Streamlined Low-Input Transcriptomics through EASY-RNAseq

Yiwen Zhou [1,†], Hao Xu [1,†], Haiyang Wu [2,†], Haili Yu [1], Peng Zhou [2], Xin Qiu [2], Zihan Zheng [1], Qin Chen [2], Fa Xu [2], Gang Li [3], Jianzhi Zhou [4], Gang Cheng [1], Wei He [5], Liyun Zou [6] and Ying Wan [1,]

*1 - **Biomedical Analysis Center,** Army Medical University, Chongqing, China*
*2 - **R&D Department,** TCRCure Ltd., Chongqing, China*
*3 - **Department of Cardiology,** PIDU District People's Hospital, Chengdu, China*
*4 - **Biowavelet Ltd.,** Chongqing, China*
*5 - **Department of Gynecology and Obstetrics,** First Affiliated Hospital of Army Medical University, Chongqing, China*
*6 - **Department of Immunology,** Army Medical University, Chongqing, China*

***Correspondence to Ying Wan:** wanying516@foxmail.com*
https://doi.org/10.1016/j.jmb.2019.08.002
Edited by Dylan Taatjes

Abstract

High-throughput sequencing for transcriptome profiling is an increasingly accessible and important tool for biological research. However, accurate profiling of small cell populations remains challenging due to issues with gene detection sensitivity and experimental complexity. Here we describe a streamlined RNAseq protocol (EASY RNAseq) for sensitive transcriptome assessment starting from low amount of input materials. EASY RNAseq is technically robust enough for sequencing small pools of cells, recovering information on larger amounts of genes and with a more even expression distribution pattern than other commonly used methods. Application of EASY RNAseq to single-human embryos at the 8-cell stage led to detection of 70% of currently annotated protein-coding genes. This workflow may thus serve as a useful tool for sensitive interrogation of rare cell populations.

Introduction

RNA sequencing (RNAseq) has become a potent method for transcriptome profiling, with applications that include monitoring gene expression profiles, novel transcript assembly, and investigating alternative splicing. Indeed, the newest wave of techniques has made it possible to generate substantial transcriptomic insights into single nuclei, allowing for a fuller appreciation of the true scope of cellular heterogeneity. As researchers seek to obtain increasing amounts of information from decreasing amounts of input material, there is an increasing demand for more sensitive and resource-effective sequencing methods.

One common approach to overcome the challenge of starting from low-input materials has been to perform whole transcriptome amplification (WTA) reactions to reach the DNA threshold necessary for sequencing library construction. The earliest WTA strategy was developed by performing a PCR reaction after the polyA tailing in a reverse transcription reaction [1–3], and an alternative approach has been to apply a template-switching reaction [4,5]. However, both WTA strategies are reliant on a PCR reaction, during which unavoidable biases may be introduced to the low-input material [6]. To reduce these biases, *in vitro* transcription (IVT) methods were used for single-cell RNA-Seq such as CEL-Seq and MARS-Seq. [7,8] Unfortunately, traditional IVT requires intensive laboratory work for the amplification, and is not very easily expandable without automation. These workflows were also time-intensive, leading to increased risk of potential degradation and/or contamination.

In this manuscript, we describe a rapid and efficient method to obtain sufficient DNA for sequencing without relying on traditional RNA isolation and

