BMC Genomics

RESEARCH

Open Access

siqRNA-seq is a spike-in-independent technique for quantitative mapping of mRNA landscape

Zhenzhen Wang[1†], Kehan Tao[1†], Jiaojiao Ji[1], Changbin Sun[1*†] and Wei Xu[1*]

Abstract

Background RNA sequencing (RNA-seq) is widely used for gene expression profiling and quantification. Quantitative RNA sequencing usually requires cell counting and spike-in, which is not always applicable to many samples. Here, we present a novel quantitative RNA sequencing method independent of spike-ins or cell counting, named siqRNA-seq, which can be used to quantitatively profile gene expression by utilizing gDNA as an internal control. Single-stranded library preparation used in siqRNA-seq profiles gDNA and cDNA with equal efficiency.

Results To quantify mRNA expression levels, siqRNA-seq constructs libraries for total nucleic acid to establish a model for expression quantification. Compared to Relative Quantification RNA-seq, siqRNA-seq is technically reliable and reproducible for expression profiling but also can sequence reads from gDNA which can be used as an internal reference for accurate expression quantification. Applying siqRNA-seq to investigate the effects of actinomycin D on gene expression in HEK293T cells, we show the advantages of siqRNA-seq in accurately identifying differentially expressed genes between samples with distinct global mRNA levels. Furthermore, we analyzed factors influencing the downward trend of gene expression regulated by ActD using siqRNA-seq and found that mRNA with m^6A modification exhibited a faster decay rate compared to mRNA without m^6A modification. Additionally, applying this technique to the quantitative analysis of seven tumor cell lines revealed a high degree of diversity in total mRNA expression among tumor cell lines.

Conclusions Collectively, siqRNA-seq is a spike-in independent quantitative RNA sequencing method, which creatively uses gDNA as an internal reference to absolutely quantify gene expression. We consider that siqRNA-seq provides a convenient and versatile method to quantitatively profile the mRNA landscape in various samples.

Keywords Transcriptome, Gene expression, Quantification, RNA sequencing, Next-generation sequencing

Background

RNA sequencing (RNA-seq) and almost 100 derivatives have revolutionized our understanding of gene expression in various aspects of biology [1], such as common markers for multiple cancers [2]. Studies of single-cell omics techniques and recent spatial transcriptomics are increasingly being explored to reshape the current cell-type classification system and preserve spatial information [3–5]. Nonetheless standard bulk RNA-seq remains a convenient and routine research tool for studies on

†Zhenzhen Wang, Kehan Tao and Changbin Sun contributed equally to this work.

*Correspondence:
Changbin Sun
sunchangbin@caas.cn
Wei Xu
xuwei01@caas.cn
[1] Shenzhen Branch, Guangdong Laboratory of Lingnan Modern Agriculture, Key Laboratory of Livestock and Poultry Multi-omics of MARA, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen 518120, China

gene expression, especially for a large number of samples to be sequenced [1].

For RNA-seq analysis, gene expression is quantified to identify differentially expressed genes (DEGs), to infer regulatory networks, and/or to reveal cellular states and function [6]. Generally, sequencing reads are mapped to the reference genome after quality control and assigned to each feature to quantify gene expression [7]. To account for differences in read depth, genes are typically normalized to RPKM (Reads Per Kilobase Million) or FPKM (Fragments Per Kilobase Million) for comparative studies. These normalization methods are largely based on assumptions that most genes are at the same expression level and that the total mRNA levels do not show much difference across samples [8]. However, when groups with high heterogeneity of gene expression levels are sequenced for comparison, these assumptions fail to hold and sequencing depth-based normalization methods may result in erroneous conclusions [9], such as cells with high levels of c-Myc showing two to three times more total RNA than their low-Myc counterparts [10, 11]. To address such issues, one approach is the use of spike-in control RNAs that are added during library construction [12, 13]. Spike-in controls are a useful resource for evaluating the sensitivity and accuracy of RNA-seq experiments for transcriptome discovery and quantification [9]. Although spike-in controls are commercially available, such as a common set of external RNA controls that has been developed by the External RNA Controls Consortium (ERCC), they are not yet widely adopted [1].

Here, we developed a spike-in independent quantitative RNA sequencing (siqRNA-seq) method, which uses genomic DNA (gDNA) as an internal reference to normalize mRNA expression levels. By comparison, we showed that the relative gene expression pattern profiled by siqRNA-seq is similar to that profiled by Relative Quantification RNA-seq. We further demonstrated that siqRNA-seq enables us to assess the copy number of mRNA per cell/genome without cell counting, RNA quantification, or spike-ins. Finally, we exemplify the application of siqRNA-seq for differential gene expression analysis on samples with distinct global mRNA levels. Together, siqRNA-seq provides a convenient and versatile method to quantitatively profile the mRNA landscape.

Results

Design of siqRNA-seq

To develop a spike-in-independent method for gene expression quantification, we designed siqRNA-seq that utilized genomic DNA as an internal reference for normalization (Fig. 1A). As we know, in general, each cell during the G1 phase has a complete set of gDNA, so we considered that the gDNA can be taken as a stable internal reference for normalizing mRNA copy number to "per cell". For non-interphase cells or other special samples, the copy number of gDNA in each cell is not constant; thus, the "per genome" normalization method is suitable. With that goal, in siqRNA-seq, total nucleic acids were extracted from samples to construct two libraries in parallel, an mRNA library and an mRNA&gDNA library for each sample, as shown in Fig. 1A. The only difference between the mRNA library and the mRNA&gDNA library preparation was that DNase I digestion was performed after nucleic acid extraction to remove gDNA in mRNA library (Fig. 1A). For both libraries, mRNA was reverse-transcribed with oligo(dT) primers to synthesize complementary DNA (cDNA). To reduce bias and simplify the pipeline of library preparation, reverse transcription products were fragmented by sonication and denatured by heat, following library preparation using a highly efficient single-strand DNA (ssDNA) ligation technique, Adaptase™ from xGen™ ssDNA&Low-Input DNA Library Prep Kit (Integrated DNA Technologies). Denaturation makes cDNA-mRNA hybrid and gDNA to ssDNA. Adaptase is a commercial enzyme mixture with very high ligation efficiency and low bias for single-stranded DNA (ssDNA) library construction with strand-specific information, which we have used to develop ssDRIP-seq and ULI-ssDRIP-seq for R-loop profiling [14, 15] and DEtail-seq for DNA break detection [16]. Compared to Relative Quantification RNA-seq, the mRNA library in siqRNA-seq was directly prepared from single-stranded cDNA, therefore we named this method ssRNA-seq (single-strand DNA ligation-basedR NA sequencing) hereafter.

As shown above, we can obtain two data sets from siqRNA-seq after sequencing, ssRNA-seq for mRNA profiling and the mRNA&gDNA library with reads from gDNA and mRNA. The next step is to integrate the two data sets to quantify gene expression with reads from gDNA as an internal reference for normalization (Fig. 1B). In the mRNA&gDNA library, reads that are mapped to intergenic regions can be used to calculate the sequencing depth of gDNA, while reads that are assigned to exons include mRNA reads and gDNA reads. Consequently, the mRNA count per diploid genome (RCPG) for each gene can be obtained from the ratio of mRNA read depth to gDNA read depth multiplied by four (gDNA of diploid with two strands). If cells are at interphase with constant diploid genomic DNA, we can infer that the mRNA count per cell (RCPC) is equal to RCPG (Fig. 1B). Considering that it is difficult to distinguish cDNA reads on low-expression genes from the gDNA background, we use ssRNA-seq data to calculate FPKM (Fragments Per Kilobase Million) for all genes and combine FPKM and RCPG

