration of sequencing library loaded into the Illumina flow cell was calibrated by the sequencing facility. We typically obtained good cluster density with 5 pM of library concentration on HiSeq v3 kit.

Reverse transcription and quantitative RT-PCR (qPCR). Total RNA was extracted from cells using Trizol (Invitrogen) according to the manufacturer's instructions. About 10 μg of total RNA was treated for DNA removal and converted into first strand cDNA using Quantitect Reverse Transcription kit (Qiagen). SYBR Green qPCR was run on a LightCycler 480 (Roche) using the LightCycler 480 SYBR Green Master Kit (Roche). All primers were designed with a T_m of 60°C. Data was analyzed using the $\Delta\Delta C_t$ method, using GAPDH as the normalization control, which was determined as a valid reference in mouse ESC differentiation. The primer sequences can be found as Supplementary Table S5 online.

Data analysis. *Mapping reads.* All sequencing libraries were sequenced on HiSeq 2000 platform ((TruSeq SR Cluster Kit v3-cBot-HS and TruSeq SBS Kit v3-HS). The sequencing libraries obtained from Std. RNA-seq, Smart-seq and DP-seq were sequenced to obtain 100 bp single-end reads. For CEL-seq, paired-end 100 bp sequencing was performed. The read 1 of the CEL-seq libraries were used to identify the barcodes of the pooled libraries (same mRNA dilutions) and demultiplex the reads coming from different CEL-seq primers (#37–40)[6]. The reads were demultiplexed while allowing up to 2 mismatches in the CEL-seq barcodes. For all methods, the first 7 bp of the reads (including Read 2 for CEL-seq sequencing libraries) were truncated and next 32 bp sequences were aligned to the mouse NCBI RefSeq mRNA database (Version 41 mRNA RefSeq database, May 9 2010) using an in-house mapping software while allowing up to 2 mismatches. The 44 DP-seq primers were designed for the same version of the NCBI RefSeq mRNA database where about 26,566 transcripts with NM and NR ids were selected. The transcripts with XM and XR ids were removed from the database. The reads that did not map to the mRNA database were further aligned to mouse genomic locations including intronic and intergenic locations (Build 37) using Bowtie[37] while allowing ≤ 2 mismatches.

Differential gene expression analysis. Reads mapping uniquely to the known transcripts in the biological samples were used for further analysis. DEseq[25], an R package to analyze count data from high-throughput sequencing assays such as RNA-Seq and test for differential expression, was used for sequencing library normalization and identification of differentially expressed genes. To estimate dispersions, we used "pooled-CR" method with a "fit-only" sharing mode. A P-value cutoff of 0.01 was used to identify the differentially expressed transcripts.

Accession Code. Gene Expression Omnibus: GSE50856 (sequencing read data).