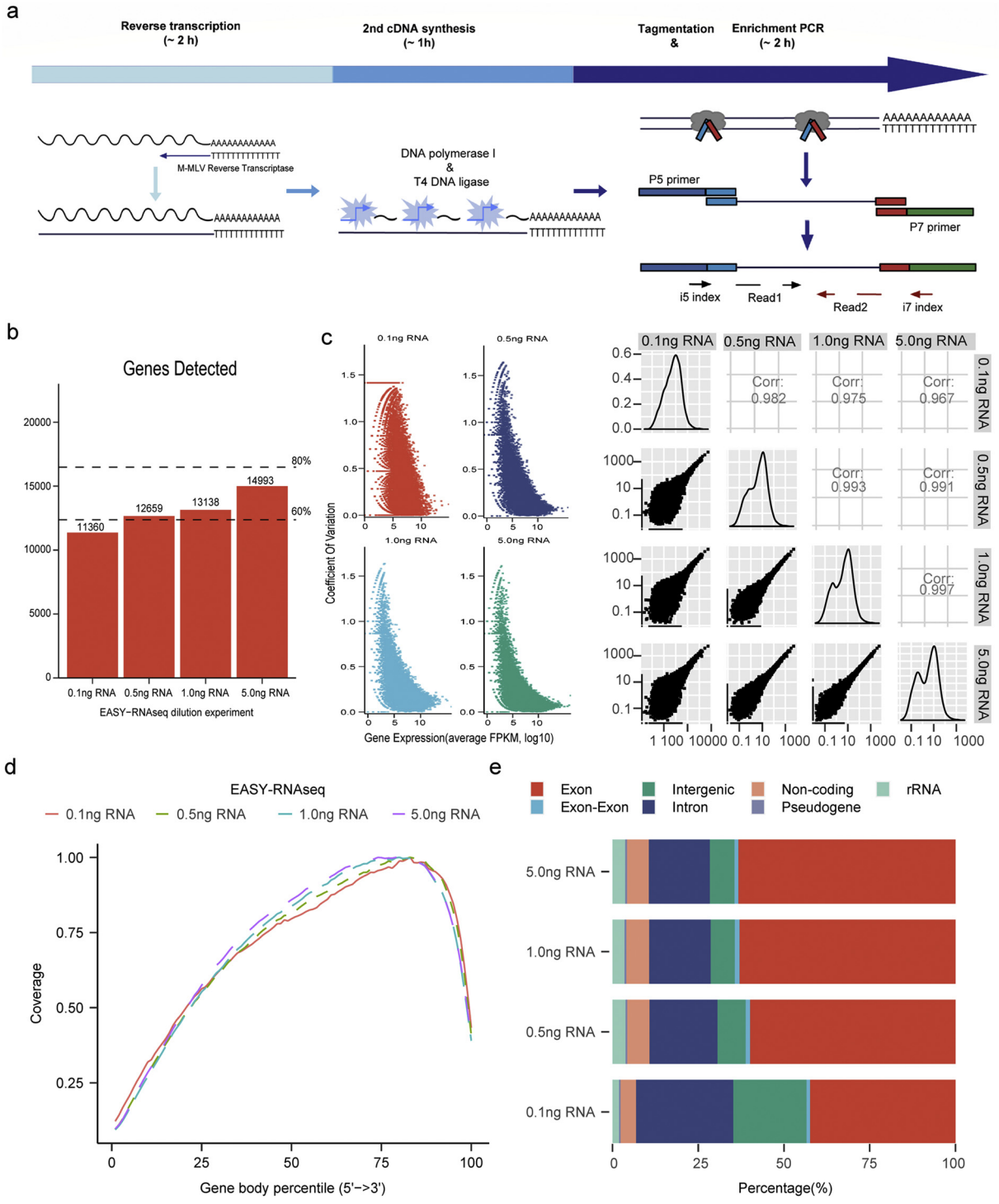


Fig. 1. EASY RNAseq is stable for diluted inputs. (A) Schematic depicting the workflow of EASY-RNAseq library construction. Serial dilution of from a single RNA sample derived from whole splenocytes could be sequenced consistently using our workflow, with over 12,000 unique genes detected from 0.1 ng of RNA, and increasing based on RNA input, reflecting the increased detection of rare transcripts (B). High sample correlation (C) and low CoV (C) demonstrate that the technique is robust, with appreciable tailing off of COV among highly expressed genes (log10 FPKM > 5). Approximately 23% gene body bias was observed that is independent of input quantity (E), and the samples had similar mapping percentages to gene features (F), suggesting that these results are characteristic.

PCR amplification. Through comparison of EASY RNAseq performed on input cell counts of varying orders of magnitude, we observed that the procedure was highly robust, and could detect significantly more protein-coding genes than in datasets generated by conventional methods, while efficiently recovering transcripts associated with smaller subpopulations of cells. EASY RNAseq is thus suited for whole transcriptome profiling of rare cell populations.

Results

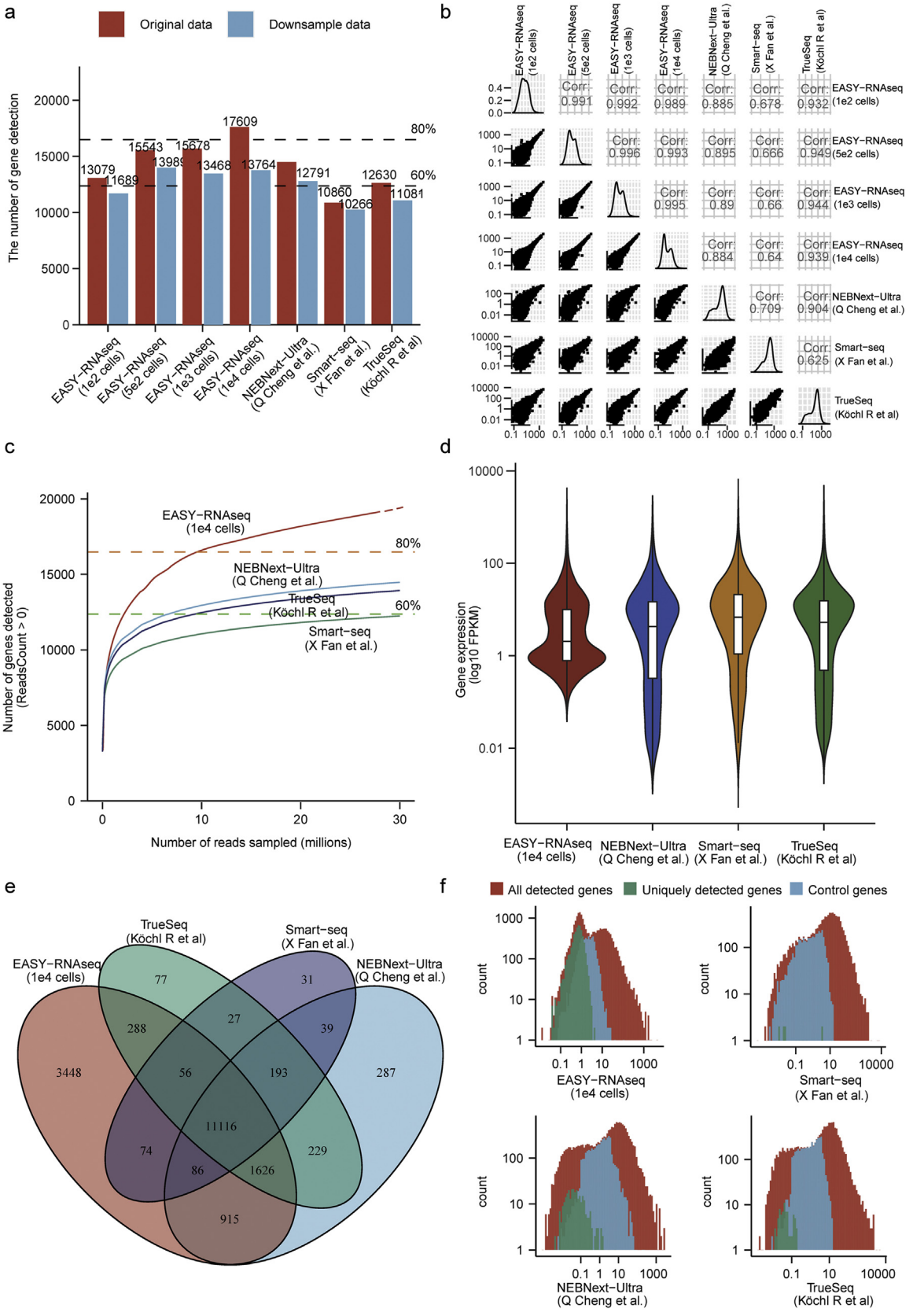
EASY RNAseq is highly reproducible at low sample input levels