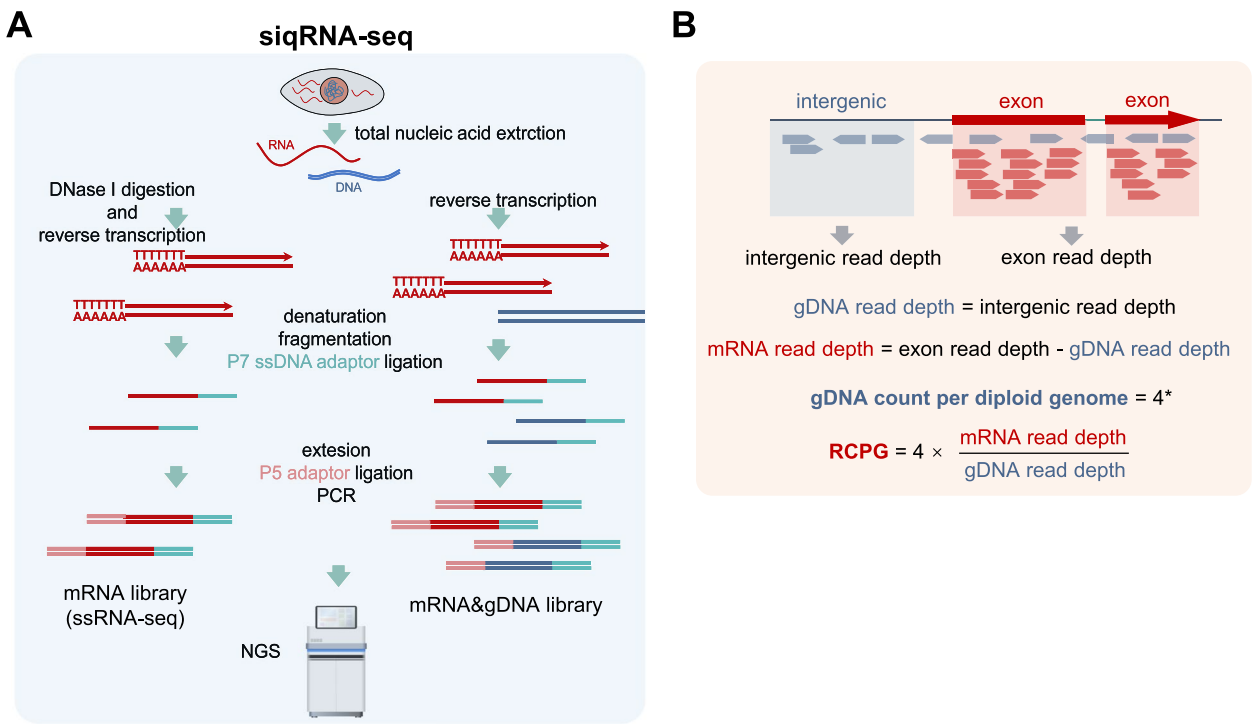

Fig. 1 Design of siqRNA-seq. **A** Flowchart of siqRNA-seq. Total nucleic acids were extracted and two types of libraries, mRNA library (ssRNA-seq) and mRNA&gDNA library, were constructed in parallel and sequenced for each sample in siqRNA-seq. **B** Principle of siqRNA-seq for RCPG calculation in mRNA&gDNA library. Depth of gDNA can be assessed by intergenic read depth. RCPG is equal to four times the ratio of mRNA read depth to gDNA depth. *: gDNA of diploid with two strands. RCPG: mRNA count per genome

values of highly expressed genes to establish a model for normalization (Fig. 1B). After normalization, FPKM values of ssRNA-seq data are transferred to RCPG values for all expressed genes. Together, siqRNA-seq is a ssDNA ligation-based method that constructs libraries for cDNA and gDNA in parallel, which can be applied to quantify gene expression by using gDNA as an internal reference.

Validation of siqRNA-seq for expression profiling

To investigate the reliability and reproducibility of siqRNA-seq for gene expression profiling, we performed siqRNA-seq on the cell lines HEK293T, IOSE-80, and HCT116 (Fig. 2A). The results of correlation analysis for ssRNA-seq showed that the expression profiles of replicates were highly correlated with each other (Pearson correlation coefficients were 0.991, 0.985, and 0.987 for HEK293T, IOSE-80, and HCT116, respectively), suggesting that whole nucleic acids can supply high-quality materials for siqRNA-seq library preparation (Additional file 1: Fig.S1A). By comparison, high correlations were also showed between ssRNA-seq and public RNA-seq data (Additional file 1: Fig.S1B), suggesting that ssRNA-seq is reliable and reproducible for gene expression profiling.

Next, we analyzed the ability of the mRNA&gDNA library to perform gene expression profiling. Although there was a high gDNA background, approximately 11.84% of the total mapped reads could be assigned to exons in mRNA&gDNA libraries, while only approximately 6.74% reads could be assigned to exons in gDNA libraries (Not all shown) (Additional file 2: Table S1), suggesting high efficiency for RNA profiling in the mRNA&gDNA library with gDNA. Similar to the ssRNA-seq results, RNA profiles from the mRNA&gDNA libraries remained well-correlated with each replicate (Fig. 2B), and correlated well with the ssRNA-seq data (Fig. 2C). These results demonstrated that our mRNA&gDNA library not only can be used as a reliable and reproducible method for gene expression profiling but also supply information on reads from gDNA for expression quantification.

Taken together, our data validated that both the mRNA&gDNA library and ssRNA-seq in siqRNA-seq are technically reliable and reproducible, which are the basis for accurate qualification of gene expression.

Pipeline and model of siqRNA-seq for gene expression quantification

Normally, gDNA in a diploid cell includes two copies for each site in double-stranded DNA formation. However,

