Received: 29 September 2023 | Revised: 20 January 2024 | Accepted: 30 January 2024

DOI: 10.1002/biot.202300516

Biotechnology Journal

RESEARCH ARTICLE

# A single-cell genomic strategy for alternative transcript start sites identification

Yanling Peng[1] | Qitong Huang[1,2] | Danli Liu[1] | Siyuan Kong[1] | Rui Kamada[3] | Keiko Ozato[4] | Yubo Zhang[1,5] | Jun Zhu[6]

[1]Animal Functional Genomics Group, Shenzhen Branch, Guangdong Laboratory for Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, China

[2]Animal Breeding and Genomics, Wageningen University & Research, Wageningen, Netherlands

[3]Department of Chemistry, Faculty of Science, Hokkaido University, Sapporo, Japan

[4]Division of Developmental Biology, National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, Maryland, USA

[5]Kunpeng Institute of Modern Agriculture at Foshan, Foshan, China

[6]DNA Sequencing and Genomics Core, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, Maryland, USA

**Correspondence**
Yubo Zhang, Animal Functional Genomics Group, Shenzhen Branch, Guangdong Laboratory for Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen 518120, China.
Email: ribon_001@163.com

Jun Zhu, DNA Sequencing and Genomics Core, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD 20892, USA.
Email: jz@mokobious.com.

**Funding information**
National Key Research and Development Program of China, Grant/Award Number: 2018YFA0903201; National Natural Science Foundation of China, Grant/Award Number: 2017M620977; NICHD, Grant/Award Number: ZIA HD001310-35; China Postdoctoral Science Foundation, Grant/Award Numbers: BX2021367, 2021M703543; the Science, Technology and Innovation Commission of Shenzhen Municipality, Grant/Award Number: JCYJ20180306173714935; National Key Research and Development Program of China, the National Natural Science Foundation of China, Grant/Award Number: 32202653

## Abstract

Alternative transcription start sites (TSSs) usage plays a critical role in gene transcription regulation in mammals. However, precisely identifying alternative TSSs remains challenging at the genome-wide level. We report a single-cell genomic technology for alternative TSSs annotation and cell heterogeneity detection. In the method, we utilize Fluidigm C1 system to capture individual cells of interest, SMARTer cDNA synthesis kit to recover full-length cDNAs, then dual priming oligonucleotide system to specifically enrich TSSs for genomic analysis. We apply this method to a genome-wide study of alternative TSSs identification in two different IFN-$\beta$ stimulated mouse embryonic fibroblasts (MEFs). The data clearly discriminate two IFN-$\beta$ stimulated MEFs. Moreover, our results indicate 81% expressed genes in these two cell types containing multiple TSSs, which is much higher than previous predictions based on Cap-Analysis Gene Expression (CAGE) (58%) or empirical determination (54%) in various cell types. This indicates that alternative TSSs are more pervasive than expected and implies our strategy could position them at an unprecedented sensitivity. It would be helpful for elucidating their biological insights in future.

KEYWORDS
analytical biotechnology, bioinformatics

**Abbreviations:** *ALK*, anaplastic lymphoma kinase; *CAGE*, Cap-Analysis Gene Expression; *DPO*, dual priming oligonucleotides; *ISG*, IFN-stimulated gene; *MEF*, mouse embryonic fibroblast; *ScISOr-Seq*, single-cell isoform RNA-Seq; *scTSS-seq*, single cell TSSs sequencing; *SUPeR-seq*, single-cell universal poly(A)-independent RNA sequencing; *TP*, tumor protein; *t-SNE*, t-distributed Stochastic Neighbor Embedding; *TSS*, transcription start site.

Yanling Peng, Qitong Huang, and Danli Liu contributed equally to this study.

  

## 1 | INTRODUCTION

Alternative transcription start sites (TSSs) usage is a common phenomenon and more than 50% genes contain multiple TSSs, and it plays an important role in gene transcription regulation in mammals.[1–3] For example, a novel transcript of the anaplastic lymphoma kinase (ALK) initiates from a de novo alternative transcription initiation site in ALK intron 19, which leads to the expression of a novel ALK isoform and results in melanomas.[1] Similarly, alternative TSSs usage of tumor protein (TP) p53 family gene (p73) could result in full-length or shorter isoform. Elevated expression of the shorter isoform of p73 (deltaNp73), which inhibits apoptosis, has been associated with tumor progression and poor prognosis in several human cancers, including neuroblastoma, lung and ovarian carcinomas.[2] Therefore, profiling of genome-wide alternative TSSs is expected to facilitate a better understanding of transcriptome complexity.

Cap-Analysis Gene Expression (CAGE) had the high performance of the 5′ RNA-seq methods.[4] Based on the CAGE method, a series of 5′ end-capture techniques have been derived. For example, the nAnT-iCAGE,[5] the less biased method for genome-wide identification of TSSs, the nanocage,[6,7] an approach for with limited amounts of RNA, and the SLIC-CAGE, an super-low input carrier-CAGE approach.[8] As technology continues to evolve and integrate, the accuracy and comprehensiveness of single cell TSS identification continues to improve. For example, scCAT-seq,[9] a method for sequencing the 5′ end of a single cell to infer the location of the TSS and CamoTSS,[10] a computational method that effectively detects alternative TSS. During the past decades, single-cell sequencing methods have made rapid advances on profiling genomics at unprecedented resolution.[11] Traditionally, bulk cell sequencing methods can only reveal the average expression signal from an ensemble of cells assumed to be one homogeneous population/tissue status.[11] Therefore, signals from small cell numbers or low densities could hardly be detected. Using single-cell technologies, numerous novel transcripts/cell types thereby have been identified.[12] For examples, based on single-cell universal poly(A)-independent RNA sequencing (SUPeR-seq), 913 novel linear transcripts and 2891 circRNAs in mouse preimplantation embryos have been found.[12] Based on single-cell isoform RNA-Seq (ScISOr-Seq), 16,872 novel isoforms and 18,173 known isoforms in thousands of cerebellar cells have been detected.[13] Thus, single cell sequencing is able to detect cell heterogeneity and lowly expressed alternative TSSs which may be un-achievable by bulk cell technologies.