Starting with total RNA, the protocol of EASY RNAseq is divided into five steps that can be completed within 5 h (Fig. 1A). As a preliminary evaluation of our workflow, we assessed the robustness of EASY RNAseq across a range of starting material levels. Four samples of 5.0, 1.0, 0.5, and 0.1 ng RNA derived from mixed murine splenocytes were generated through serial dilution to generate 11 libraries for sequencing. Fragment analysis of the libraries could identify clear RNA peaks between 200 and 400 bp in length in each (Supplementary Fig. 1C). Sequencing of these libraries yielded a total of 308 million 150-nt paired-end reads, with an average of 28 million reads per library. A preliminary analysis of the data following the pipeline described by Sahraeian *et al*.[9] suggested that our method led to a high enrichment for intergenic sequences, despite having very little apparent rRNA (Supplementary Fig. 1B). However, more detailed investigation revealed that a substantial portion of the intergenic and intronic reads mapped onto likely rRNA sequences (as per BLAST alignment) that were not obviously identified as such within the GENCODE database (Supplementary Fig. 1B–C). In order to correct for these contaminating sequences, we implemented an alignment-based search to clear putative rRNA sequences via HISTAT2. As a result of these rRNA cleaning procedure, we could more accurately recover exonic reads while reducing the percentage of intergenic and intronic sequences (Fig. 1B). A table of the pre- and post-removal read counts is included as Table S5.

From our modified analysis pipeline, over 12,000 protein-coding genes could be found in the 0.5-ng sample , representing greater than 60% of the protein-coding genes in the murine transcriptome (Fig. 1B). Genes with low expression exhibited greater variance, suggesting the possibility that the aliquoting may have led to loss of some rare transcripts as well as a potential minimum input requirement (Fig. 1C). This phenomenon was especially obvious in the 0.1-ng sample s, where a thousand fewer protein-coding genes could be found, while intergenic and intronic reads accounted for greater than 35% of all cleaned reads. These results suggest that a minimum amount of 0.5 ng input RNA is likely necessary under this procedure for successful implementation of the protocol, and that lower amounts of input may not have sufficient quantities of cDNA to tagment in our procedure. However, libraries constructed by EASY RNAseq protocol were highly similar overall, with pairwise Pearson's correlation coefficients exceeding 0.95 after averaging the expression of the three replicates, particularly in the samples with higher initial input (Fig. 1D). Notably, nearly all detected genes had a coefficient of variance less than 1 within replicates of the 5-ng sample, with only a slight increase of variance observed in those of the 1- and 0.5-ng sample s (47 and 78 genes fluctuating, respectively). From visualization of the alignment of the reads against their gene location, we could see that EASY RNAseq data could generally span across full-length transcripts, with a coverage bias favoring the 3′ end of genes (Fig. 1D). This 3′ bias is likely caused by incomplete reverse transcription when using polyT primers and was not unanticipated [10]. At the same time, a significant proportion of data still mapped to both intronic and intergenic regions. These reads may be partially explained by more complex post-transcriptional regulation mechanisms occurring in the cells that has been previously underappreciated, but which EASY RNA-seq may be able to successfully detect [11–14].

EASY RNAseq can recover weakly expressed transcripts from small inputs

While the sequencing of aliquoted RNA samples suggested that EASY RNAseq is robust enough to work with small input material, most biological questions instead involve starting from small populations of cells. As such, we next generated individual samples of sorted T cells and checked them against our mixed splenocyte samples. Consistent with our expectations, most of the genes detected between the two methods were the same (Supplementary Fig. 2A). At the same time, genes specifically expressed in T cells (such as *Cd3e*, *Cd3g*, *Cd4*, and *Cd8a*) were found at higher FPKM levels in the sorted cells than in the splenocyte mixture (Supplementary Fig. 2B), suggesting that the sorting was successful. We then compared these results with public T-cell sequencing data generated via other methods. Comparison of gene counts revealed that EASY RNAseq was able to detect over 13,000 unique protein-coding genes from just 100 input cells, exceeding the amount recoverable from the use of other protocols (Fig. 2A), while over 17,000 genes could be recovered from 10,000 cells. This effect was even more striking when the results were downsampled to an equivalent number of input


a
Original data
Downsample data
The number of gene detection
20000
15000
10000
5000
0
80%
60%
13079
11689
15543
13989
15678
13468
17609
13764
12791
10860
10266
12630
11081
EASY−RNAseq (1e2 cells)
EASY−RNAseq (5e2 cells)
EASY−RNAseq (1e3 cells)
EASY−RNAseq (1e4 cells)
NEBNext−Ultra (Q Cheng et al.)
Smart−seq (X Fan et al.)
TrueSeq (Köchl R et al)
b
Corr: 0.991
Corr: 0.992
Corr: 0.989
Corr: 0.885
Corr: 0.678
Corr: 0.932
Corr: 0.996
Corr: 0.993
Corr: 0.895
Corr: 0.666
Corr: 0.949
Corr: 0.995
Corr: 0.89
Corr: 0.66
Corr: 0.944
Corr: 0.884
Corr: 0.64
Corr: 0.939
Corr: 0.709
Corr: 0.904
Corr: 0.625
c
Number of genes detected (ReadsCount > 0)
Number of reads sampled (millions)
EASY−RNAseq (1e4 cells)
NEBNext−Ultra (Q Cheng et al.)
TrueSeq (Köchl R et al)
Smart−seq (X Fan et al.)
80%
60%
d
Gene expression (log10 FPKM)
e
TrueSeq (Köchl R et al)
Smart−seq (X Fan et al.)
EASY−RNAseq (1e4 cells)
NEBNext−Ultra (Q Cheng et al.)
3448
77
31
288
27
39
56
193
287
11116
74
229
86
1626
915
f
All detected genes
Uniquely detected genes
Control genes
count
EASY−RNAseq (1e4 cells)
Smart−seq (X Fan et al.)
NEBNext−Ultra (Q Cheng et al.)
TrueSeq (Köchl R et al)

reads. Indeed, the saturation curve of the EASY-RNAseq sample very quickly surpasses the others in gene count, but appears to approach saturation more slowly (Fig. 2B).