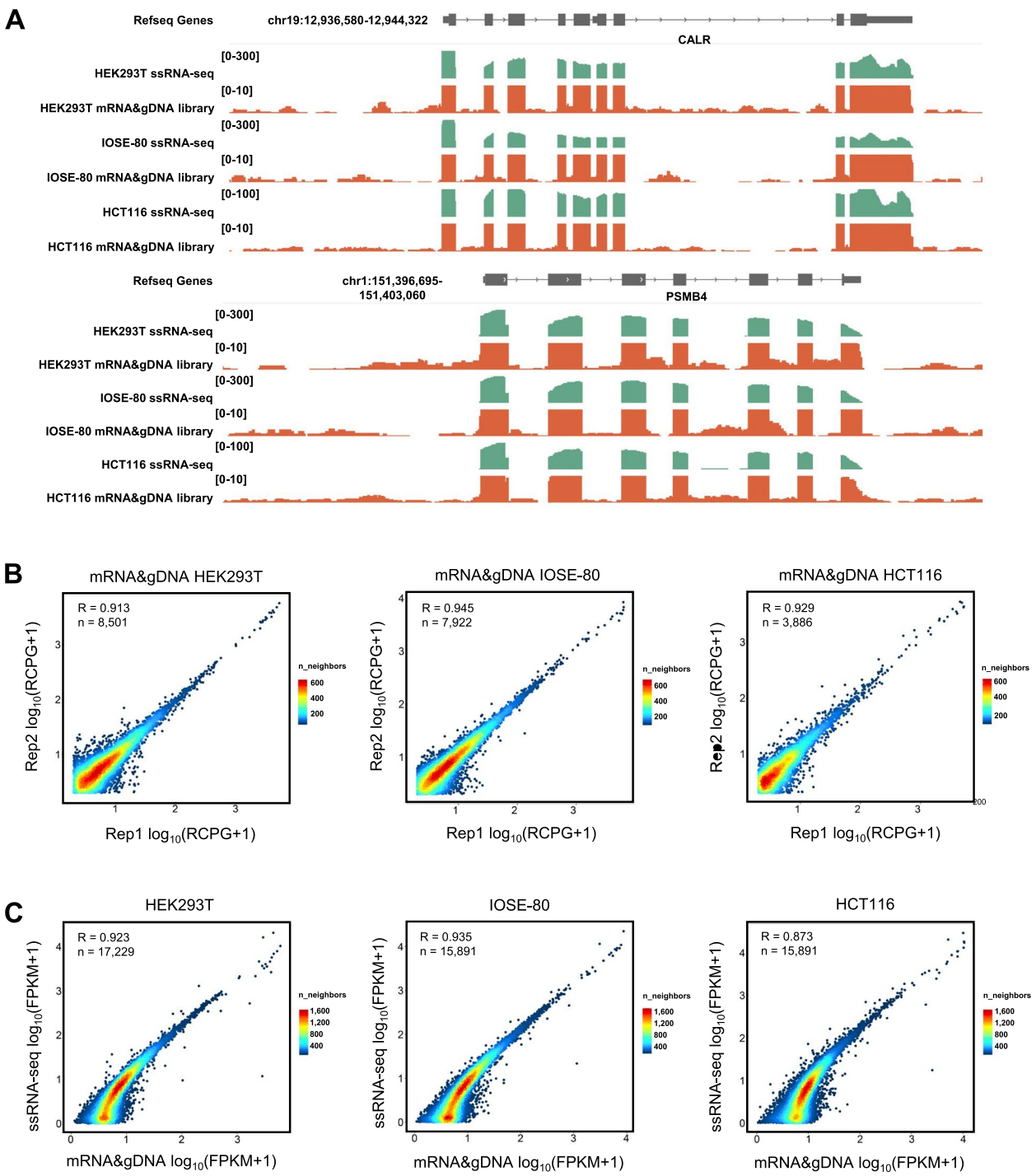

Fig. 2 Validation of siqRNA-seq for quantitative expression profiling. **A** IGV showing snapshots of siqRNA-seq signals (RPCG) in the human genome. Both RNA and gDNA signals can be sequenced in the mRNA&gDNA libraries. **B** Scatter plots showing the correlation between mRNA&gDNA library repeats. **C** Scatter plots showing the correlation of ssRNA-seq with mRNA&gDNA libraries

repeated sequences [17] and regions such as blacklists of ChIP-seq data in the genome may produce erroneous signals after alignment [18]. Therefore, we designed a pipeline to extract reliable regions to accurately assess the sequencing depth of gDNA (Fig. 3A). First, the genome was divided into consecutive windows with a 10 kb bin size and windows overlapping with any gene and downstream of the gene were removed. Then, counts of reads

from the gDNA library on each window were calculated. After sorting by sequencing depth, windows only with sequencing depths higher or lower than 10% of the median were selected to exclude windows with extreme depth. Considering that intergenic transcripts may lead to overestimation of gDNA depth, windows that could be mapped by reads from ssRNA-seq were finally removed. The remaining windows, named intergenic regions (IRs), were used to assess gDNA depth in the mRNA&gDNA library. Applying the pipeline for HEK293T cells, 125.78 Mb IRs were screened to calculate gDNA depth in the mRNA&gDNA library (Additional file 1: Fig.S2A).

Next, the model for siqRNA-seq was established to quantify gene expression as shown in Fig. 3B. First, the mapped reads of the mRNA&gDNA library were assigned to IRs, and genesto calculate their depth. Then, the RCPG of each gene can be calculated according to the formula shown in Fig. 3B (see Methods). The diploid genome is dsDNA with two copies for each region, while cDNA reverse-transcribed from mRNA is ssDNA. Thus, one mRNA read is equal to four gDNA reads sequenced in the mRNA&gDNA library. Due to the gDNA background, lowly expressed genes might not be accurately calculated by data from the mRNA&gDNA library. Then, we built a linear model to calibrate the RCPG values of all expressed genes through the integration of the mRNA&gDNA library and the ssRNA-seq (Not all shown) (Fig. 3B, Additional file 1: Figs.S2C and S4B). To establish the linear model with RANSAC [19] for siqRNA-seq, we selected a set of genes according to criteria described in Methods to obtain values of RCPG from the mRNA&gDNA library and expression levels from the ssRNA-seq for each sample (Not all shown) (Additional file 1: Figs.S2C, S4B and Additional file 2: Table S2). Eventually, the model can be used to calibrate the RCPG of each expressed gene using expression data from ssRNA-seq (Fig. 3B).

To validate the reliability of our model, we applied siqRNA-seq in the HEK293T and IOSE-80 cell lines to estimate the number of mRNA molecules per genome. In these cells, we detected approximately 118,15 and 153,469 mRNA molecules (Fig. 3C). In addition, we performed qPCR of seven genes with moderate gene expression and three intergenic regions as references to validate our sequencing data (see Methods). The results showed that the expression of these genes quantified by siqRNA-seq was consistent with the signals from qPCR (Fig. 3D), further demonstrating the reliability of our siqRNA-seq for gene expression quantification. Together, we established a reliable analysis pipeline and model in siqRNA-seq to quantify gene expression.

Example of siqRNA-seq for differential gene expression analysis

RNA-seq remains the primary tool for differential expression gene (DEG) analysis to study the change in expression of genes or transcripts under different conditions [1]. To exemplify the merits of siqRNA-seq for quantitative mapping of the mRNA landscape and DEG analysis, we performed siqRNA-seq on HEK293T cells treated with actinomycin D (ActD). ActD, which has been widely used to study mRNA stability, is a transcription inhibitor that does not significantly affect DNA replication or protein synthesis at low concentrations [20, 21]. Compared to the untreated control, the abundance of mRNA quantified by siqRNA-seq was greatly reduced in treated cells as expected (Fig. 4A, Additional file 1: Fig.S4A, and Additional file 2: Table S4). The results of DEG analysis using the siqRNA-seq method showed that almost all DEGs were downregulated (4,746 genes) except one gene that showed upregulation in our data (Fig. 4B). In contrast, if we use the traditional FPKM normalization method for ssRNA-seq data to quantify gene expression, the fold change trend of all genes in the traditional normalization method was higher than that in the siqRNA-seq method (Fig. 4C and Additional file 1: Fig.S4C). In summary, approximately 50% of genes showed an upregulation trend (Fig. 4C), and 2,119 genes were identified as upregulated genes by the FPKM normalization method (Fig. 4D). In addition, some downregulated genes of siqRNA-seq were assigned to upregulated genes in the FPKM normalization method, such as FUCA2, ARF5, GGCT, CCDC124, RPS20, CSDE1, MDH1, FHL1, and GRN. In line with siqRNA-seq, RT-qPCR confirmed that these genes were downregulated in ActD-treated cells (Additional file 1: Fig.S4D). These results together suggested that siqRNA-seq shows advantages in accurately identifying

(See figure on next page.)

Fig. 3 Pipeline and model of siqRNA-seq for gene expression quantification. **A** Schematic diagram showing the pipeline for intergenic regions (IRs) screening in the gDNA library. There IRs are used for assessment of gDNA depth in siqRNA-seq quantitative analysis. **B** Schematic diagram showing the model of siqRNA-seq for gene expression quantification. In the mRNA&gDNA library, a set of genes was selected for constructing a linear model using RCPG values from the mRNA&gDNA library and FPKM values from ssRNA-seq. Then, the established linear model was applied to transform the FPKM values of all genes in the ssRNA-seq to RCPG. **C** Bar plot showing the number of mRNA molecules per cell in HEK293T and IOSE-80 cells based on siqRNA-seq quantification. **D** The quantitative results of siqRNA-seq were verified by RT-qPCR in HEK293T cells. n.s.: not significant