Here, to enable genome-wide alternative TSSs annotation and cell heterogeneity detection, we have developed single cell TSSs sequencing (scTSS-seq) method (Figure 1). In this method, we isolate single cells in Fluidigm C1 system, capture full-length cDNAs using SMARTer cDNA synthesis kit, and enrich 5′ end cDNA tags specifically with customized dual priming oligonucleotides (DPO)[14] in Nextera XT DNA Library Prep Kit. DPO contains a longer 5′-segment, a shorter 3′-segment, a poly(I) linker between these two segments.[14] This enables enrichment of 5′end cDNA tags precisely even under sub-optimal PCR conditions.[14] We then apply this method to naïve 1 h and IFN-$\beta$ pretreated 1 h mouse embryonic fibroblasts (MEFs) (Materials and Methods) (Figure 2). Furthermore, in conjunction with RNA-seq, we examine whether pervasive alternative TSSs contribute to transcription memory following IFN-$\beta$ re-stimulation. This study detects a larger proportion of genes containing alternative TSSs than empirical estimations, and finds that alternative TSSs are cell-type-specific and may be involved in transcriptional memory.

## 2 | MATERIALS AND METHODS

### 2.1 | Cell preparation

The preparation of MEFs was similar to the protocol of Kamada et al.[15] In brief, MEFs were separated from day 13.5 mouse embryos, and then cultured in Dulbecco's minimal essential medium (DMEM) with 10% fetal bovine serum (FBS) at 37°C with 5% $CO_2$. Later, they were treated with 100 units $mL^{-1}$ of murine recombinant IFN-$\beta$ (PBL Interferon Source) with two different ways for indicated periods (Figure 2). For naïve 0 h MEFs, MEFs were treated without IFN-$\beta$ for 6 h, washed, and left without IFN-$\beta$ for 24 h, and then harvested. For pretreated 0 h MEFs, MEFs were treated with IFN-$\beta$ for 6 h, washed, and then left without IFN-$\beta$ for 24 h, and then harvested. In this study, both naïve and pretreated MEFs were restimulated with IFN-$\beta$ for 1 h and then harvested for scTSS-seq, which were termed as naïve 1 h, and pretreated 1 h MEFs, respectively (Figure 2). For bulk cell RNA-seq, besides naïve 1 h and pretreated 1 h MEFs, naïve 0 h and pretreated 0 h MEFs were used.

### 2.2 | scTSS-seq experiment procedure

A total of 96 naïve 1 h and 96 pretreated 1 h MEFs were captured using the C1 Fluidigm system according to the manufacturer's instructions. Upon capture, reverse transcription and cDNA preamplification were performed in the C1 Fluidigm system using the SMARTer PCR cDNA Synthesis Kit (Clontech 634926) according to manufacturer's recommendations. Full-length cDNAs were harvested and diluted to a range of 0.2 ng $\mu l^{-1}$, and Nextera libraries were prepared using the Nextera DNA Sample Preparation Kit (Catalog No. FC-131-1096). Nextera P5 (Illumina) and custom-made P7 index primers (Table S1) were used to amplify the tagmented fragments. The P7 index primers were designed as DPO. The details of the scTSS-seq library construction procedure were given in Supplementary File 1. Libraries were pooled and single-end sequencing was performed on HiSeq 2000 (Illumina).

### 2.3 | scTSS data processing

Reads were subjected to adapter and quality trimming with Trim Galore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/; v0.4.4; options: default parameters). The trimmed reads

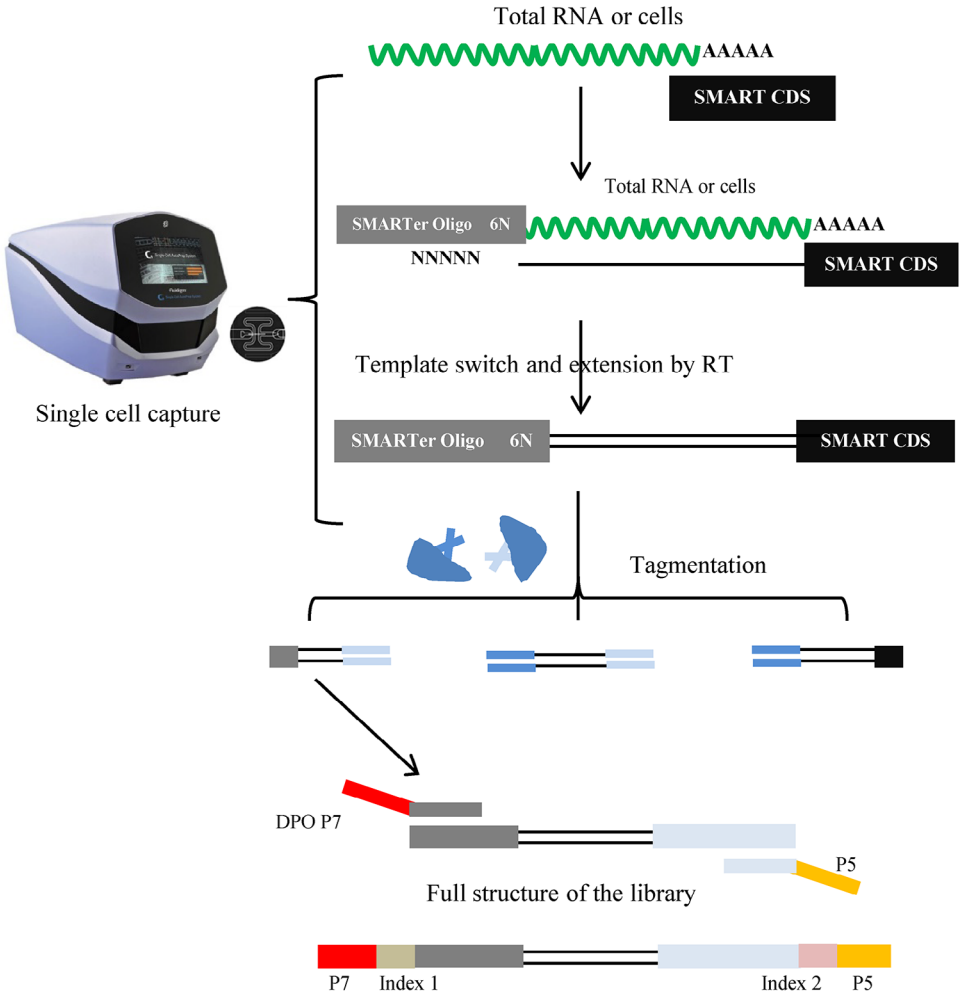


**FIGURE 1** Schematic of the scTSS-seq method. P5, Nextera XT illumina P5 primers; DPO P7, custom-made DPO P7 primers. DPO, dual priming oligonucleotides; scTSS-seq, single cell TSSs sequencing.