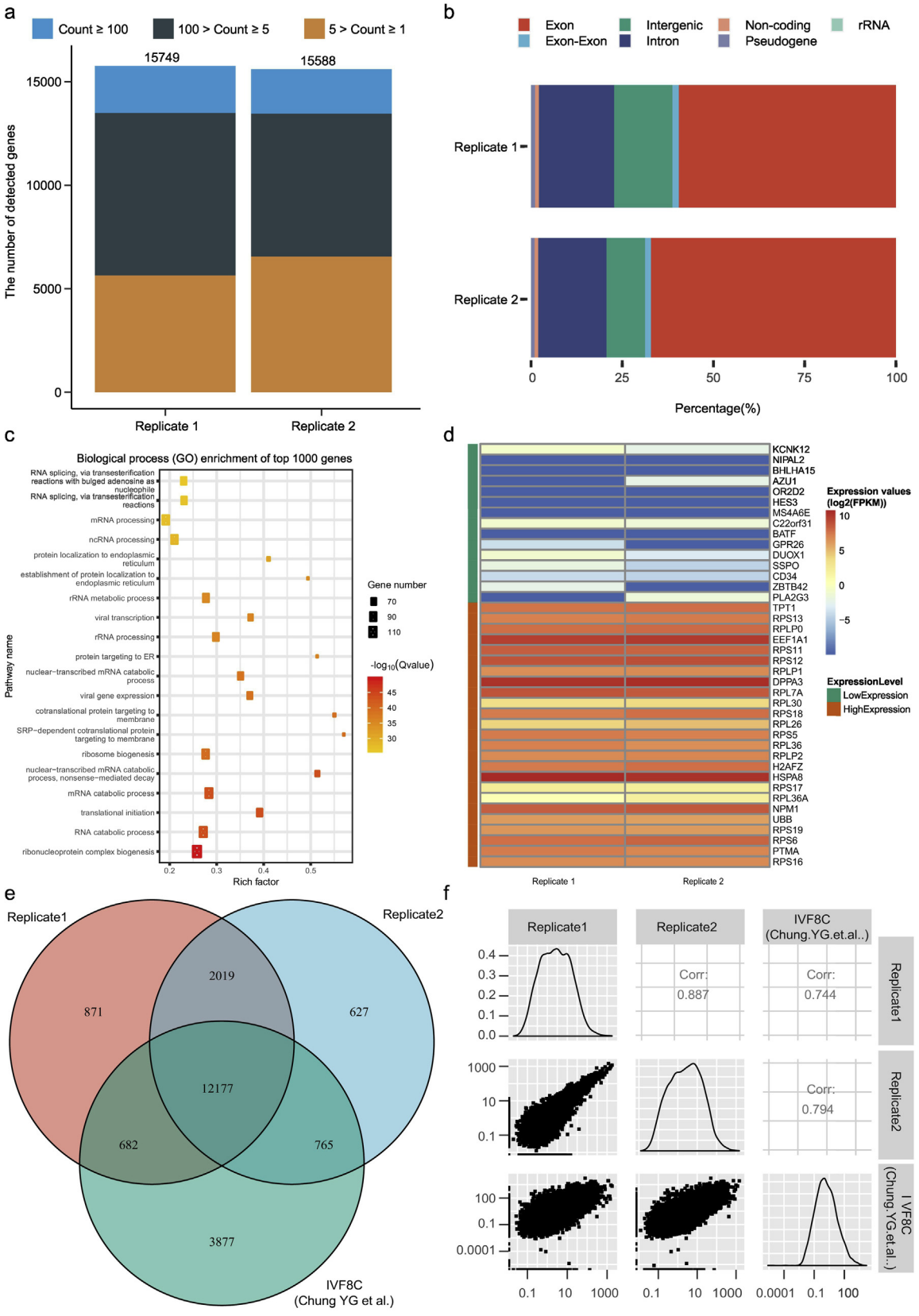
Since the quantified expression of any individual gene in RNAseq is generated through competition to be sequenced, one of the main draws/difficulties of RNAseq is its ability to evaluate relative expression of different transcripts within a single sample. Violin plots of the distribution of gene expression illustrated that EASY RNAseq profiled gene expression with a more centered density which favored genes expressed at around 1 FPKM, suggesting that EASY RNAseq could recover both more weakly expressed transcripts and also find them at higher absolute FPKM than traditional methods (Fig. 2C). At the same time, the genes that were detected had high expression correlation when compared with the other datasets, except for the SMARTseq set that sequenced regulatory T cells (Fig. 2D), indicating that our sequencing approach was still reliable for working with small fractions of sorted cells. EASY RNAseq was able to successfully quantify 4683 more protein-coding genes at any level, of which 1307 could be found at FPKM greater than 1 (Table S3, Fig. 2E). To further verify if those genes were true positives, the frequency distribution of relative expression was split into two categories: all detected genes and the uniquely detected genes. As expected, the distribution of unique genes expression was centered at low expression values, with the few genes missed in EASY RNAseq data also being ones with relatively low expression levels (and potentially attributable to random sampling bias) (Fig. 2F). At the same time, comparison of the expression levels across all four datasets of a list of 4000 genes matched in expression level to the uniquely found genes in our dataset showed that these genes also low expression in other datasets.

To further confirm that the uniquely recovered genes we found were actually expressed in mRNA form, we then selected 20 of the uniquely expressed genes for validation via qPCR. Using primers designed to span exon−exon junctions, we were able to observe that 19/20 could be found within 35 cycles, and 6/20 within 32 cycles (Fig. S2C). Furthermore, inspection of the reads mapping to the mostly highly expressed of these genes, *Klra10*, demonstrated that a significant portion of the reads mapping to the gene spanned exon−exon junctions (Fig. S2B). In order to assess the generality of this observation, we further evaluated the list of uniquely detected and commonly detected genes for potential differences in the proportion of reads crossing splices sites, exon−exon junctions, exon/intron ratio, and containment within a single gene. From our analyses, we could observe that the uniquely detected genes shared generally similar sequencing characteristics compared to the commonly recovered genes, albeit at slightly lower levels (Fig. S3C−E).