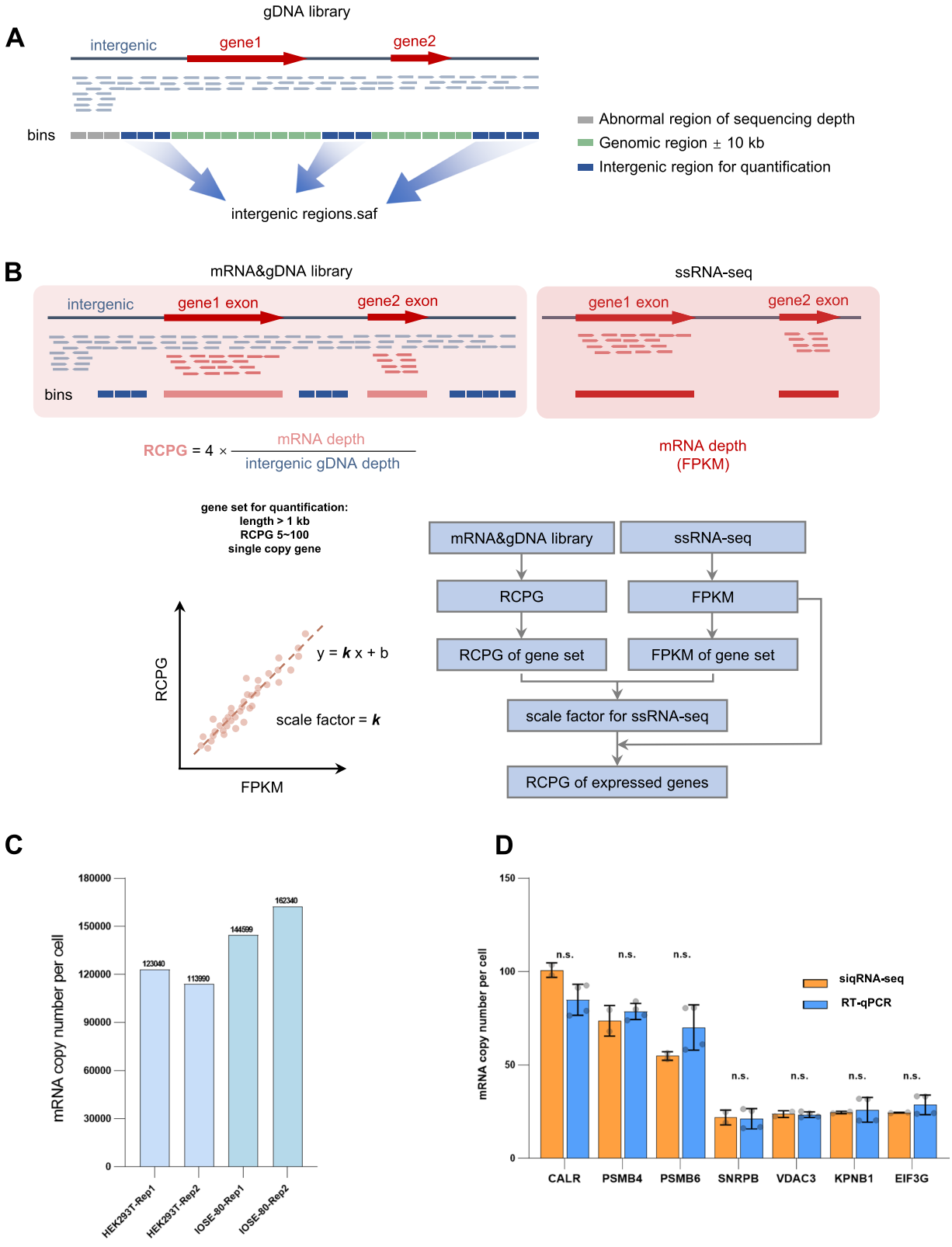

Fig. 3 (See legend on previous page.)

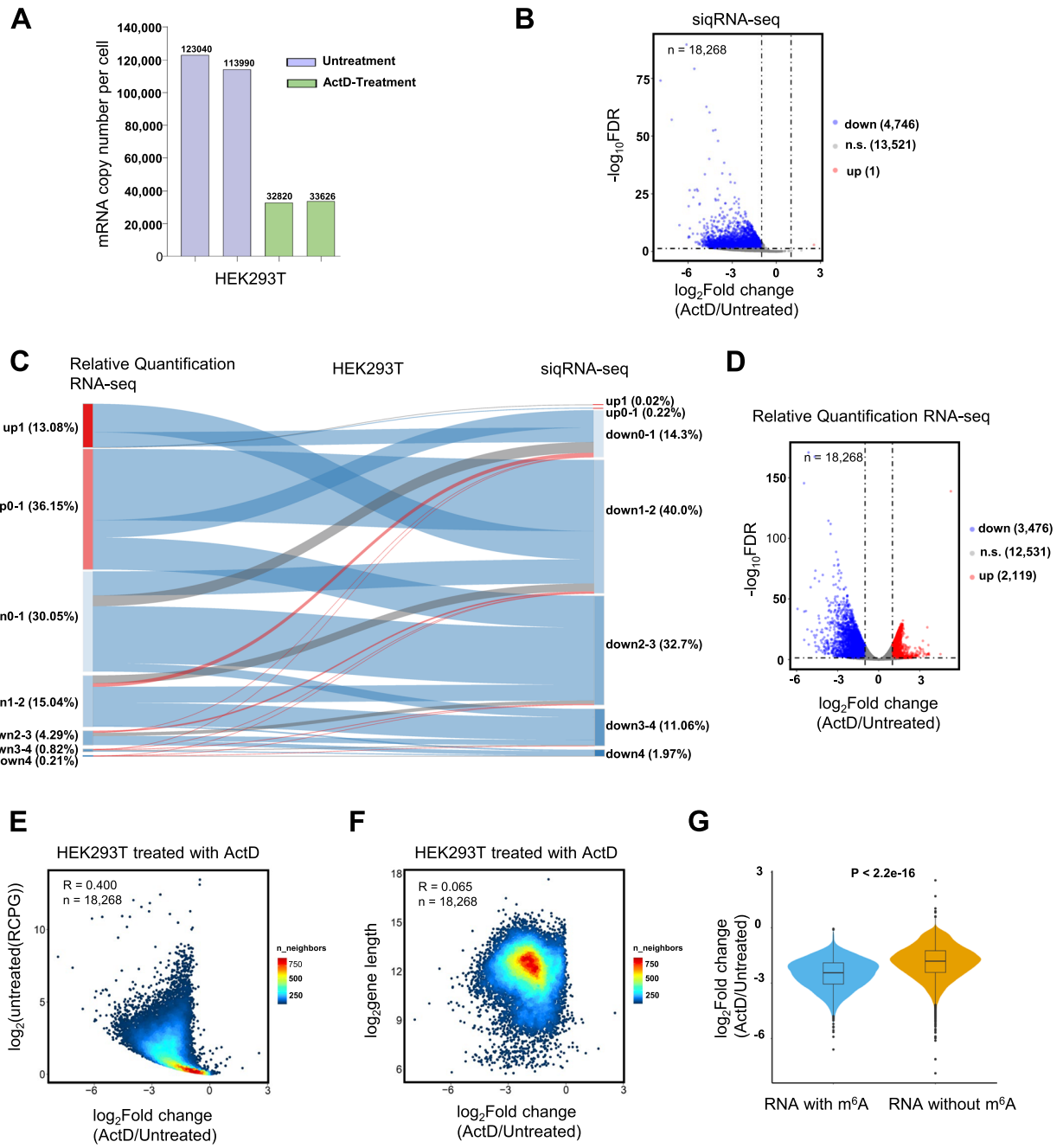

Fig. 4 Analysis of factors influencing the downward trend of ActD-regulated gene expression by siqRNA-seq. **A** Bar plot showing absolute quantification of mRNA molecules per cell for ActD-treated and untreated HEK293T cells by siqRNA-seq. **B** Volcano plot showing the results of the DEGs identified by siqRNA-seq for HEK293T cells treated with ActD compared to untreated control. Genes with fold change greater than 2 and FDR less than 0.01 were assigned as DEGs. **C** The Sankey diagram showing the trend of genes with different fold changes(log_2) of Relative Quantification RNA-seq and siqRNA-seq for HEK293T cells treated with ActD. In Relative Quantification RNA-seq, nearly 50% of genes showed an upregulation trend, while siqRNA-seq showed almost no upregulation. **D** Volcano plot showing the results of DEGs identified by Relative Quantification RNA-seq analysis for HEK293T cells treated with ActD compared to untreated control. Genes with a fold change greater than 2 and FDR less than 0.01 were assigned as DEGs

differentially expressed genes between samples with distinct global mRNA levels using gDNA as an internal reference (Additional file 2: Table S5).

