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

Yanling Peng, Qitong Huang, Rui Kamada, Keiko Ozato, Yubo Zhang, Jun Zhu

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. Here, 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- β stimulated mouse embryonic fibroblasts (MEFs). We quantify the performance of our method and find it as accurate as other single cell methods for the detection of TSSs. Our data are also clearly discriminate two IFN- β stimulated MEFs. Moreover, our results indicate 82% expressed genes in these two cell types containing multiple TSSs, which is much higher than previous predictions based on 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.

Main

Alternative transcription start sites (TSSs) usage is a common phenomenon and more than 50% genes contain multiple TSSs in human and mouse genomes ¹⁻⁴. It plays an important role in gene transcription regulation in mammals ^{5,6}. 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 ⁵. 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 ⁶. Therefore, profiling of genome-wide alternative TSSs is expected to facilitate a better understanding of transcriptome complexity.

During the past decades, single-cell sequencing methods have made rapid advances on profiling genomics, transcriptomics, and epigenomics at unprecedented resolution ⁷⁻⁹. 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 ¹⁰. 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 ^{11, 12}. For example, 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 ¹². Based on single-cell universal poly(A)-independent RNA sequencing (SUPeR-seq), 913 novel linear transcripts and 2,891 circRNAs in mouse preimplantation embryos have been found ¹¹. Furthermore, single-cell technologies are capable of dissecting heterogeneity in complex tissues ¹³⁻¹⁵. For example, single-cell methods enable classifying splenic cells into transcriptionally distinct groups ¹³ and identifying cell types of the central nervous system ¹⁶. Thus, single cell sequencing is able to detect cell heterogeneity and lowly expressed alternative TSSs which may be un-achievable by bulk cell technologies. To date, alternative TSSs annotation largely depends on computational predictions which have been available in Refseq ¹, UCSC ², and Genecode ³, as well as experimental identification by bulk cell technologies such as cap analysis of gene expression (CAGE) ¹⁷, oligo-capping ¹⁸, RNA annotation and mapping of promoters for the analysis of gene expression (RAMPAGE)

¹⁹. However, a paucity of single cell sequencing technologies are developed for alternative TSSs identification ^{20, 21}.

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 first 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) ²² 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 ²². This enables enrichment of 5'end cDNA tags precisely even under sub-optimal PCR conditions ²². We then apply this method to naïve 1h and IFN- β pretreated 1h 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- β 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.

Results

The performance of scTSS-seq

Based on quality metrics, we removed six naïve and three pretreated cells (Materials and Methods). In the median, we detected 5,775 TSSs and 2,712 known TSSs (47% in total) per cell (Figure 3a). While, by applying C1 CAGE to 151 A514 cells following TGF- β stimulation (40 cells for 0h, 41 cells for 6h, and 70 cells for 24 h), Kouno et al. ²⁰ detected a median of 2,788 CAGE clusters and 948 known TSSs (34% in total) per cell ²⁰. When all single cells were pooled together, our study detected 518,097 TSSs in total and 143,923 known TSSs (28% in total). In comparison, Carninci et al. ⁴ detected 159,075 CAGE clusters in 145 different mouse libraries and 28,443 (18% in total) known TSSs using bulk-cell CAGE. Moreover, our results identified a median of 1,693 TSSs (30% in total) per cell and a total of 192,208 TSSs (37%) 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. All the results suggested that our method was efficient at determining TSSs.

1 Later, we investigated the correlation of gene expression between scTSS-seq and bulk
2 RNA-seq dataset for each cell type. For scTSS-seq data, gene expression levels were
3 calculated as the sum of genic TSSs tags within the genes (cpm). The expression
4 profiles of scTSS-seq data and bulk RNA-seq data were well correlated ($\rho = 0.64$ for
5 naïve 1h MEFs, $\rho = 0.70$ for pretreated 1h MEFs, $p < 0.001$, Spearman correlation,
6 Figure 3c,d) compared to previous observations in other immune cells^{23 24}. This
7 indicated that TSSs data were in good quality. To test whether scTSS-seq was able to
8 differentiate cell types as well as other single-cell technologies, we employed the t-
9 distributed Stochastic Neighbor Embedding (t-SNE) method for clustering. Two
10 prominent clusters consistent with the two well-predefined cell types were classified
11 (Figure 3e), suggesting that scTSS-seq could detect cell heterogeneity.