EASY RNAseq profiling of embryos

Having verified that EASY RNAseq is sufficiently robust for transcriptome analysis of low-input mouse samples, we next sought to apply the approach to even smaller counts of human cells as another likely use scenario. RNA from two pairs of 12 cells harvested from rejected IVF embryos were processed using the EASY RNAseq workflow, leading to recovery of 21.48 million reads. The proportion of the reads that mapped to reference genome were 82.86% and 90.35%, with 55% of them being exonic in the former and 43% exonic in the latter (Table S4). Interestingly, the difference of exonic mapping percentage did not seem to affect the overall gene detection sensitivity, as those exonic data detected 15,749 and 15,588 genes, respectively. Among them, 10,893 protein-coding genes were quantified with at least 5 reads in both replicates, accounting for 83% of all detected genes (Fig. 3A). Both samples had relatively high percentages of exonic reads as before (Fig. 3B). While functional analyses of the top 1000 expressed genes unsurprisingly unveiled high enrichment for genes associated with basic cellular machinery (Fig. 3C), comparison with a previously published set of genes varying during early embryonic development [15] confirmed that a range of randomly selected high and low expressed genes with our data had the same general expression pattern (Fig. 3D). Indeed, integration of our sequencing data with a previously published 8-cell stage dataset [16] revealed that 12,177 genes were found to be shared between our two samples and the previous data (Fig. 3E). Correlational analysis demonstrated that our samples had a 0.7 correlation with that set in terms of expressed genes (Fig. 3F).

Fig. 2. Comparison of EASY RNAseq with other sequencing methodologies. (A and B) Downsampling of each sequencing run to 1E7 reads and full comparisons demonstrate that EASY RNAseq is able to more evenly recover unique genes than other sequencing workflows, but still has very high correlation with the data generated by other methods in terms of expression levels. (C and D) Saturation analysis and violin plots of the expression distribution of each method reveal a notable bulge in genes with low expression that is not present in the other datasets, and is also shifted upward with a higher minimum. (E) Venn diagram of the unique genes found by each method. (F) Distribution plot of the expression levels of the uniquely detected genes within each method shows that most of the unique genes still have expression above 0.1 FPKM in EASY RNAseq, and that commonly detected genes have similar expression levels across all four workflows.


a
Count ≥ 100
100 > Count ≥ 5
5 > Count ≥ 1
15749
15588
The number of detected genes
Replicate 1
Replicate 2
b
Exon
Exon-Exon
Intergenic
Intron
Non-coding
Pseudogene
rRNA
Replicate 1
Replicate 2
Percentage(%)
c
Biological process (GO) enrichment of top 1000 genes
Pathway name
Rich factor
Gene number
-log10(Qvalue)
d
Expression values (log2(FPKM))
ExpressionLevel
LowExpression
HighExpression
e
Replicate1
Replicate2
IVF8C (Chung YG et al.)
f
Replicate1
Replicate2
IVF8C (Chung.YG.et.al..)
Corr: 0.887
Corr: 0.744
Corr: 0.794

Collectively, these results demonstrate that EASY RNAseq is suitable for detailed transcriptomic investigations of rare/small cells populations.

Discussion

The true breadth of cellular heterogeneity is becoming increasingly appreciated as new technologies for assaying single cells have been developed, with advanced platforms now capable of identifying up to 8000 genes per cell in hundreds of cells. Despite these advances however, the need for accurate transcriptome measurements of small mixtures of cells has not gone away. Due to gene dropout and other technical limitations, most approaches for differential analysis of gene expression in single-cell data require for the *in silico* clustering of the most similar cells and are critically dependent on bulk RNAseq for validation. Our recovery of over 13,000 shared genes from our application of EASY RNAseq on human embryos as the 8-cell stage represents a much higher detection rate than what is currently achievable by single-cell techniques, and could thus serve as a useful scaffold for validating single cell results. The simplification offered by EASY RNAseq may also allow it to be more amenable for automation, as most steps are reduced to simple liquid handling.

Successful transcriptome profiling of rare cell population requires overcoming the hurdle of needing to amplify low input materials while insulating against technical noise. To meet these criteria, EASY RNAseq applies an IVT-based method to replace the widely used WTA for second strand synthesis. At the same time, one potential weakness of this approach is the possibility of DNA contamination, with genomic DNA also being tagmented and processed along with the rest of the RNA and leading to high amounts of intergenic and intronic reads. Libraries generated using this method also tended to have slightly higher rates of duplicate reads. This form of noise may be particularly impactful in genes with low expression levels, as the gDNA captured would likely be near-randomly distributed across the genome. Indeed, this phenomenon could be seen in the initial 0.1-ng sample s we analyzed, while being less evident at higher levels of starting input. However, based on our qPCR validation and detailed analysis of read characteristics, it seems clear that the bulk of the low-expressed genes we detected in samples originating from higher inputs corresponded to real mRNA, even those that were not detected in other similar datasets. Collectively, our results suggested that genomic DNA contamination is not an overriding factor that would prevent extraction of biologically meaningful information in our method.

Other factors, such as tagmentation bias, may also have some influence on the efficacy of our protocol, and future refinement of the workflow may require further optimization [17–19]. Interference caused by RNA secondary structures may also influence our assessment of short sequences with complex structures, such as iron response elements or small RNAs [20,21]. It is possible that the method described here may also be further refined in the future to incorporate in mRNA enrichment or RNA fragmentation steps at the early stages to shrink the gDNA noise currently observed. However, these additions may also come at the expense of total RNA content and workflow simplicity, and may not be entirely feasible at present when starting for minute amounts of input material. More detailed investigations will be necessary to clarify this potential. Further explorations in this direction may also be able to clarify the number of reads required for meaningful biological interpretation of the results [22].

With only a dozen cells, EASY RNAseq is capable of detecting more than 15,000 of protein-coding genes. Among these detected genes, nearly 80% entries have strong signals supported by more than 5 independent non-duplicate reads. This level of sensitivity can greatly enhance both sensitivity and accuracy of distinguishing subgroups from rare cells, and aid in feature selection of cell subsets for development of novel biomarkers. EASY RNAseq may also allow for easier application of deconvolution techniques for characterizing subpopulations of cells within pooled populations as a result of the advantages offered by the more linear distribution pattern of transcript expression compared to other existing techniques. The ability of EASY RNAseq to successfully sequence starting from low amounts of heterogeneous input suggests that it may also be capable of sequencing RNA from rare clinical specimens. Future explorations and refinements are still necessary to clarify the full potential of this approach.