ActD intercalates into DNA to form a very stable complex with DNA to prevent the unwinding of the DNA double-helix; thus, transcription could be globally

inhibited [21]. However, the degradation rates may be distinct among these expressed genes, as our data shown (Fig. 4B and D). Next, we investigated some potential factors that may impact the degradation of genes responding to ActD. First, we performed a correlation analysis between the fold change (FC) of gene expression after ActD treatment and the gene expression levels in the control. Our results showed that the Pearson correlation coefficient was approximately 0.400 between them (Fig. 4E), suggesting that the gene degradation rate is positively correlated with the expression level. We also observed that the FC of genes downregulated by ActD was slightly positively correlated with gene length (Pearson $r=0.065$) (Fig. 4F). Additionally, we investigated the influence of N6-methyladenosine (m^6A) modification on changes in expression induced by ActD. m^6A, a chemical derivative of adenosine in RNA, has been increasingly reported as an important RNA modification involved in RNA metabolism, such as altering pre-mRNA processing, promoting mRNA nuclear export, changing mRNA stability, and increasing translation efficiency [22, 23]. It has been reported that some genes, such as YTHDF1, YTHDF2, and YTHDF3, trigger the rapid decay of m^6A-modified transcripts [24, 25]. Comparing m^6A-modified mRNA with mRNA without m^6A modification, we observed a larger change in m^6A modified mRNA than mRNA without m^6A modification, indicating that m^6A mediates mRNA decay as previously reported (Fig. 4G) [24, 25].

High diversity of total mRNA expression in tumor cell lines

In recent years, due to the continuously rising incidence of cancer, it has become a significant global public health issue. Consequently, research on tumor cells remains a hotspot [26–28]. We applied siqRNA-seq to tumor cell lines HCT116, A498, DU145, NCI-H226, SK-OV-3, MDA-MB-468, and SW620 to estimate the number of mRNA molecules per genome. In these cells, we detected approximately 64,410, 76,500, 142,008, 55,742, 56,865, 72,380, and 47,147 mRNA molecules (Fig. 5A), indicating a high degree of diversity in total mRNA expression among tumor cell lines. Through analysis of the mRNA quantification data from these tumor cell lines, we found that the total mRNA content in DU145 tumor cell line was significantly higher than in other tumor cell lines. Therefore, we further investigated the underlying reasons for this phenomenon and discovered that it was due to the overall higher levels of gene expression rather than a few specific genes (Fig. 5B). Together, with the assistance of siqRNA-seq technology, it was revealed that there is a rich diversity in total RNA expression among tumor cells. Regarding the diversity of total RNA expression in tumor cells, we speculate that it may be attributed to changes in genomic polyploidy [29–31].

Discussion

In this study, we present a novel method, siqRNA-seq, to profile and quantify global gene expression (Figs. 1A and 3A). Without spike-ins, siqRNA-seq integrates reads from gDNA and mRNA to construct a model for quantification of mRNA copy number per genome or per cell (Figs. 1B and 3B). Thus, siqRNA-seq allows the accurate identification of differentially expressed genes between samples with distinct mRNA abundances to study RNA dynamics under different conditions (Additional file 1: Fig.S4A).

Although RNA-seq has been applied in various fields for decades [32], DEG analysis remains at the essential stage for gene expression studies [1]. Current normalization methods for DEG analysis generally assume that cells have similar RNA abundance, however, this can result in erroneous conclusions for cells globally expressing different RNA levels, such as increased mRNA abundance correlating with high metabolic activity in active cell cycling cells [33–35]. To quantify the abundance of gene expression, several methods have been reported, such as counting cell numbers to measure different RNA abundances based on the Quant-iT assay [33], counting cell numbers combined with adding a known quantity of spike-in RNAs [36], or constructing a linear correlation between sequencing read count and RNA abundance [37]. However, these methods use absorbance or fluorescence measurements to estimate RNA content per cell with risks of large technical error. The process of cell counting might delay the time point and change the gene expression pattern. Although spike-in RNAs, such as commercially available ERCC, can be used to measure sensitivity, accuracy, and biases in RNA-seq experiments as well as to derive standard curves for quantifying the abundance of transcripts, they are still highly expensive [9]. Besides, RNase is ubiquitous. RNA standards must be carefully processed to prevent them degradation. In contrast, siqRNA-seq, which uses reads from gDNA as the internal reference to quantify gene expression levels, enables us to assess the copy number of mRNA per cell or per genome, with the advantages of no requirement of cell counting and spike-in controls. In addition, the ratio of mRNA and DNA in whole nucleic acids is stable during dilution to a needed concentration for library construction. For materials, such as multinucleated skeletal muscle or polyploid plants, it may be interesting to investigate their expression profiles per genome. Therefore, we consider siqRNA-seq to be an extended RNA-seq