were then aligned to the mouse genome (mm9) using bowtie2[16] (v2.29, options: default parameters). HOMER[17] tag directory was created with "makeTagDirectory" function (options: default parameters) for each cell. For each tag directory, peaks were called with HOMER "findpeaks" function[17] (options: -style tss -ntagThreshold 200). For each cell, peaks were then annotated with HOMER "annotatePeaks.pl" function,[17] three different mm9 transcript models were used for annotations, including Refseq (HOMER default), UCSC (https://genome.ucsc.edu/cgi-bin/hgTables), and Genecode V1 (https://www.gencodegenes.org/mouse/release_M1.html). In these models, annotated promoters were defined as regions between downstream 1000 bp and upstream 100 bp of TSSs. For each cell, peaks overlapped with annotated promoters in any of three models were termed as known TSSs, while the remaining were novel TSSs (Figure 2). Furthermore, we downloaded mouse mm10 G-quadruplex (G4) data from EndoQuad (https://endoquad.chenzxlab.cn/). Since we utilized the mm9 version of the reference sequence, we employed liftover to convert the database to the mm9 version. Subsequently, we performed intersect analysis on our TSCs using bedtools (v2.17.0). We removed the novel TSCs overlapped with G4 database.

Low quality cells would have lost transcripts prior to cell lysis and their exon numbers would be larger in broken cells.[18] Thus, we used the ratios between known TSSs numbers and identified TSSs numbers in total for each cell (known TSS ratios) for filtering low quality cells. In this study, we used threshold as z-score of known TSSs ratios of 1 for removing cells. The remaining cells were considered as high quality cells and used for downstream analysis.

To produce the matrix of TSSs tags across cells, we merged TSSs when their 5′ end were located around ±2 bp across all high quality cells. For further analysis, the expression levels of TSSs were normalized to counts per million, with calculation by counts per mapped reads per 10 million. The expression levels of genes were calculated as the expression levels of TSSs tags within Refseq genes on average. t-Distributed Stochastic Neighbor Embedding (t-SNE) analysis and differential analysis for TSSs were performed by using R packages Seurat.[19] To know how many novel TSSs were overlapped in other TSSs datasets across difference cell types, we downloaded TSSs coordinates (bed format) from different cell types or tissues such as 10Thalf, 3T3, ATDC5, and embryos (ftp://ftp.hgc.jp/pub/hgc/db/dbtss/dbtss_ver8/mm9/TSSseq/bed/).

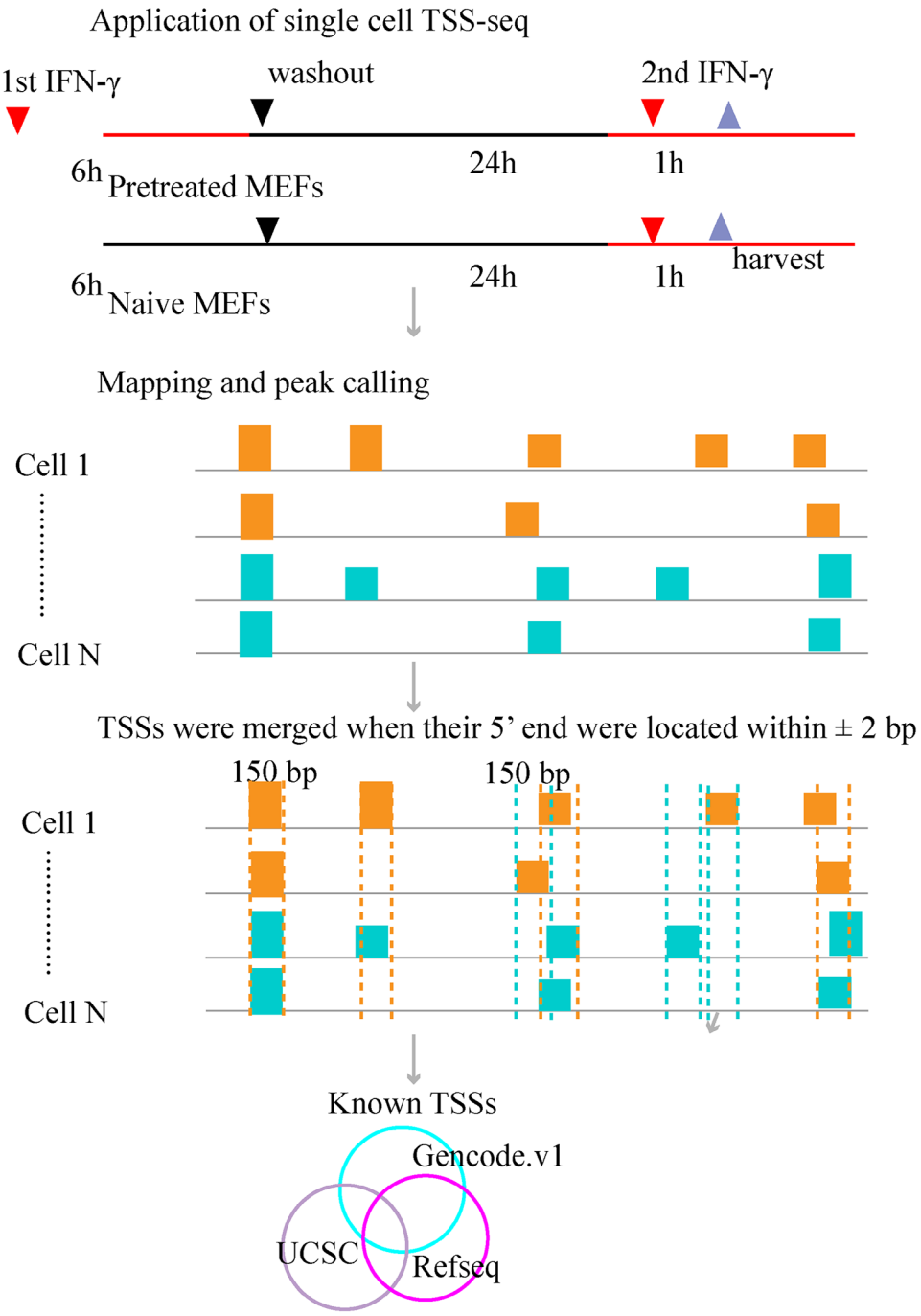


**FIGURE 2** scTSS-seq applications and TSSs identification schematic in naïve 1 h and pretreated 1 h MEFs. MEF, mouse embryonic fibroblast; scTSS-seq, single cell TSSs sequencing; TSS, transcription start site.