Fig. 3. EASY RNAseq can efficiently sequence 8-cell embryo samples. (A and B) A majority of genes in both samples were detected at robust expression levels, matching the distribution expected based on our previous sequencing, and with similar mapping ratios to exonic sequences. (C and D) Functional analysis and gene list comparison confirm that our sequencing data encompasses expected biological activities and factors previously identified. (E and F) Comparison of our samples with a previously published report on 8-cell stage embryos shows that our approach is able to identify a similar number of common genes, as well as a significant amount of unique factors. Spearman's ranked correlation of the samples confirms that the genes are expressed in similar patterns in both.

Materials and Methods

Sample preparation

All T-cell samples were obtained from the spleens of 8-week-old C57/BL6 wild-type female mice bought from Beijing Huafukang Bioscience. Samples of murine CD3 positive T cells (1e2/5e2/1e3/1e4) were sorted by flow cytometry with the BD Jazz. Use and housing of animals followed the institutional guidelines of Army Medical University as approved for W.Y. All human oocyte samples for EASY RNAseq profiling were obtained after *in vitro* fertilization and had been discarded following observation of tripronuclear (3PN) cells in the zygotes at day 1 culture. Use of the human samples was permitted by the Ethics Committee of the First Affiliated Hospital, Third Military Medical University, under Approval No. 201554 to W.H.

RNA acquisition

All cell samples were stored in TRIzol reagent (Invitrogen, cat. no. 15596026) at −80 °C prior to RNA isolation. Bulk quantities of murine RNA were isolated by classic liquid phase separation methods [23], and then diluted to 10 ng/μl for each sample. The bulk sample was then further divided into different starting material amounts (0.1/0.5/1.0/5.0 ng) by serial dilution. For low-input samples, RNA was isolated by using the Direct-zol™ RNA MiniPrep Kit (ZYMO, cat. R2050) according to the manufacturer's instructions. RNA concentration and the absorbance ratios at 260/280 and 260/230 nm of each sample were measured by NanoDrop™One (Thermo, cat. ECS000493).

EASY-RNAseq Library preparation

The qualified RNA samples were added with 1 μl of SuperScript III reverse transcriptase (200 units/μl; Invitrogen, cat. 18080044), 4 μl of Superscript III first strand synthesis buffer (5 ×; Invitrogen, cat. 18080044), 1 μl of dNTP mixture (10 mM), 1 μl of Oligo 30 (dT) primer (10 μM; polyT primer), 1 μl of RNase inhibitor (40 units/μl; Thermo, cat. k1622), and 1 μl of DTT (0.1 M) to construct 20-μl reaction system for RNA reverse transcription by heating the mixtures for 50 °C for 90 min, 70 °C for 15 min, and a final hold at 4 °C. For second-strand cDNA synthesis, the 20-μl products of reverse transcription were directly mixed into reaction buffer (NEB, cat. E6111S/L) on ice to obtain dsDNA samples with the help of DNA polymerase I and T4 DNA ligase, following the manufacturer's recommendations [24]. dsDNA products were purified with 144 μl (1.8 ×) of Agencourt AMPure XP Beads (Beckman, cat. A63880−A63882) through magnetic separation (Invitrogen, cat. no. 123.21D) [24]. The purified dsDNA samples were then diluted to 1 ng/μl following Qubit 2.0 measurement (Invitrogen, cat. Q32866), and 1 ng of dsDNA was used for Tn5 tagmentation. Reaction was performed for 10 min at 55 °C and quenched with 5 μl of pre-mixed 5 × TS to avoid excessive DNA fragmentation (tagmentation protocol). Samples were then barcoded following the protocol of TruePrep DNA Library Prep Kit with P5/P7 adapter primers (Vazyme, cat. TD503), and amplicon was purified via the VAHTS™ DNA clean beads kit (Vazyme, cat. N411-02). For the initial serial dilution experiment described in Fig. 1, 12 libraries were constructed from the sample original splenocyte mixture, with three technical replicates at each of the input levels. One of the libraries, originating from the 0.1-ng sample , showed very poor quality under fragment analysis and was consequently omitted for sequencing. Biological triplicates were generated for the purified T-cell experiments shown in Fig. 2.

Analysis of RNA-Seq data

All EASY-RNAseq libraries were sequenced through Illumina HiSeq platform with 150-bp pair-end model. Three published datasets were used for method comparison (GEO Accession: GSE63961, GSE121482, GSE111066). All sequencing reads were passed through adapter filtration by using Trimmomatic [25]. To clear potential rRNA contaminants, we then collected all rRNA sequences associated with the organism of interest through NCBI Nucleotide. For the mice samples we analyzed, 53 sequences found in Nucleotide were downloaded and used as a reference file for alignment-based mapping through HISAT2 under default parameters. Unmapped reads were then recovered and used as the mRNA information. The clean data were then aligned to respective genome reference (GRCm38 and GCRh38, respectively) [26]. Mapped reads uniquely assigned to one genomic location and one gene were counted as real gene expression, which was carried out by FeatureCounts [27]. Normalization of gene expression was performed by transferring the read counts to FPKM values, and both of the read counts and FPKM values were used in visualization. The coverage of gene body and percentage of data features were calculated by RseQC [28]. For more direct comparison of the datasets via downsampling, raw reads were sampled at random from the corresponding source analyzed by same protocol afterward.

The reproducibility was assessed within samples and replicates, respectively. Pearson's coefficient was used to measure the similarity between samples, which was calculated with average FPKM values. Similarly, the coefficient of variance was applied to represent the technical robustness within replicates. The saturation curves were created by

the regression of downsampled datasets. Each curve was based on 25 different volumes of downsampled data, ranging from 0.1 million to the corresponding original data size. The saturation plot was cutoff that only showed the maximum sequencing depth of 30 million. The Venn plot was draw with those quantified (reads count > 0) protein-coding genes of each original dataset. The uniquely quantified genes of each dataset were selected, together with all quantified protein-coding genes, and were visualized as expression frequency spectrums. Gene set enrichment analysis was performed using the GSEA app (Broad Institute) on the KEGG gene lists.