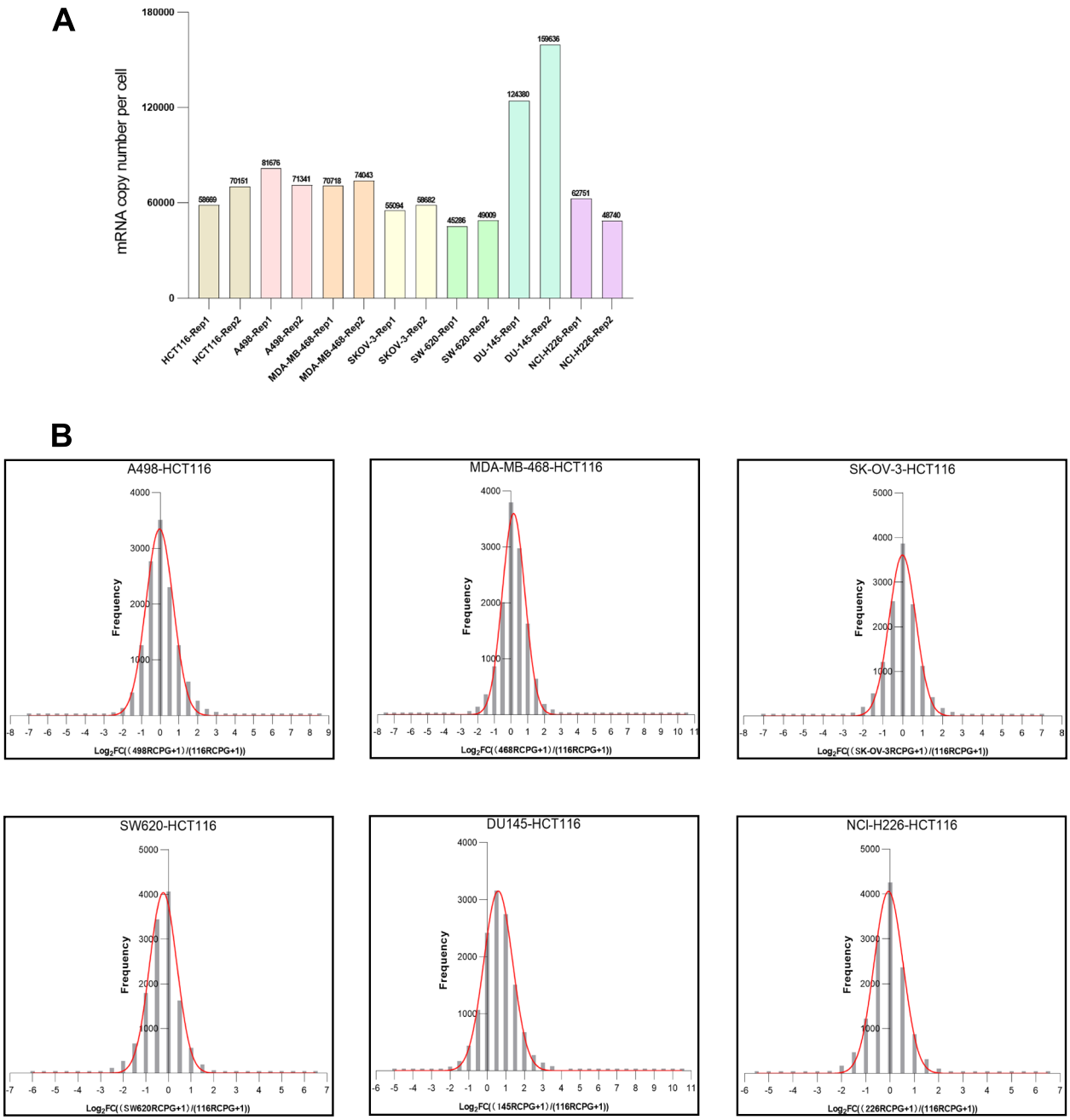

Fig. 5 High diversity of total mRNA expression in tumor lines. **A** The bar chart showing the quantitative results of total mRNA content in seven tumor cell lines obtained through siqRNA-seq. **B** The bar charts showing the density distribution of fold changes in differential expression compared to tumor HCT116 across these tumor cell lines

method to quantify expression in cells with multiple nuclei.

As with all high-throughput sequencing methods, siqRNA-seq has limitations. In siqRNA-seq, RNAs were reverse-transcribed with oligo(dT) to synthesize cDNA. The efficiency of reverse transcription greatly impacts the accuracy of quantification. However, the efficiency can be improved by high-efficient reverse transcriptase in optimized reaction buffer. In addition, the efficiency is equal for samples in the same experiment without influencing the expression comparison. Another limitation of our siqRNA-seq is that we only tested the whole nucleic acid extraction method on mammalian cells. For samples, such as microorganisms or plant tissues with cell walls,

we may need to optimize the extraction method for different materials. Additionally, cells in the G1 phase of meiosis are required to normalize mRNA copy number to "per cell" by siqRNA-seq, it might be not easy for many samples to isolate cells in the G1 phase. Thus, the "per genome" method may be more practical to normalize data for comparative analysis of gene expression among conditions.

Conclusion

In summary, siqRNA-seq is a spike-in independent quantitative RNA sequencing method, which creatively uses gDNA as an internal reference to quantify gene expression. siqRNA-seq enables us to assess the copy number of mRNA per genome or cell with no requirement of cell counting and spike-ins. We consider that siqRNA-seq can be supplied as a complementary tool to profile and quantify gene expression in many fields.

Methods

Cell preparation

Human embryonic kidney cells (HEK293T), human normal ovarian cells (IOSE-80), human colon cancer cells (HCT116), Human renal cell carcinoma cells (A498), human prostate cancer cells (DU145), human squamous lung cancer cells (NCl-H226), human ovarian cancer cells (SK-OV-3), human breast cancer cells (MDA-MB-468), and human colorectal adenocarcinoma cells (SW-620) were purchased from Xiaofan Technology, Guangzhou. They were cultured simultaneously in a medium containing 10% fetal bovine serum (Gibco) and 1% penicillin–streptomycin (Gibco) at 37 °C with 5% CO_2. When cells reached approximately 90% confluence, they were collected for nucleic acid extraction.

For actinomycin D treatment, cells were cultured to 80% convergence and changed to fresh medium with actinomycin D at a final concentration of 5 μg/ml for 12 h. Then, the medium was removed, and the cells were carefully washed once with PBS. To collect cells, cultures were digested with 0.25% Trypsin–EDTA (Gibco) at 37 °C for 5 min. Untreated HEK293T cells were cultured simultaneously as controls.

Nucleic acid extraction from cells

Total nucleic acids were extracted from the collected cells by SDS and proteinase K methods. Briefly, approximately 1×10^6 cells were resuspended in 4 ml TE buffer with 0.5% SDS and 0.1 mg/ml proteinase K and incubated at 37 °C in a shaker at 200 rpm for 4 h. Then, 1/4 volume 5 M potassium acetate was added, mixed well, and placed on ice for 15 min. Finally, the total nucleic acids were purified by the phenol–chloroform extraction method and precipitated with 1 volume of isopropanol. Purified total nucleic acids can be stored at -80 °C until later use.

Genomic DNA (gDNA) library construction

During siqRNA-seq quantification, the gDNA depth of the mRNA&gDNA libraries was assessed by intergenic regions (IRs) that were picked out from gDNA libraries. To construct the gDNA library, the extracted total nucleic acid was dissolved in enzyme-free water and digested with RNase A at 37 °C for 40 min to remove RNA. Then, the gDNA was fragmented by sonication for library preparation by the Accel-NGS 1S Plus DNA Library Kit (Swift Accel-NGS) according to the manual. We compared the IRs of three cell lines (HEK293T, IOSE-80, and HCT116) and found that their IRs are almost identical. Therefore, this study utilized the IRs identified in HEK293T cells for quantitative analysis of siqRNA-seq (Additional file 1: Fig.S2B).

siqRNA-seq library preparation

siqRNA-seq includes two libraries for each sample, the mRNA&gDNA library and ssRNA-seq library. Different from the mRNA&gDNA library, total nucleic acids need to be digested with DNase I at 37 °C for 40 min to remove DNA for the ssRNA-seq library. To reduce potential DNA contamination due to incomplete DNase I digestion as possible, we have used the same amount of the total nucleic acid in the digestion reaction and monitored by Qubit™ 1X dsDNA HS Assay Kit. Then, reverse transcription was performed with oligo(dT) to synthesize the first strand of complementary DNA (cDNA) for both libraries. After extraction and purification, approximately one-fifth of reversed products without sonication were left for qPCR and others were ultrasonically fragmented to 300 bp using ME220 (Covaris, 70 W, 20% Duty factor, 1,000 cycles per burst, 130 s, at 4 °C) for library preparation using the Accel-NGS 1S Plus DNA Library Kit according to the manual, which can use ssDNA as substrate for library preparation.

Routine transcriptome analysis

The libraries were sequenced on NovaSeq 6000 platform. The data were checked by Fastp (version 0.23.1) software [38] and mapped to the GRCh38 genome via HISAT2 (version 2.2.1) software [39]. Using two normalization methods to output BigWig files, one is RPCG normalization: BamCoverage from Deeptools (version 3.5.1) [40] is used to convert Bam format to BigWig format and normalize it to 1×sequencing depth (RPGC). Another type of normalization is based on IRs: the multiBigwigSummary BED-file (version 3.5.1) for IRs was used to calculate the average score of the BigWig files and the following formula was used to calculate the scale factor.

Corrected BigWig files were generated using the scale Factor parameter in BamCoverage.

SRs_{avg} is the average of the average scores for all regions. *IRs* are the intergenic regions. IRs_{avg} represents the average depth of *IRs*.

$$scale\ factor\ = \frac{SRs_{avg}}{IRs_{avg}}$$

The gene expression profiles were quantified to obtain the fragments per kilobase of the exon model per million mapped fragments (FPKM). Read counts were quantified with FeatureCounts software [41]. PCA in R was performed using the prcomp function. The correlation of the samples was determined by Pearson's method, using the cor function. The differential expression analysis was performed using edgeR (version 3.26) [42]. Cut-off values $|\log2(FC)|>1$ and $Padj<0.05$ were used for differential gene expression analysis. The Python package pyecharts was used to draw Sankey diagrams.