## 2.4 | RNA-seq procedure and analysis

The library construction of RNA-seq was similar to Kamada et al.[15] In contrast, single-end sequencing on a HiSeq2000 (Illumina) was used in our study. The obtained sequences were aligned to the mouse genome mm9 using Tophat[20] (v2.29, options: default parameters). Transcript abundance was quantified using Cufflinks[21] (v1.2.1, options: default parameters). Differential expressed genes (DEGs) are defined as the Fold change (FC) ≥ 2 of gene expression levels and the $p$ values ≤ 0.05 using the program Cuffdiff.[22] To determine IFN-stimulated genes (ISGs), we first identified genes showing significantly ($p < 0.01$, FC ≥ 2) higher transcript expression (RPKM) in treated cells following IFN treatment, that is, treated 1 h versus treated 0 h, which detected 472 upregulated genes; then, we identified genes showing significantly ($p < 0.01$, FC ≥ 2) higher transcript expression (RPKM) in untreated cells following IFN treatment, that is, untreated 1 h and untreated 0 h, which included 346 upregulated genes; while, 88 genes were defined as ISGs as their gene

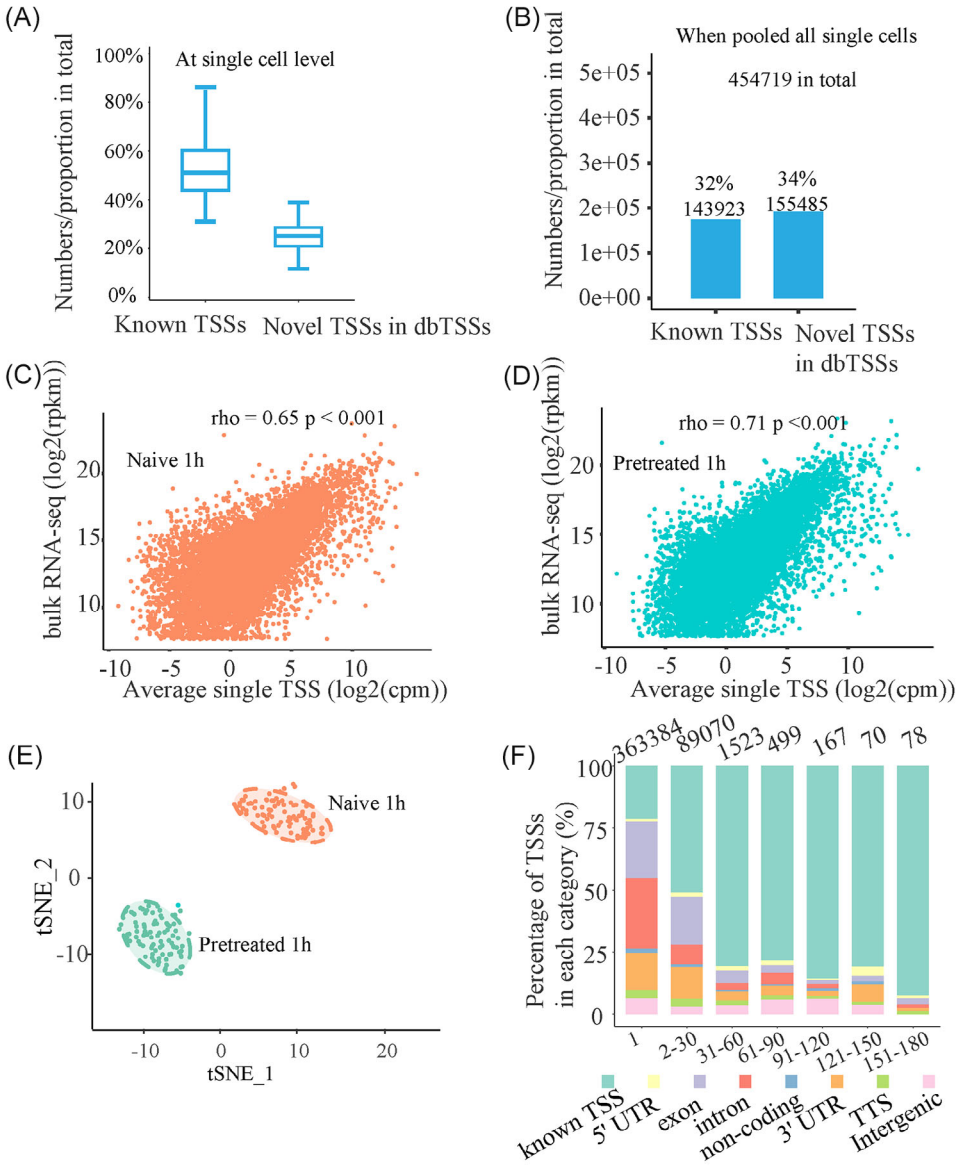


**FIGURE 3** The performance of scTSS-seq. (A, B) TSSs identification at single cell level and when all single cells were pooled together. (C, D) The Spearman correlation of gene expression profile based on scTSS-seq data and bulk RNA-seq data for naïve 1 h and pretreated 1 h MEFs. (E) t-SNE analysis based on TSSs expression levels. (F) Distribution of genomic annotation for TSSs. Note: 1, TSSs occurring in one cell; 2–30 TSSs occurring between 2 and 30 cells; 31–60, TSSs occurring between 31 and 60 cells and the remaining followed the similar rule to name. MEF, mouse embryonic fibroblast; scTSS-seq, single cell TSSs sequencing; t-SNE, t-distributed Stochastic Neighbor Embedding; TSS, transcription start site.

expression levels higher upregulated in treated cells than untreated cells (expression of treated upregulated genes/untreated upregulated genes > 1). We also identified 188 genes downregulated in treated cells and 871 genes upregulated in untreated cells during IFN treatment. A total of 24 was defined as non-ISGs as their gene expression levels higher downregulated in treated cells than untreated cells (expression of treated downregulated/untreated downregulated genes < 1). Gene Ontology (GO) analysis of different categories of ISGs were performed by using R packages ClusterProfiler,[23] respectively. Terms with $p < 0.05$ were regarded as significant enrichment.