Transcriptome profiling of embryonic cells

The expression values of two embryonic samples were obtained by the same RNA-Seq analysis workflow mentioned above. The expression values were defined into three categories for visualization. Among those quantified genes, genes with reads count greater than or equal to 5 were treated as confidently detected. To decrease the false positives, only those confidently detected genes were used in the following analysis. The Gene Set Enrichment Analysis (GSEA) was performed with the intersection of expression results through clusterProfiler packages [29,30].

Visualization

All figures were produced by R [31] and ggplot2 [32]. The color board was using ggsci [33]. Among them, those plots of Pearson's coefficient in Supplementary Fig. 2B and Fig. 3B were created by GGally [34], and the Venn plot was produced through VennDiagram [35]. Heatmaps were visualized through pheatmap [36].

Data Availability

Raw sequencing data generated in this paper are accessible on the Sequence Read Archive (SRA) under project accession number PRJNA542941. Computed values generated based on the raw data are available as GSE135290. All scripts used for the analysis are available upon request to the authors.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmb.2019.08.002.

Acknowledgments

This work was supported by grants from the National Key Research and Development Program of China (2017YFA0700404) to Y.W., and the National Natural Science Foundation of China (91642119) to Y.W. We would like to thank the other members of the Wan laboratory and Dr. Lin He for insightful discussions during the drafting of the manuscript.

Declaration of Competing Interest

J. Z. is affiliated with Biowavelet Ltd., which is developing automated methods for performing biological experiments.

Received 13 March 2019;
Received in revised form 7 August 2019;
Accepted 7 August 2019
Available online 3 September 2019

Keywords:
T cells;
embryo;
mRNA;
low input;
RNAseq

Co-first authors.

Abbreviations used:
RNAseq, RNA sequencing; WTA, whole transcriptome amplification; IVT, *in vitro* transcription.

†Co-first authors.

References

[1] K. Kurimoto, Y. Yabuta, Y. Ohinata, Y. Ono, K.D. Uno, R.G. Yamada, H.R. Ueda, M. Saitou, An improved single-cell cDNA amplification method for efficient high-density oligonucleotide microarray analysis, Nucleic Acids Res. (2006), https://doi.org/10.1093/nar/gkl050.

[2] K.Q. Lao, F. Tang, C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, B. Tuch, J. Bodeau, A. Siddiqui, M.A. Surani, mRNA-sequencing whole transcriptome analysis of a single cell on the solidTM system, J. Biomol. Tech. 20 (2009) 266–271.

[3] Y. Sasagawa, I. Nikaido, T. Hayashi, H. Danno, K.D. Uno, T. Imai, H.R. Ueda, Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals nongenetic gene-expression heterogeneity, Genome Biol. 14 (2013) 1–17.

[4] D. Ramsköld, S. Luo, Y.C. Wang, R. Li, Q. Deng, O.R. Faridani, G.A. Daniels, I. Khrebtukova, J.F. Loring, L.C. Laurent, G.P. Schroth, R. Sandberg, Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells, Nat. Biotechnol. 30 (2012) 777–782.

[5] S. Islam, U. Kjällquist, A. Moliner, P. Zajac, J.B. Fan, P. Lönnerberg, S. Linnarsson, Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq, Genome Res. 21 (2011) 1160–1167.