To determine the minimal number of reads required for mRNA and gDNA libraries to confidently normalize FPKM values in ssRNA-seq to RCPG values, we randomly subsampled raw fastq files with different numbers of reads and compared their correlations to the original data (Additional file 1: Fig.S5). According to the results, we found that at least 10 million reads are needed for mRNA & gDNA libraries to establish a reliable normalization model.

Quantitative analysis of siqRNA-seq

The siqRNA-seq quantification process was initiated with the step of splitting the gDNA library into 10 kb windows using Bedtools make windows (version 2.30.0) [43]. Then, windows with sequencing depths 10% above or below the median were sorted out after statistical calculation by multiBamSummary of Deeptools (version 3.5.1) [40]. In particular, the windows overlapping with any gene or its 10 kb flanking region were discarded, and those overlapping with regions that could be mapped by the ssRNA-seq reads were also removed. The remaining windows were named intergenic regions (IRs). Mapped reads of the mRNA&gDNA library for the genomic region, IRs, and genes were counted to gain the depth of each feature. The RCPG was calculated as follows:

$$RCPG\ i = \frac{\left(G_i - IRs_{avg}\right)}{IRs_{avg}} \times 4$$

In the formula, "*i*" represents gene i, "*Gi*" represents the depth of gene i, and IRs_{avg} represents the average depth of *IRs*. The diploid genome is double-stranded DNA, which means that there will be two copies of each region. cDNA reversely transcribed from mRNA has single-stranded DNA. Thus, one mRNA read is equal to four gDNA sequencing reads in the mRNA&gDNA library as shown in the formula.

The gDNA reads in the mRNA&gDNA library may result in some low-expressed genes not being accurately calculated from the data in the mRNA&gDNA library. Therefore, correction of the expressed gene RCPG in ssRNA-seq was performed by constructing a linear model. To construct a linear model, we selected a set of genes that met the following criteria.

(1) Considering that some genes have multiple copies or pseudogenes in the genome. Genes with outlier depth in the gDNA library were removed.
(2) Lengths of genes less than 1,000 bp were removed to reduce bias in genes with as short a length as possible.
(3) Genes with RCPG values obtained from mRNA&gDNA library should be between 5 and 100 to reduce interference from possible outliers. The number of genes with RCPG values between 5 and 100 is consistent with the expected moderate expression levels of genes that exclude quantitative biases caused by under- or overexpression of genes (Additional file 2: Table S6).

Overall, we used RANSAC [19] to screen eligible genes in mRNA&gDNA libraries and construct a linear model (Additional file 1: Fig. 2C). Then the FPKM of all expressed genes in ssRNA-seq was calculated. and corrected by this linear model. The final result was the RCPG expressed by each gene.

Quantitative verification by RT-qPCR

To test the accuracy of our siqRNA-seq quantification, RT-qPCR was used for validation. According to the results of the siqRNA-seq sequencing, seven genes with moderate gene expression and three intergenic regions were selected. Primers were designed for genes and intergenic regions for RT-qPCR validation. First, a standard curve was established for the ten pairs of primers (Additional file 1: Fig.S3). Since the gene primers are located in the exon, gDNA can be used as the amplification template. Using gDNA as a template, five gradients (20 ng/μl, 10 ng/μl, 5 ng/μl, 2.5 ng/μl, 1.25 ng/μl) were diluted by 2× dilution. qPCR was performed with designed primers, and each primer was repeated 3 times. According to the output Cq value of each dilution gradient, the standard curve and the standard curve equation of each primer

were calculated. Our experimental samples were then diluted within the dilution gradient range (1.25—20 ng/μl), and RT-qPCR was performed on the samples using gene and intergenic region primers. One Cq value was output for each gene and intergenic region primer, and the Cq values of the two primers were substituted into the equation for their respective primers. The amount of gDNA per diploid genome is four, so the ratio of four times the gene primer values to intergenic regional primer values is RCPG. RT-qPCR quantification values and siqRNA-seq quantification values differed within 10%. Primers for HEK293T cell genes and intergenic regions were designed according to the design principle of fluorescence primers (Additional file 2: Table S7).

Verification of gene differential expression analysis

The transcriptome of HEK293T cells treated with actinomycin D for 12 h was analyzed by Relative Quantification RNA-seq and siqRNA-seq. We found that many genes were upregulated in Relative Quantification RNA-seq, while the expression of these genes in siqRNA-seq was downregulated. Examples include FUCA2, ARF5, GGCT, CCDC124, RPS20, CSDE1, MDH1, FHL1, and GRN. Primers were designed for these nine genes and a housekeeping gene, GAPDH, and RT-qPCR confirmed that these genes were downregulated in ActD-treated cells. The results show that siqRNA-seq shows advantages in accurately identifying differentially expressed genes between samples with different global mRNA levels. Primer design was performed on selected genes according to fluorescent primer design principles (Additional file 2: Table S8).

Abbreviations

siqRNA-seq	Spike-in-independent quantitative RNA sequencing
DEGs	Differentially expressed genes
gDNA	Genomic DNA
cDNA	Complementary DNA
ssDNA	Single-strand DNA
ssRNA-seq	Single-strand DNA ligation-based RNA sequencing
IRs	Intergenic regions
RCPG	The mRNA count per diploid genome
RCPC	The mRNA count per cell
FPKM	Fragments Per Kilobase Million
ActD	Actinomycin D
FC	The fold change
m^6A	N6-methyladenosine
PBS	Phosphate buffered saline
EDTA	Ethylene Diamine Tetraacetie Acid
DMEM	Dulbecco's Modified Eagle Medium
TE	Tris–EDTA
RT-qPCR	Quantitative reverse transcription PCR
SDS	Sodium dodecyl sulfate
PCA	Principal component analysis
RPGC	Reads Per Genomic Content, defined as (total number of mapped reads * fragment length) / effective genome size

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-024-10650-2.

Additional file 1: Fig. S1. Quality control of siqRNA-seq. A Scatter plots showing the correlation between repeats of ssRNA-seq. B Scatter plots showing the correlation of ssRNA-seq and data from public databases. Fig. S2. siqRNA-seq for gene expression quantification. A Pie chart showing the size of abnormal sequencing depth regions, ssRNA-seq signal regions, and IRs in the genome. B Bar graph showing the region of IRs of the three cell lines HEK293T, IOSE-80, and HCT116. C Scatter plots showing the construction of linear models in the quantitative process. Fig. S3. RT-qPCR standard curve and standard curve equation. Validate the standard curve and standard curve equation for siqRNA-seq quantification. Fig. S4. siqRNA-seq for differential gene expression analysis. A IGV snapshots showing siqRNA-seq signals (IRs) in the human genome. Gene expression was greatly reduced in ActD drug-treated HEK293T cells compared with untreated controls. B Scatter plots showing the construction of linear models for gene expression quantification of untreated HEK293T cells and cells treated with ActD for 12 h. C Scatter plot showing the trend of fold change for all genes analyzed by siqRNA-seq compared to Relative Quantification RNA-seq for HEK293T cells treated with ActD for 12 h. D Bar plot showing RT-qPCR validation of FUCA2, ARF5, GGCT, CCDC124, RPS20, CSDE1, MDH1, FHL1, and GRN genes downregulated in HEK293T cells with ActD treatment. **: *p*- Value < 0.01; ***: *p*- Value < 0.001. Figure S5. Subsampling analysis showing the minimal number of reads required for siqRNA-seq. A Correlation analysis for data between subsampling mRNA & gDNA libraries and the raw mRNA & gDNA library. B Correlation analysis for data between subsampling ssRNA-seq and the raw ssRNA-seq. C Correlation analysis for data between subsampling mRNA & gDNA libraries and the raw ssRNA-seq.

Additional file 2.

Acknowledgements

We greatly appreciate all the Wei Xu Lab for useful discussions. We would like to thank Jinjin Li and Jianli Yan for laboratory support.