## 3 | RESULTS

### 3.1 | The performance of scTSS-seq

In this study, we filtered the possible false positive TSSs based on known G4 database, and removed six naïve and three pretreated cells based on quality metrics (Section 2). At the single cell level, we detected 5002 TSSs in median and 2713 known TSSs (54.2% in total) per cell (Figure 3A). When all single cells were pooled together, our study detected 454,719 TSSs in total and 143,923 known TSSs (34% in total) (Figure 3B). Compared to C1 CAGE, an alternative TSS identification

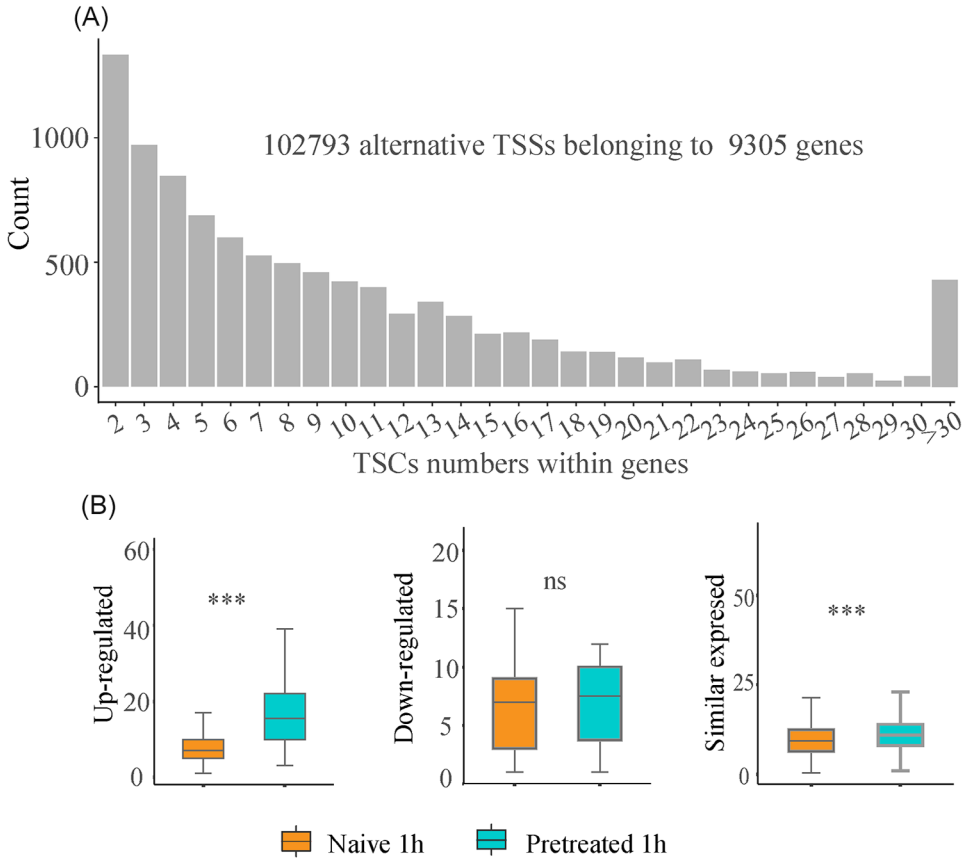


**FIGURE 4** Alternative TSSs usage at single cell resolution. (A) The distribution of alternative TSSs. (B) The comparison of TSSs numbers within genes between naïve 1 h and pretreated 1 h MEFs for unregulated, downregulated and similar expressed genes. Note: ***$p < 0.001$ the paired Wilcoxon signed-rank test are performed. MEF, mouse embryonic fibroblast; TSS, transcription start site.

method, which detected a median of 2788 CAGE clusters and 948 known TSSs (34% in total) per cell by applying C1 CAGE to 151 A514 cells following TGF-$\beta$ stimulation (40 cells for 0 h, 41 cells for 6 h, and 70 cells for 24 h).[26] Moreover, our results identified a median of 1150 novel TSSs (23% in total of TSSs) per cell (Figure 3A) and a total of 155,485 novel TSSs (34%) across all single cells overlapped with other TSSs in dbTSSs (Figure 3B), indicating a large part of TSSs that were not overlapped with annotated promoters were identified in other cell types. In addition, to verify the stability of the identified signals, we used a logistic regression-based approach to train the parameters peak score, focus ratio, distance to nearest RefSeq TSS, and cluster volume (Supplementary 2). All the results suggested that our method was efficient at determining TSSs.

Later, we investigated the correlation of gene expression between scTSS-seq and bulk RNA-seq dataset for each cell type. For scTSS-seq data, gene expression levels were calculated as the sum of genic TSSs tags within the genes (CPM). The expression profiles of scTSS-seq data and bulk RNA-seq data were well correlated (rho = 0.65 for naïve 1 h MEFs, rho = 0.71 for pretreated 1 h MEFs, $p < 0.001$, Spearman correlation, Figure 3C, D). This indicated that TSSs data were in good quality. To test whether scTSS-seq was able to differentiate cell types as well as other single-cell technologies, we employed the t-SNE method for nonlinear dimension reduction and data visualization method. The result has suggested that two prominent clusters consistent with the two well-predefined cell types were classified (Figure 3E), suggesting that scTSS-seq could detect cell heterogeneity.

Furthermore, we characterized the distribution of TSSs across 183 high quality MEFs. About 363,384 TSSs were presented in one cell (cell-unique TSSs) while 91,407 occurred in more than one cell (cell-shared TSSs) (Figure 3F). About 24% of cell-unique TSSs were known TSSs, while more than 62% cell-shared TSSs were known TSSs (Figure 3F). These results further supported that single cell technologies were able to detect numerous novel TSSs or transcripts.[12]

## 3.2 | Alternative TSSs detection at single cell level

To characterize the distribution of alternative TSSs and to compare with other bulk cell 5′end cDNA technologies, we first excluded the TSSs occurring in one cell and then counted alternative TSSs across naïve and pretreated MEFs. In this study, we found 91,407 TSSs occurring in more than two cells, of which 89,159 genic TSSs belonged to 11,557 Refseq genes. Among these genes, 9305 (81% in total) had more than one TSS, and most genes contained 2–10 TSSs (Figure 4A). With bulk CAGE, 11,264 (58% in total) genes had alternative TSSs in mammals.[24] With empirical determination according to the UCSC