12

13 Furthermore, we characterized the distribution of TSSs across 183 high quality MEFs.
14 About 408,743 TSSs were presented in one cell (cell-unique TSSs) while 109,354
15 occurred in more than one cell (cell-shared TSSs) (Figure 3f). About 25% of cell-unique
16 TSSs were known TSSs, while more than 50% cell-shared TSSs were known TSSs
17 (Figure 3f). These results further supported that single cell technologies were able to
18 detect numerous novel TSSs or transcripts^{11, 12}.

19

20 **Alternative TSSs detection at single cell level**

21 To characterize the distribution of alternative TSSs and to compare with other bulk cell
22 5'end cDNA technologies, we first excluded the TSSs occurring in one cell and then
23 counted alternative TSSs across naïve and pretreated MEFs. In this study, we found
24 109,354 TSSs occurring in more than two cells, of which 104,842 genic TSSs belonged
25 to 11,761 Refseq genes. Among these genes, 9,712 (82% in total) had more than one
26 TSS, and most genes contained 2~10 TSSs (Figure 4a). With bulk CAGE, 11,264 (58%
27 in total) genes had alternative TSSs in mammals⁴. High-density DNA microarrays had
28 identified 1,609 (24% in total) of expressed genes having alternative TSSs in human
29 fibroblasts cells²⁵. With empirical determination according to the UCSC database, 54%
30 of these genes exhibited alternative TSSs²⁶. This suggested our method was capable of
31 detecting alternative TSSs at genome-wide scale sensitively.

32

33 To investigate whether alternative TSSs contributed to transcription regulation, we first
34 assessed the alternative TSSs numbers for their significantly differentially expressed

1 genes between two well-defined cell types. Based on single cell differential gene
2 expression analysis, we detected 103 significant differentially expressed genes across
3 two cell types, including 30 up-regulated and 73 down-regulated genes. The up-
4 regulated genes included 264 alternative TSSs (Figure 4b). The TSSs numbers within
5 these genes were significantly larger in pretreated 1h than in naïve 1h MEFs ($p < 0.001$,
6 the paired wilcoxon signed-rank test with continuity correction). In contrast, down-
7 regulated genes included 386 alternative TSSs. The TSSs numbers within these genes
8 were significantly smaller in pretreated 1h than in naïve 1h MEFs ($p < 0.001$, the paired
9 wilcoxon signed-rank test with continuity correction). Later, we investigated the
10 changes of alternative TSSs numbers for similar expressed genes between two cell
11 types. A total of 1,516 genes exhibited similar expression levels (fold change < 1.2 , p
12 value > 0.05), that included 19,545 TSSs. The numbers of alternative TSSs within these
13 genes were significantly larger in pretreated 1h than naïve 1h MEFs ($p < 0.001$, paired
14 wilcoxon signed-rank test with continuity correction). These results indicated that
15 alternative TSSs may be involved in transcription regulation using varying mechanisms,
16 such as dependent or independent of changes in gene expression.

17

18 **Alternative TSSs usage is correlated with transcriptional memory**

19 With IFN- β re-stimulation in MEFs, it has been reported that the recruitment of RNA
20 Pol II in memory IFN-stimulated genes (ISGs) (Materials and Methods) is faster and
21 greater than that in non-memory ISGs²⁷. To test whether alternative TSSs usage
22 affected transcription memory, we combined bulk RNA-seq data to analyze the changes
23 of alternative TSSs numbers for ISGs across naïve and pretreated 1h MEFs. Taking
24 similar strategies with Kamada et al.²⁷, we identified 112 up-regulated ISGs based on
25 bulk RNA-seq data. These ISGs included 75 (67% in total) memory ISGs (*fold change* $>$
26 1.2 in pretreated 1h than naïve 1h MEFs) and 37 (33% in total) non-memory ISGs (*fold*
27 *change* ≤ 1.2 and *fold change* > -1.2 in pretreated 1h than naïve 1h MEFs) (Figure 5a).
28 GO analyses showed that memory and non-memory ISGs shared related categories,
29 such as innate immune and defense responses. The GO results were similar in naïve
30 and pretreated MEFs re-stimulated by IFN- β in previous study²⁷. Among these 75
31 memory ISGs, 910 alternative TSSs were detected (Figure 5b). The alternative TSSs
32 numbers within memory ISGs were significantly larger in pretreated 1h than in naïve
33 1h MEFs ($p < 0.001$, the paired wilcoxon signed-rank test with continuity correction).
34 For example, memory IGS *H2-K1* in pretreated 1h have more prevalent alternative

TSSs usage than in naïve 1h MEFs (Figure 5c). A total of 226 alternative TSSs were identified belonging to 37 non-memory ISGs (Figure 5b). Alternative TSSs numbers within non-memory ISGs showed no significant differences between pretreated 1h and naïve 1h MEFs ($p > 0.01$, the paired wilcoxon test with continuity correction). These results supported that TSSs provided platform for RNA Pol II binding, while the rapid recruitments of RNA Pol II binding were necessary for transcription memory^{27, 28}.