1. Furusawa, C. & Kaneko, K. Zipf's law in gene expression. *Phys Rev Lett* **90**, 088102 (2003).
2. Ueda, H. R. *et al.* Universality and flexibility in gene expression from bacteria to human. *Proc Natl Acad Sci U S A* **101**, 3765–3769 (2004).
3. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**, 57–63 (2009).
4. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**, 621–628 (2008).
5. Nookaew, I. *et al.* A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae. *Nucleic Acids Res* **40**, 10084–10097 (2012).
6. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep* **2**, 666–673 (2012).
7. Gertz, J. *et al.* Transposase mediated construction of RNA-seq libraries. *Genome Res* **22**, 134–141 (2012).
8. Ramskold, D. *et al.* Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat Biotechnol* **30**, 777–782 (2012).
9. Sasagawa, Y. *et al.* Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. *Genome Biol* **14**, R31 (2013).
10. Pan, X. *et al.* Two methods for full-length RNA sequencing for low quantities of cells and single cells. *Proc Natl Acad Sci U S A* **110**, 594–599 (2013).
11. Bhargava, V., Ko, P., Willems, E., Mercola, M. & Subramaniam, S. Quantitative transcriptomics using designed primer-based amplification. *Sci Rep* **3**, 1740 (2013).
12. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* **6**, 377–382 (2009).
13. Islam, S. *et al.* Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res* **21**, 1160–1167 (2011).
14. Tang, F., Lao, K. & Surani, M. A. Development and applications of single-cell transcriptome analysis. *Nat Methods* **8**, S6–11 (2011).
15. Qiu, S. *et al.* Single-neuron RNA-Seq: technical feasibility and reproducibility. *Front Genet* **3**, 124 (2012).
16. Gadue, P., Huber, T. L., Paddison, P. J. & Keller, G. M. Wnt and TGF-beta signaling are required for the induction of an in vitro model of primitive streak formation using embryonic stem cells. *Proc Natl Acad Sci U S A* **103**, 16806–16811 (2006).
17. Willems, E. & Leyns, L. Patterning of mouse embryonic stem cell-derived pan-mesoderm by Activin A/Nodal and Bmp4 signaling requires Fibroblast Growth Factor activity. *Differentiation* **76**, 745–759 (2008).
18. Armes, N. A. & Smith, J. C. The ALK-2 and ALK-4 activin receptors transduce distinct mesoderm-inducing signals during early Xenopus development but do not co-operate to establish thresholds. *Development* **124**, 3797–3804 (1997).
19. Gurdon, J. B., Harger, P., Mitchell, A. & Lemaire, P. Activin signalling and response to a morphogen gradient. *Nature* **371**, 487–492 (1994).
20. Jones, C. M., Kuehn, M. R., Hogan, B. L., Smith, J. C. & Wright, C. V. Nodal-related signals induce axial mesoderm and dorsalize mesoderm during gastrulation. *Development* **121**, 3651–3662 (1995).
21. Tam, P. P., Kanai-Azuma, M. & Kanai, Y. Early endoderm development in vertebrates: lineage differentiation and morphogenetic function. *Curr Opin Genet Dev* **13**, 393–400 (2003).
22. Sulzbacher, S., Schroeder, I. S., Truong, T. T. & Wobus, A. M. Activin A-induced differentiation of embryonic stem cells into endoderm and pancreatic progenitors-the influence of differentiation factors and culture conditions. *Stem Cell Rev* **5**, 159–173 (2009).
23. Vallier, L. *et al.* Early cell fate decisions of human embryonic stem cells and mouse epiblast stem cells are controlled by the same signalling pathways. *PLoS One* **4**, e6082 (2009).
24. Griffith, M. *et al.* Alternative expression analysis by RNA sequencing. *Nat Methods* **7**, 843–847 (2010).
25. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol* **11**, R106 (2010).
26. McIntyre, L. M. *et al.* RNA-seq: technical variability and sampling. *BMC Genomics* **12**, 293 (2011).
27. Bullard, J. H., Purdom, E., Hansen, K. D. & Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* **11**, 94 (2010).
28. Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M. & Gilad, Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* **18**, 1509–1517 (2008).
29. Shalek, A. K. *et al.* Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498**, 236–240 (2013).
30. Baker, S. C. *et al.* The External RNA Controls Consortium: a progress report. *Nat Methods* **2**, 731–734 (2005).
31. Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. Stochastic gene expression in a single cell. *Science* **297**, 1183–1186 (2002).
32. Bar-Even, A. *et al.* Noise in protein expression scales with natural protein abundance. *Nat Genet* **38**, 636–643 (2006).
33. Kim, J. K. & Marioni, J. C. Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. *Genome Biol* **14**, R7 (2013).
34. Kong, H., Kucera, R. B. & Jack, W. E. Characterization of a DNA polymerase from the hyperthermophile archaea Thermococcus litoralis. Vent DNA polymerase, steady state kinetics, thermal stability, processivity, strand displacement, and exonuclease activities. *J Biol Chem* **268**, 1965–1975 (1993).
35. Ujvari, A. & Martin, C. T. Thermodynamic and kinetic measurements of promoter binding by T7 RNA polymerase. *Biochemistry* **35**, 14574–14582 (1996).
36. Schroeder, A. *et al.* The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC Mol Biol* **7**, 3 (2006).
37. Langmead, B. Aligning short sequencing reads with Bowtie. *Curr Protoc Bioinformatics* **Chapter 11**, Unit 11 17 (2010).

Acknowledgments

This work was supported by National Institutes of Health (NIH), National Institute of Diabetes and Digestive and Kidney Diseases Grant P01-DK074868 (S.S.), National Heart, Lung and Blood Institute Grant 5 R33 HL087375-02 (S.S. and M.M.), NIH grants HL106579 (S.S.) and HL108735 (S.S.), National Institute of Allergy and Infectious Diseases Grant U19 A1063603 (S.R.H.). and NIH/National Institute of General Medical Sciences Grant GM078005-05 (S.S.). V.B. was recipient of a California Institute for Regenerative Medicine (CIRM) Graduate Fellowship (T1-00003 and TG2-01154, Interdisciplinary Stem Cell Training Program at UCSD). The authors would like to thank John Shimashita who helped them with sequencing libraries preparation, Lana Schaffer for providing bioinformatics support, Erik Willems for providing qPCR primers, Andrew Richards and Gaurav Agrawal for many useful discussions.

Author contributions

S.S. and V.B. conceived the research. V.B. carried out mouse embryonic stem cell culture and differentiation. V.B. prepared sequencing libraries using Smart-seq, DP-seq and CEL-seq. S.R.H. and P.O. assisted in preparing Std. RNA-seq and Smart-seq libraries and

performed sequencing. V.B. performed the computational analysis. V.B. wrote the first draft of the manuscript and S.S. and M.M. supervised the work and revised the manuscript.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Bhargava, V., Head, S.R., Ordoukhanian, P., Mercola, M. & Subramaniam, S. Technical Variations in Low-Input RNA-seq Methodologies. *Sci. Rep.* **4**, 3678; DOI:10.1038/srep03678 (2014).