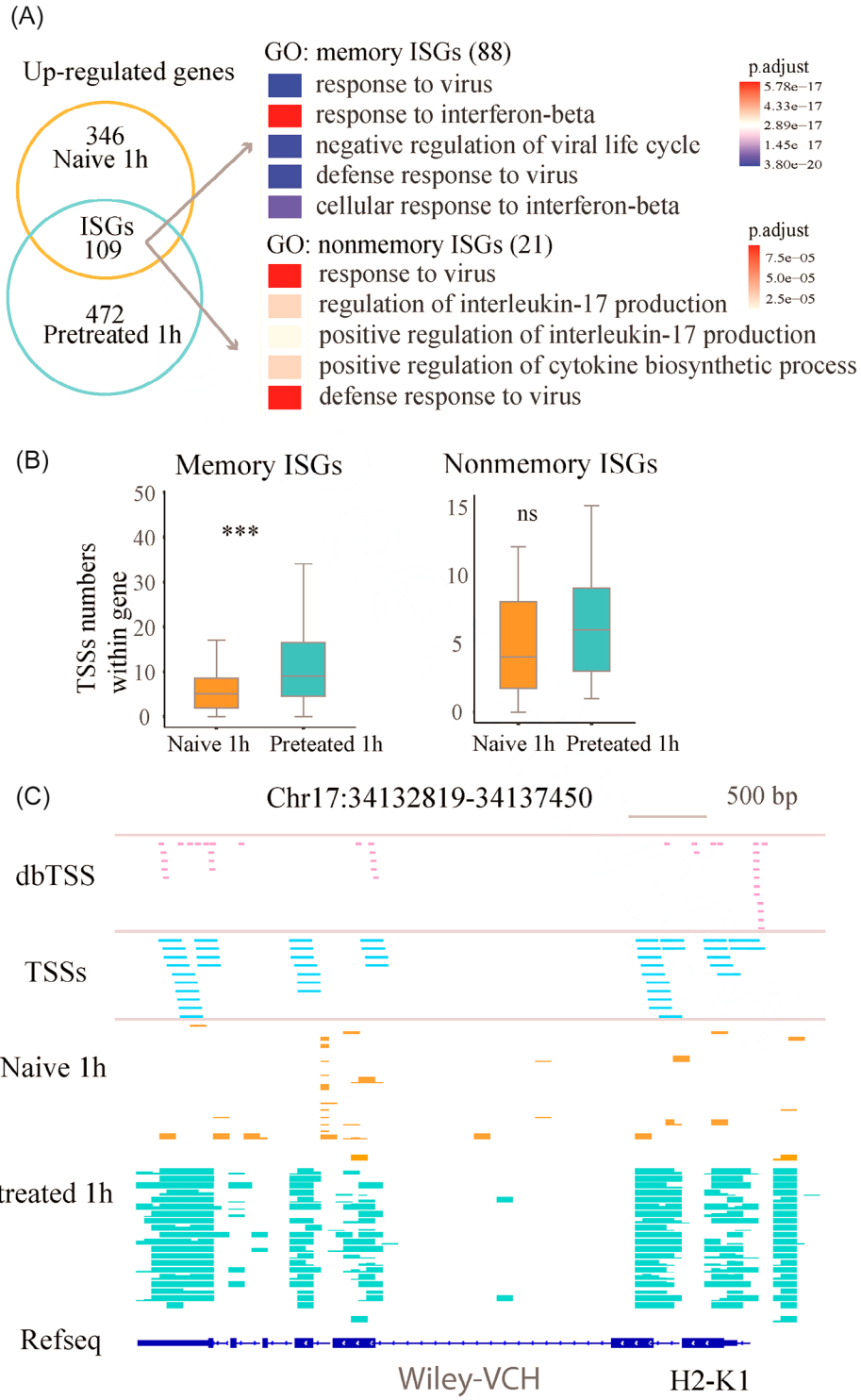


**FIGURE 5** Memory ISGs related-TSSs biological process. (A) The identification of memory and non-memory ISGs and GO analysis for the identified memory and nonmemory ISGs. (B) Comparison of alternative TSSs numbers per gene between naïve 1 h and pretreated 1 h MEFs, the paired Wilcoxon signed-rank test are performed, (C) example of TSSs distribution for ISGs, H2-K1. GO, Gene Ontology; ISG, IFN-stimulated gene; MEF, mouse embryonic fibroblast; TSS, transcription start site.

database, 54% of these genes exhibited alternative TSSs.[3] This suggested our method was capable of detecting alternative TSSs at genome-wide scale sensitively.

To explored whether the alternative TSSs numbers changes were correlated with their gene expression. Based on the matrix of the average TSSs expression level per gene using Seurat[25] and differential gene expression analysis of bulk RNA-seq, we detected 67 significant differentially expressed genes between pretreated 1 h and naïve 1 h following IFN stimulation, including 52 genes upregulated (average $\log_2$FC > 1, $p$ value adjust < 0.05) and 17 genes downregulated (average $\log_2$FC < −1, $p$ value adjust < 0.05) in pretreated 1 h compared to naïve 1 h cells (Figure 4B). The alternative TSSs numbers within upregulated genes were significantly larger in pretreated 1 h than in naïve 1 h MEFs ($p < 0.001$, the paired Wilcoxon signed-rank test with continuity correction). In contrast, the alternative TSSs numbers within downregulated genes were significantly smaller in pretreated 1 h compared to naïve 1 h MEFs ($p > 0.05$, the paired Wilcoxon signed-rank test with continuity correction). Later, we investigated the changes of alternative TSSs numbers for similar expressed genes between two cell types. A total of 1178 genes exhibited similar expression levels (abs($\log_2$(FC)) < 1, $p > 0.05$) between pretreated 1 h and naïve 1 h MEFs, The numbers of alternative TSSs within these genes were significantly larger in pretreated 1 h than naïve 1 h MEFs ($p < 0.001$, the paired Wilcoxon signed-rank test with continuity correction). These results indicated that alternative TSSs may be involved in transcription regulation using varying mechanisms, such as dependent or independent of changes in gene expression.