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 to increase detection sensitivity of transcripts and genes²⁹, SMART technology to increase full-length rate of cDNAs³⁰ and DPO system to increase enrichment specificity²². Meanwhile, the results show that the scTSS-seq method is efficient at determining TSSs compared to C1 CAGE²⁰. In addition, this method is in good performance by estimating the correlation between single cell TSSs and bulk RNA-seq expression profile. Furthermore, higher rates (82%) of alternative TSSs across naïve 1h and pretreated 1h MEFs are detected in comparison with previous 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 1h than in naïve 1h MEFs, implying that alternative TSSs usage are cell-type-specific and may be contributed to transcriptional memory²⁷. 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^{31, 32}, which would be affected during single cell separation^{32, 33}. Secondly, the amount of protein and RNA of single cells may be not enough for histochemical experiments^{34, 35}. Thirdly, the quality and amount of antibodies would be limited for validation^{36, 37}. Despite these factors, we develop a reliable method for genome-wide alternative TSSs detection and also dissect a more pervasive alternative TSSs phenomenon in cells. Meanwhile, we provide useful alternative TSSs resources for future biological functional research.

MATERIALS AND METHODS

Cell preparation

The preparation of MEFs was similar to Kamada et al ²⁷. 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% CO₂. Later, they were treated with 100 units/mL of murine recombinant IFN- β (PBL Interferon Source) with two different ways for indicated periods (Figure 2). On the one hand, MEFs were treated without IFN- β for 6h, washed, and left without IFN- β for 24h (naïve 0h MEFs). On the other hand, MEFs were treated with IFN- β for 6h, washed, and then left without IFN- β for 24h (pretreated 0h MEFs). In this study, both naïve and pretreated MEFs were re-stimulated with IFN- β for 1 hour (naïve 1h and pretreated 1h MEFs) and harvested for scTSS-seq (Figure 2). For bulk cell RNA-seq, besides naïve 1h and pretreated 1h MEFs, naïve 0h and pretreated 0h MEFs were harvested.

scTSS-seq experiment procedure

96 naïve 1h and 96 pretreated 1h 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/ul, and Nextera libraries were prepared using the Nextera DNA Sample Preparation Kit (Catalog No. FC-131-1096). A Nextera P5 (Illumina) and a custom-made P7 index primer (Supplementary Table S1) were used to amplify the tagmented fragments. The P7 index primers were designed as DPO systems. 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).

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 were then aligned to the mouse genome (mm9) using bowtie2 ³⁸ (v2.29, options: default parameters). HOMER ³⁹ tag directory was created with 'makeTagDirectory' function (options: default parameters) for each cell. For each tag directory, peaks were called with HOMER 'findpeaks' function ³⁹ (options: -style tss -ntagThreshold 200). For each cell, peaks were then annotated with HOMER

‘annotatePeaks.pl’ function ³⁹, 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.

Low quality cells would have lost transcripts prior to cell lysis and their exon numbers would be larger in broken cells ⁴⁰. 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 ≤ -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. TSSs clustering and differential analysis were performed by using R packages Seurat ⁴¹. 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/).

RNA-seq procedure and analysis

The library construction of RNA-seq was similar to Kamada et al. ²⁷. 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 ⁴². Transcript abundance was quantified using Cufflinks 1.2.1 ⁴³. Genes showing significantly (p value < 0.01) higher transcript expression (RPKM) in treated cells following IFN treatment (1h) over untreated cells (0h) were considered up-regulated genes, termed as IFN stimulated genes (ISGs) and were analyzed further. GO analysis of different

categories of ISGs were performed by using R packages ClusterProfiler⁴⁴, respectively.
Terms with $p < 0.05$ were regarded as significant enrichment.

Data availability

scTSS-seq and RNA-seq raw data has been submitted to the Gene Expression Omnibus (GEO) under accession GSE174480.