[6] S. Parekh, C. Ziegenhain, B. Vieth, W. Enard, I. Hellmann, The impact of amplification on differential expression analyses by RNA-seq, Sci. Rep. (2016), https://doi.org/10.1038/srep25533.
[7] T. Hashimshony, F. Wagner, N. Sher, I. Yanai, CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification, Cell Rep. 2 (2012) 666–673.
[8] D.A. Jaitin, E. Kenigsberg, H. Keren-Shaul, N. Elefant, F. Paul, I. Zaretsky, A. Mildner, N. Cohen, S. Jung, A. Tanay, I. Amit, Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types, Science 343 (2014) 776–779.
[9] S.M.E. Sahraeian, M. Mohiyuddin, R. Sebra, H. Tilgner, P.T. Afshar, K.F. Au, N. Bani Asadi, M.B. Gerstein, W.H. Wong, M.P. Snyder, E. Schadt, H.Y.K. Lam, Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis, Nat. Commun. (2017), https://doi.org/10.1038/s41467-017-00050-4.
[10] Z. Wang, M. Gerstein, M. Snyder, RNA-Seq: a revolutionary tool for transcriptomics, Nat. Rev. Genet. (2009), https://doi.org/10.1038/nrg2484.
[11] A. Ameur, A. Zaghlool, J. Halvardson, A. Wetterbom, U. Gyllensten, L. Cavelier, L. Feuk, Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain, Nat. Struct. Mol. Biol. 18 (2011) 1435–1440.
[12] D. Gaidatzis, L. Burger, M. Florescu, M.B. Stadler, Analysis of intronic and exonic reads in RNA-seq data characterizes transcriptional and post-transcriptional regulation, Nat. Biotechnol. 33 (2015) 722–729.
[13] S.H. Lee, I. Singh, S. Tisdale, O. Abdel-Wahab, C.S. Leslie, C. Mayr, Widespread intronic polyadenylation inactivates tumour suppressor genes in leukaemia, Nature. 561 (2018) 127–131.
[14] Y. Song, B. Milon, S. Ott, X. Zhao, L. Sadzewicz, A. Shetty, E.T. Boger, L.J. Tallon, R.J. Morell, A. Mahurkar, R. Hertzano, A comparative analysis of library prep approaches for sequencing low input translatome samples, BMC Genomics 19 (2018) 1–16.
[15] S. Petropoulos, D. Edsgärd, B. Reinius, Q. Deng, S.P. Panula, S. Codeluppi, A.P. Reyes, S. Linnarsson, R. Sandberg, F. Lanner, Single-cell RNA-Seq reveals lineage and X chromosome dynamics in human preimplantation embryos, Cell. 167 (1) (2016) 285, https://doi.org/10.1016/j.cell.2016.08.009.
[16] L. Yan, M. Yang, H. Guo, L. Yang, J. Wu, R. Li, P. Liu, Y. Lian, X. Zheng, J. Yan, J. Huang, M. Li, X. Wu, L. Wen, K. Lao, R. Li, J. Qiao, F. Tang, Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells, Nat. Struct. Mol. Biol. 20 (9) (2013 Sep) 1131–1139, https://doi.org/10.1038/nsmb.2660.
[17] B. Ason, W.S. Reznikoff, DNA sequence bias during Tn5 transposition, J. Mol. Biol. 335 (5) (2004 Jan 30) 1213–1225.
[18] M. Steiniger, C.D. Adams, J.F. Marko, W.S. Reznikoff, Defining characteristics of Tn5 transposase non-specific DNA binding, Nucleic Acids Res. 34 (9) (2006 May 22) 2820–2832.
[19] A. Kia, C. Gloeckner, T. Osothprarop, N. Gormley, E. Bomati, M. Stephenson, I. Goryshin, M.M. He, Improved genome sequencing using an engineered transposase, BMC Biotechnol. 17 (1) (2017 Jan 17) 6, https://doi.org/10.1186/s12896-016-0326-1.
[20] R.T. Fuchs, Z. Sun, F. Zhuang, G.B. Robb, Bias in ligation-based small RNA sequencing library construction is determined by adaptor and RNA structure, PLoS One 10 (5) (2015 May 5), e0126049, https://doi.org/10.1371/journal.pone.0126049.
[21] T.J. Jackson, R.V. Spriggs, N.J. Burgoyne, C. Jones, A.E. Willis, Evaluating bias-reducing protocols for RNA sequencing library preparation, BMC Genomics 15 (2014 Jul 7) 569, https://doi.org/10.1186/1471-2164-15-569.
[22] R. Lei, K. Ye, Z. Gu, X. Sun, Diminishing returns in next-generation sequencing (NGS) transcriptome data, Gene. 557 (1) (2015 Feb 15) 82–87, https://doi.org/10.1016/j.gene.2014.12.013.
[23] A.B. Hummon, S.R. Lim, M.J. Difilippantonio, T. Ried, Isolation and solubilization of proteins after TRIZOL® extraction of RNA and DNA from patient material following prolonged storage, Biotechniques. (2007), https://doi.org/10.2144/000112401.
[24] J. Cao, J.S. Packer, V. Ramani, D.A. Cusanovich, C. Huynh, R. Daza, X. Qiu, C. Lee, S.N. Furlan, F.J. Steemers, A. Adey, R.H. Waterston, C. Trapnell, J. Shendure, Comprehensive single-cell transcriptional profiling of a multicellular organism, Science. 357 (2017) 661–667.
[25] A.M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data, Bioinformatics. (2014), https://doi.org/10.1093/bioinformatics/btu170.
[26] Frankish, A., Diekhans, M., Ferreira, A.-M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J. M., Sisu, C., Wright, J., Armstrong, J., Barnes, I., Berry, A., Bignell, A., Carbonell Sala, S., Chrast, J., Cunningham, F., Di Domenico, T., Donaldson, S., Fiddes, I. T., García Girón, C., Gonzalez, J. M., Grego, T., Hardy, M., Hourlier, T., Hunt, T., Izuogu, O. G., Lagarde, J., Martin, F. J., Martínez, L., Mohanan, S., Muir, P., Navarro, F. C. P., Parker, A., Pei, B., Pozo, F., Ruffier, M., Schmitt, B. M., Stapleton, E., Suner, M.-M., Sycheva, I., Uszczynska-Ratajczak, B., Xu, J., Yates, A., Zerbino, D., Zhang, Y., Aken, B., Choudhary, J. S., Gerstein, M., Guigó, R., Hubbard, T. J. P., Kellis, M., Paten, B., Reymond, A., Tress, M. L., and Flicek, P. (2018) GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. https://doi.org/10.1093/nar/gky955.
[27] Y. Liao, G.K. Smyth, W. Shi, FeatureCounts: an efficient general purpose program for assigning sequence reads to genomic features, Bioinformatics. (2014), https://doi.org/10.1093/bioinformatics/btt656.
[28] L. Wang, S. Wang, W. Li, RSeQC: quality control of RNA-seq experiments, Bioinforma. Oxford Engl. (2012), https://doi.org/10.1093/bioinformatics/bts356.
[29] G. Yu, clusterProfiler: Universal Enrichment Tool for Functional and Comparative Study, bioRxiv (2018), https://doi.org/10.1101/256784.
[30] A. Subramanian, P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette, A. Paulovich, S.L. Pomeroy, T.R. Golub, E.S. Lander, J.P. Mesirov, Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, Proc. Natl. Acad. Sci. (2005), https://doi.org/10.1073/pnas.0506580102.
[31] Team, R. D. C., and R Development Core Team, R, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2016, https://doi.org/10.1007/978-3-540-74686-7.
[32] Ginestet, C. (2011) ggplot2: elegant graphics for data analysis. J. R. Stat. Soc. Ser. A (Statistics Soc. https://doi.org/10.1111/j.1467-985X.2010.00676_9.x.
[33] Xiao, N. (2018) ggsci: Scientific Journal and Sci-Fi Themed Color Palettes for "ggplot2".

[34] Schloerke, B., Briatte, F., bigbeardesktop, Crowley, J., justsomeone1001, Cook, D., Ibanez, E., Ross, Ogden, K., Sievert, C., Joseph, Spiller, T., Gilligan, J., elbamos, Beck, M. W., Richter, J., FabianRoger, Thoen, E., Schmidt, C., Muschelli, J., Müller, K., Bolker, B., Xie, Y., Badger, T.G., Hofmann, H., Eraslan, G., Le Pennec, E., & Chuanxin (2017). ggobi/ggally: v1.3.2. https://doi.org/10.5281/zenodo.838362

[35] H. Chen, Venn diagram: generate high-resolution Venn and Euler plots, 2018.

[36] R. Kolde, Pheatmap: Pretty Heatmaps. R Package Version 1.0.10. https://CRAN.R-project.org/package=pheatmap, 2018.