### 3.3 | Alternative TSSs usage is correlated with transcriptional memory

With IFN-$\beta$ restimulation in MEFs, it has been reported that the recruitment of RNA Pol II in memory ISGs is faster and greater than that in non-memory ISGs.[15] To test whether alternative TSSs usage affected transcription memory, we combined bulk RNA-seq data to analyze the changes of alternative TSSs numbers for ISGs across naïve and pretreated 1 h MEFs. Taking similar strategies with Kamada et al.,[15] we identified 112 upregulated ISGs based on bulk RNA-seq data. Of these, 88 (79% in total) ISGs mRNA levels were >1-fold higher in pretreated 1 h than naïve 1 h MEFs (>1 in RPKM), which was defined as memory ISGs; while 24 (21% in total) ISGs mRNA levels were ≤1-fold higher (≤1 in RPKM), which was defined as non-memory ISGs (Figure 5A). GO analyses showed that memory and non-memory ISGs shared related categories, such as innate immune and defense responses. The GO results were similar in naïve and pretreated MEFs restimulated by IFN-$\beta$ in previous study.[15] For these 88 memory ISGs, the alternative TSSs numbers within memory ISGs were significantly larger in pretreated 1 h than in naïve 1 h MEFs (Figure 5B) ($p < 0.001$, the paired Wilcoxon signed-rank test with continuity correction). For example, memory IGS H2-K1 in pretreated 1 h have more prevalent alternative TSSs usage than in naïve 1 h MEFs (Figure 5C). For 37 non-memory ISGs (Figure 5B), alternative TSSs numbers within non-memory ISGs showed no significant differences between pretreated 1 h and naïve 1 h MEFs ($p > 0.01$, the paired Wilcoxon test with continuity correction). These results supported that TSSs provided a platform for RNA Pol II binding, while the rapid recruitments of RNA Pol II binding were necessary for transcription memory.[15]

## 4 | DISCUSSION

In this study, we have developed scTSS-seq for genome-wide alternative TSSs detection at single cell level. This method utilizes Fluidigm C1 system, SMART technology, and DPO system to increase detection sensitivity of transcripts and genes, full-length rate of cDNAs, and enrichment specificity,[14,26,27] respectively. The results indicate scTSS-seq as accurate as C1 CAGE.[28] The previous study indicated the six 5′ RNA-seq methods included CAGE, RAMPAGE, STRT, NanoCAGE, oligo-capping, and GRO-cap, the study indicated that CAGE performed best overall by comparing RNA input, per-sample cost, the amount of time and number of steps per library for each method, and the rate of false positives.[4] C1-CAGE method allows 5′-end transcript profiling with strand information in a single cell.[28] In addition, this method is able to detect cell heterogeneity. Furthermore, higher rates (82%) of alternative TSSs are detected compared to predictions based on CAGE (58%) or empirical determination (54%) in various cells. Thus, scTSS-seq is an efficient method for positioning alternative TSSs at genome-wide scale.

At the same time, we have detected that memory ISGs have larger numbers of alternative TSSs in pretreated 1 h than in naïve 1 h MEFs, implying that alternative TSSs usage are cell-type-specific and may be contributed to transcriptional memory.[15] However, it still remains challenging to validate the role of alternative TSSs on memory ISGs in single cells so far. Firstly, cell viability is required when isolating single-cells for the purpose of production of monoclonal cell culture, which would be affected during single cell separation.[29] Secondly, the amount of protein and RNA of single cells may be not enough for histochemical experiments.[30] Thirdly, the quality and amount of antibodies would be limited for validation.[31] Despite these factors, we have developed a reliable method for genome-wide alternative TSSs detection and also dissect a more pervasive alternative TSSs phenomenon in cells. Meanwhile, we have provided useful alternative TSSs resources for future biological functional research.

### AUTHOR CONTRIBUTIONS

Yubo Zhang and Jun Zhu conceived the project, supervised, and designed the study. Yubo Zhang and Rui Kamada performed experiments. Keiko Ozato supervised the culture study. Yanling Peng and Qitong Huang carried out computational analyses. Yanling Peng, Danli Liu, Siyuan Kong, Yubo Zhang, and Jun Zhu wrote the manuscript.

### ACKNOWLEDGMENTS

This work was supported by National Key Research and Development Program of China [2018YFA0903201], National Natural

Science Foundation of China [2017M620977], the Science, Technology and Innovation Commission of Shenzhen Municipality [JCYJ20180306173714935], NIH Intramural research program at NICHD [ZIA HD001310-35], Funding for open access charge: National Key Research and Development Program of China, the National Natural Science Foundation of China (No.32202653), and the Project Funded by the China Postdoctoral Science Foundation (No. BX2021367 and 2021M703543).

## CONFLICT OF INTEREST STATEMENT

The authors declare no competing interests.

## DATA AVAILABILITY STATEMENT

All raw and processed sequencing data generated in this study have been submitted to the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) under accession number GSE174480.