Authors' contributions

WX and CS conceived, designed, and supervised the experiments. ZW, CS, and KT contributed equally to this work. ZW developed the new technique with the help of CS. Z W and JJ assisted KT in completing the data analysis. WX, ZW, and CS wrote the manuscript. All authors read and approved the final manuscript.

Funding

This study was supported by National Key R&D Program of China (Grant No. 2021YFF1000600), the Youth Innovation Program of Chinese Academy of Agricultural Sciences (No. Y2022QC33), National Natural Science Foundation of China (Grant Nos. 32071437 and 32100423), and China Postdoctoral Science Foundation (Grant No. 2022M713420).

Availability of data and materials

All data for this study are publicly available and can be accessed via a link. (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE223145).

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 4 April 2024 Accepted: 22 July 2024
Published online: 30 July 2024

References

1. Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. Nat Rev Genet. 2019;20:631–56
2. Kaczkowski B, Tanaka Y, Kawaji H, Sandelin A, Andersson R, Itoh M, et al. Transcriptome Analysis of Recurrently Deregulated Genes across Multiple Cancers Identifies New Pan-Cancer Biomarkers. Cancer Res. 2016;76:216–26.
3. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol. 2015;33:495–502.
4. Wen L, Li G, Huang T, Geng W, Pei H, Yang J, et al. Single-cell technologies: From research to application. Innovation (Camb). 2022;3:100342.
5. Zhang Y, Lin X, Yao Z, Sun D, Lin X, Wang X, et al. onvolution algorithms for inference of the cell-type composition of the spatial transcriptome. Comput Struct Biotechnol J. 2023;21:176–84.
6. Janjic A, Wange LE, Bagnoli JW, Geuder J, Nguyen P, Richter D, et al. Prime-seq, efficient and powerful bulk RNA sequencing. Genome Biol. 2022;23:88.
7. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016;17:13.
8. Zyprych-Walczak J, Szabelska A, Handschuh L, Górczak K, Klamecka K, Figlerowicz M, et al. The Impact of Normalization Methods on RNA-Seq Data Analysis. Biomed Res Int. 2015;2015: 621690.
9. Jiang L, Schlesinger F, Davis CA, Zhang Y, Li R, Salit M, et al. Synthetic spike-in standards for RNA-seq experiments. Genome Res. 2011;21:1543–51.
10. Lin CY, Lovén J, Rahl PB, Paranal RM, Burge CB, Bradner JE, et al. Transcriptional amplification in tumor cells with elevated c-Myc. Cell. 2012;151:56–67.
11. Nie Z, Hu G, Wei G, Cui K, Yamane A, Resch W, et al. c-Myc Is a Universal Amplifier of Expressed Genes in Lymphocytes and Embryonic Stem Cells. Cell. 2012;151:68–79.
12. Lovén J, Orlando DA, Sigova AA, Lin CY, Rahl PB, Burge CB, et al. Revisiting Global Gene Expression Analysis. Cell. 2012;151:476–82.
13. Risso D, Ngai J, Speed TP, Dudoit S. Normalization of RNA-seq data using factor analysis of control genes or samples. Nat Biotechnol. 2014;32:896–902.
14. Xu W, Li K, Li Q, Li S, Zhou J, Sun Q. Quantitative, Convenient, and Efficient Genome-Wide R-Loop Profiling by ssDRIP-Seq in Multiple Organisms. Methods Mol Biol. 2022;2528:445–64.
15. Xu W, Liu X, Li J, Sun C, Chen L, Zhou J, et al. ULI-ssDRIP-seq revealed R-loop dynamics during vertebrate early embryogenesis. Cell Insight. 2024;3:100179.
16. Xu W, Liu C, Zhang Z, Sun C, Li Q, Li K, et al. DEtail-Seq is an Ultra-efficient and Convenient Method for Meiotic DNA Break Profiling in Multiple Organisms. SCLS. 2023. https://doi.org/10.1007/s11427-022-2277-y.
17. Alexander RP, Fang G, Rozowsky J, Snyder M, Gerstein MB. Annotating non-coding regions of the genome. Nat Rev Genet. 2010;11:559–71.
18. Amemiya HM, Kundaje A, Boyle AP. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. Sci Rep. 2019;9:9354.
19. Martínez-Otzeta JM, Rodríguez-Moreno I, Mendialdua I, Sierra B. RANSAC for Robotic Applications: A Survey. Sensors (Basel). 2022;23:327.
20. Khanna KK, Jackson SP. DNA double-strand breaks: signaling, repair and the cancer connection. Nat Genet. 2001;27:247–54.
21. Ratnadiwakara M, Änkö M-L. mRNA Stability Assay Using transcription inhibition by Actinomycin D in Mouse Pluripotent Stem Cells. Bio Protoc. 2018;8:e3072.
22. He PC, He C. m6 A RNA methylation: from mechanisms to therapeutic potential. EMBO J. 2021;40:e105977.
23. Lee Y, Choe J, Park OH, Kim YK. Molecular Mechanisms Driving mRNA Degradation by m6A Modification. Trends Genet. 2020;36:177–88.
24. Du H, Zhao Y, He J, Zhang Y, Xi H, Liu M, et al. YTHDF2 destabilizes m6A-containing RNA through direct recruitment of the CCR4–NOT deadenylase complex. Nat Commun. 2016;7:12626.
25. Shi H, Wang X, Lu Z, Zhao BS, Ma H, Hsu PJ, et al. YTHDF3 facilitates translation and decay of N6-methyladenosine-modified RNA. Cell Res. 2017;27:315–28.
26. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA A Cancer J Clinicians. 2018;68:394–424.
27. McGuire S. World Cancer Report. Geneva, Switzerland: World Health Organization, International Agency for Research on Cancer, WHO Press, 2015. Adv Nutr. 2014;2016(7):418–9.
28. Shen B, Wang Z, Li Z, Song H, Ding X. Circular RNAs: an emerging landscape in tumor metastasis. Am J Cancer Res. 2019;9:630–43.
29. Wei Dai YY. Genomic Instability and Cancer. J Carcinog Mutagen. 2014;05.
30. Lau TY, Poon RYC. Whole-Genome Duplication and Genome Instability in Cancer Cells: Double the Trouble. IJMS. 2023;24:3733.
31. Bielski CM, Zehir A, Penson AV, Donoghue MTA, Chatila W, Armenia J, et al. Genome doubling shapes the evolution and prognosis of advanced cancers. Nat Genet. 2018;50:1189–95.
32. Emrich SJ, Barbazuk WB, Li L, Schnable PS. Gene discovery and annotation using LCM-454 transcriptome sequencing. Genome Res. 2007;17:69–73.
33. Monaco G, Lee B, Xu W, Mustafah S, Hwang YY, Carré C, et al. RNA-Seq Signatures Normalized by mRNA Abundance Allow Absolute Deconvolution of Human Immune Cell Types. Cell Rep. 2019;26:1627-1640.e7.
34. Wagner GP, Kin K, Lynch VJ. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory Biosci. 2012;131:281–5.
35. Yu R, Vorontsov E, Sihlbom C, Nielsen J. Quantifying absolute gene expression profiles reveals distinct regulation of central carbon metabolism genes in yeast. Elife. 2021;10:e65722.
36. Schertzer MD, Murvin MM, Calabrese JM. Using RNA Sequencing and Spike-in RNAs to Measure Intracellular Abundance of lncRNAs and mRNAs. Bio Protoc. 2020;10:e3772.
37. Hu JF, Yim D, Ma D, Huber SM, Davis N, Bacusmo JM, et al. Quantitative mapping of the cellular small RNA landscape with AQRNA-seq. Nat Biotechnol. 2021;39:978–88.
38. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34:i884–90.
39. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37:907–15.
40. Ramírez F, Dündar F, Diehl S, Grüning BA, Manke T. deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res. 2014;42 Web Server issue:W187–191.
41. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30:923–30.
42. Robinson MD, McCarthy DJ, Smyth GK. edgeR : a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139–40.
43. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.

Publisher's Note