Supplementary

Table S1 Information of custom-made DPO P7 primers

File 1 The details of scTSS-seq library construction procedure

References

1. Pruitt, K.D., Tatusova, T. & Maglott, D.R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **35**, D61-D65 (2007).
2. Haeussler, M. et al. The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res* **47**, D853-D858 (2019).
3. Frankish, A. et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res* **47**, D766-D773 (2019).
4. Carninci, P. et al. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* **38**, 626-635 (2006).
5. Wiesner, T. et al. Alternative transcription initiation leads to expression of a novel ALK isoform in cancer. *Nature* **526**, 453-+ (2015).
6. Murray-Zmijewski, F., Lane, D.P. & Bourdon, J.C. p53/p63/p73 isoforms: an orchestra of isoforms to harmonise cell differentiation and response to stress. *Cell Death Differ* **13**, 962-972 (2006).
7. Wang, Y. & Navin, N.E. Advances and Applications of Single-Cell Sequencing Technologies. *Mol Cell* **58**, 598-609 (2015).
8. Zhang, X.Y. et al. Single-Cell Sequencing for Precise Cancer Research: Progress and Prospects. *Cancer Res* **76**, 1305-1312 (2016).
9. Gawad, C., Koh, W. & Quake, S.R. Single-cell genome sequencing: current state of the science. *Nat Rev Genet* **17**, 175-188 (2016).
10. Hwang, B., Lee, J.H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp Mol Med* **50** (2018).
11. Fan, X.Y. et al. Single-cell RNA-seq transcriptome analysis of linear and circular RNAs in mouse preimplantation embryos. *Genome Biol* **16** (2015).
12. Gupta, I. et al. Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. *Nat Biotechnol* **36**, 1197-+ (2018).
13. Jaitin, D.A. et al. Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types. *Science* **343**, 776-779 (2014).
14. Brennecke, P. et al. Single-cell transcriptome analysis reveals coordinated ectopic gene-expression patterns in medullary thymic epithelial cells. *Nat Immunol* **16**, 933-941 (2015).
15. Meredith, M., Zemmour, D., Mathis, D. & Benoist, C. Aire controls gene expression in the thymic epithelium with ordered stochasticity. *Nat Immunol* **16**, 942-949 (2015).
16. Ofengeim, D., Giagtzoglou, N., Huh, D., Zou, C.Y. & Yuan, J.Y. Single-Cell RNA Sequencing: Unraveling the Brain One Cell at a Time. *Trends Mol Med* **23**, 563-576 (2017).
17. Forrest, A.R.R. et al. A promoter-level mammalian expression atlas. *Nature* **507**, 462-470 (2014).
18. Yamashita, R. et al. Genome-wide characterization of transcriptional start sites in humans by integrative transcriptome analysis. *Genome Res* **21**, 775-789 (2011).