## ORCID

*Danli Liu* https://orcid.org/0000-0003-3360-7493

## REFERENCES

1. Wiesner, T., Lee, W., Obenauf, A. C., Ran, L., Murali, R., Zhang, Q. F., Wong, E. W. P., Hu, W., Scott, S. N., Shah, R. H., Landa, I., Button, J., Lailler, N., Sboner, A., Gao, D., Murphy, D. A., Cao, Z., Shukla, S., Hollmann, T. J., … Chi, P. (2015). Alternative transcription initiation leads to expression of a novel ALK isoform in cancer. *Nature*, *526*, 453–457.
2. Murray-Zmijewski, F., Lane, D. P., & Bourdon, J. C. (2006). p53/p63/p73 isoforms: An orchestra of isoforms to harmonise cell differentiation and response to stress. *Cell Death and Differentiation*, *13*, 962–972.
3. Tyner, C., Barber, G. P., Casper, J., Clawson, H., Diekhans, M., Eisenhart, C., Fischer, C. M., Gibson, D., Gonzalez, J. N., Guruvadoo, L., Haeussler, M., Heitner, S., Hinrichs, A. S., Karolchik, D., Lee, B. T., Lee, C. M., Nejad, P., Villarrea, C., … Vivian, J. (2017). The UCSC Genome Browser database: 2017 Update. *Nucleic Acids Research*, *45*, D626–D634.
4. Adiconis, X., Haber, A. L., Simmons, S. K., Levy, A. L, Moonshine, A., Ji, Z., Busby, M. A., Shi, X., Jacques, J., Lancaster, M. A., Pan, J. Q., Regev, A., & Levin, J. Z. (2018). Comprehensive comparative analysis of 5′-end RNA-sequencing methods. *Nature Methods*, *15*, 505–511.
5. Murata, M., Nishiyori-Sueki, H., Kojima-Ishiyama, M., Carninci, P., Hayashizaki, Y., & Itoh, M. (2014). Detecting expressed genes using CAGE. *Methods in Molecular Biology*, *1164*, 67–85.
6. Salimullah, M., Mizuho, S., Plessy, C., & Carninci, P. (2011). NanoCAGE: A high-resolution technique to discover and interrogate cells transcriptome. *Cold Spring Harbor Protocols*, *2011*(1), pdb.prot5559.
7. Poulain, S., Kato, S., Arnaud, O., Morlighem, J. É., Suzuki, M., Plessy, C., & Harbers, M. (2017). NanoCAGE: A method for the analysis of coding and noncoding 5′-capped transcriptomes. *Promoter Associated RNA: Methods and Protocols*, *1543*, 57–109.
8. Cvetesic, N., Leitch, H. G., Borkowska, M., Müller, F., Carninci, P., Hajkova, P., & Lenhard, B. (2018). SLIC-CAGE: High-resolution transcription start site mapping using nanogram-levels of total RNA. *Genome Research*, *28*(12), 1943–1956.
9. Liu, L., Liu, C., Quintero, A., Wu, L., Yuan, Y., Wang, M., Cheng, M., Leng, L., Xu, L., Dong, G., Li, R., Liu, Y., Wei, X., Xu, J., Chen, X., Lu, H., Chen, D., Wang, Q., Zhou, Q., … Xu, X. (2019). Deconvolution of single-cell multi-omics layers reveals regulatory heterogeneity. *Nature Communications*, *10*(1), 470.
10. Hou, R., Hon, C. C., & Huang, Y. (2023). CamoTSS: Analysis of alternative transcription start sites for cellular phenotypes and regulatory patterns from 5′ scRNA-seq data. *Nature Communications*, *14*(1), 7240.
11. Tang, X. M., Huang, Y. M., Lei, J. L., Luo, H., & Zhu, X. (2019). The single-cell sequencing: New developments and medical applications. *Cell & Bioscience*, *9*, 53.
12. Fan, X. Y., Zhang, X. N., Wu, X. L., Guo, H. S., Hu, Y., Tang, F., & Huang, Y. (2015). Single-cell RNA-seq transcriptome analysis of linear and circular RNAs in mouse preimplantation embryos. *Genome Biology*, *16*, 148.
13. Gupta, I., Collier, P. G., Haase, B., Mahfouz, A., Joglekar, A., Floyd, T., Koopmans, F., Barres, B., Smit, A. B., Sloan, S. A., Luo, W., Fedrigo, O., Ross, M. E., & Tilgner, H. U. (2018). Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. *Nature Biotechnology*, *36*, 1197–1202.
14. Chun, J. Y., Kim, K. J., Hwang, I. T., Kim, Y. J., Lee, D.-H., Lee, I.-K., & Kim, J.-K. (2007). Dual priming oligonucleotide system for the multiplex detection of respiratory viruses and SNP genotyping of CYP2C19 gene. *Nucleic Acids Research*, *35*, e40.
15. Kamada, R., Yang, W., Zhang, Y., Patel, M. C., Yang, Y., Ouda, R., Dey, A., Wakabayashi, Y., Sakaguchi, K., Fujita, T., Tamura, T., Zhu, J., & Ozato, K. (2018). Interferon stimulation creates chromatin marks and establishes transcriptional memory. *Proceedings of the National Academy of Sciences of the United States of America*, *115*, E9162–E9171.
16. Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*, 357–359.
17. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., & Glass, C. K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular Cell*, *38*, 576–589.
18. Ilicic, T., Kim, J. K., Kolodziejczyk, A. A., Bagger, F. O., McCarthy, D. J., Marioni, J. C., & Teichmann, S. A. (2016). Classification of low quality cells from single-cell RNA-seq data. *Genome Biology*, *17*, 29.
19. Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., Hao, Y., Stoeckius, M., Smibert, P., & Satija, R. (2019). Comprehensive integration of single-cell data. *Cell*, *177*, 1888–1902.e21.
20. Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics*, *25*, 1105–1111.
21. Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M. J., Salzberg, S. L., Wold, B. J., & Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, *28*, 511–515.
22. Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M. J., Salzberg, S. L., Wold, B. J., & Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, *28*(5), 511–515.
23. Yu, G., Wang, L.-G., Han, Y., & He, Q.-Y. (2012). ClusterProfiler: An R Package for comparing biological themes among gene clusters. *Omics*, *16*, 284–287.
24. Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C. A. M., Taylor, M. S., Engström, P. G., Frith, M. C., Forrest, A. R. R., Alkema, W. B., Tan, S. L., Plessy, C., Kodzius, R., Ravasi, T., Kasukawa, T., Fukuda, S., Kanamori-Katayama, M., … Hayashizaki, Y. (2006). Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genetics*, *38*, 626–635.
25. Hao, Y., Stuart, T., Kowalski, M. H., Choudhary, S., Hoffman, P., Hartman, A., Srivastava, A., Molla, G., Madad, S., Fernandez-Granda, C., & Satija, R. (2023). Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nature Biotechnology*, *42*(2), 293–304.
26. Picelli, S., Björklund, Å. K., Faridani, O. R., Sagasser, S., Winberg, G., & Sandberg, R. (2013). Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature Methods*, *10*, 1096–1098.

27. Hashimshony, T., Senderovich, N., Avital, G., & Klochendler, A. (2016). CEL-Seq2: Sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biology*, *17*, 77.
28. Kouno, T., Moody, J., Kwon, A. T. J., Shibayama, Y., Kato, S., Huang, Y., Böttcher, M., Motakis, E., Mendez, M., Severin, J., Luginbühl, J., Abugessaisa, I., Hasegawa, A., Takizawa, S., Arakawa, T., Furuno, M., Ramalingam, N., & West, J. (2019). C1 CAGE detects transcription start sites and enhancer activity at single-cell resolution. *Nature Communications*, *10*, 360.
29. Gross, A., Schoendube, J., Zimmermann, S., Steeb, M., Zengerle, R., & Koltay, P. (2015). Technologies for single-cell isolation. *International Journal of Molecular Sciences*, *16*, 16897–16919.
30. Romanov, R. A., Zeisel, A., Bakker, J., Girach, F., Hellysaz, A., Tomer, R., Alpár, A., Mulder, J., Clotman, F., Keimpema, E., Hsueh, B., Crow, A. K., Martens, H., Schwindling, C., Calvigioni, D., Bains, J. S., Máté, Z., Szabó, G., … Yanagawa, Y. (2017). Molecular interrogation of hypothalamic organization reveals distinct dopamine neuronal subtypes. *Nature Neuroscience*, *20*, 176–188.
31. Taussig, M. J., Fonseca, C., & Trimmer, J. S. (2018). Antibody validation: A view from the mountains. *New Biotechnology*, *45*, 1–8.

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Peng, Y., Huang, Q., Liu, D., Kong, S., Kamada, R., Ozato, K., Zhang, Y., & Zhu, J. (2024). A single-cell genomic strategy for alternative transcript start sites identification. *Biotechnology Journal*, *19*, e2300516. https://doi.org/10.1002/biot.202300516