19. Batut, P., Dobin, A., Plessy, C., Carninci, P. & Gingeras, T.R. High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression. *Genome Res* **23**, 169-180 (2013).
20. Kouno, T. et al. C1 CAGE detects transcription start sites and enhancer activity at single-cell resolution. *Nat Commun* **10** (2019).
21. Islam, S. et al. Highly multiplexed and strand-specific single-cell RNA 5' end sequencing. *Nat Protoc* **7**, 813-828 (2012).
22. Chun, J.-Y. et al. Dual priming oligonucleotide system for the multiplex detection of respiratory viruses and SNP genotyping of CYP2C19 gene. *Nucleic Acids Res* **35**, e40-e40 (2007).
23. Gaublomme, J.T. et al. Single-Cell Genomics Unveils Critical Regulators of Th17 Cell Pathogenicity. *Cell* **163** (2015).
24. Shalek, A.K. et al. Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* **510**, 363-+ (2014).
25. Kim, T.H. et al. A high-resolution map of active promoters in the human genome. *Nature* **436**, 876-880 (2005).
26. Tyner, C. et al. The UCSC Genome Browser database: 2017 update. *Nucleic Acids Res* **45**, D626-D634 (2017).
27. Kamada, R. et al. Interferon stimulation creates chromatin marks and establishes transcriptional memory. *Proc Natl Acad Sci U S A* **115**, E9162-E9171 (2018).
28. Chen, X. et al. Structural insights into preinitiation complex assembly on core promoters. *Science* **372** (2021).
29. Hashimshony, T. et al. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol* **17** (2016).
30. Picelli, S. et al. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat Methods* **10**, 1096-1100 (2013).
31. Chen, Y.H. & Pruett-Miller, S.M. Improving single-cell cloning workflow for gene editing in human pluripotent stem cells. *Stem Cell Res* **31**, 186-192 (2018).
32. Gross, A. et al. Technologies for Single-Cell Isolation. *Int J Mol Sci* **16**, 16897-16919 (2015).
33. Tomlinson, M.J., Tomlinson, S., Yang, X.B. & Kirkham, J. Cell separation: Terminology and practical considerations. *J Tissue Eng* **4** (2013).
34. Ginsberg, S.D. & Che, S.L. Combined histochemical staining, RNA amplification, regional, and single cell cDNA analysis within the hippocampus. *Lab Invest* **84**, 952-962 (2004).
35. Romanov, R.A. et al. Molecular interrogation of hypothalamic organization reveals distinct dopamine neuronal subtypes. *Nat Neurosci* **20**, 176-188 (2017).
36. Chames, P., Van Regenmortel, M., Weiss, E. & Baty, D. Therapeutic antibodies: successes, limitations and hopes for the future. *Brit J Pharmacol* **157**, 220-233 (2009).
37. Taussig, M.J., Fonseca, C. & Trimmer, J.S. Antibody validation: a view from the mountains. *New Biotechnol* **45**, 1-8 (2018).
38. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-U354 (2012).
39. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576-589 (2010).
40. Illicic, T. et al. Classification of low quality cells from single-cell RNA-seq data. *Genome Biol* **17** (2016).
41. Stuart, T. et al. Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-+ (2019).
42. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-1111 (2009).
43. Trapnell, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**, 511-U174 (2010).
44. Yu, G., Wang, L.G., Han, Y. & He, Q.Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics : a journal of integrative biology* **16**, 284-287 (2012).

Acknowledgements

This work has been supported by National Key Research and Development Program of

China [2018YFA0903201], National Natural Science Foundation of China (grant no. 2017M620977), the Science, Technology and Innovation Commission of Shenzhen Municipality (grant no. JCYJ20180306173714935).

Author information

Affiliations

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

Yanling Peng, Qitong Huang, Yubo Zhang

DNA Sequencing and Genomics Core, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD 20892, USA

Jun Zhu

Animal Breeding and Genomics, Wageningen University & Research, Wageningen, 6708PB, Netherlands

Qitong Huang

Division of Developmental Biology, National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, MD 20892, USA

Rui Kamada, Keiko Ozato

Contributions

Y.Z. and J.Z. conceived the project, supervised, and designed the study. Y.Z., and R.K. performed experiments. K.O. supervised the culture study. Y.P. and Q.H. carried out computational analyses. Y.P., Y.Z. and J.Z. wrote the manuscript.

Corresponding authors

Correspondence to Jun Zhu, jz@mokobious.com or Yubo Zhang, ribon_001@163.com

Ethics declarations

Competing interests

1 All authors declare no competing interests.

2

3 **Figures**

4

5 Figure 1. Schematic of the scTSS-seq method. P5, Nextera XT illumina P5 primers;

6 DPO P7, custom-made dual priming oligonucleotides (DPO) P7 primers.

7

8

9

10

11

12

13

14

15

Figure 2. scTSS-seq applications and TSSs identification schematic in naïve 1h and pretreated 1h MEFs.

1

2

3 Figure 3. The performance of scTSS-seq. (a,b) TSSs identification at single cell level
4 and when all single cells were pooled together. (c,d) The Spearman correlation of gene
5 expression profile based on scTSS-seq data and bulk RNA-seq data for naïve 1h and
6 pretreated 1h MEFs. (e) tSNE analysis based on TSSs expression levels. (f) Distribution
7 of genomic annotation for TSSs. Note: 1, TSSs occurring in one cell; 2-30 TSSs
8 occurring between 2 and 30 cells; 31-60, TSSs occurring between 31 and 60 cells; 61-
9 90, TSSs occurring between 61 to 90 cells; 91-120, TSS occurring ranged from 91 to
10 120 cells; 121-150, TSSs occurring between 121 to 150 cells; 151-180, TSS occurring
11 between 151 to 180 cells.

12

13

14

15

16

17

18

19

20

21

22

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 1h and pretreated 1h MEFs for un-regulated, down-regulated and similar expressed genes. Note: ***, $p < 0.001$ the paired wilcoxon signed-rank test are performed.

1

2

3 Figure 5. Memory ISGs related-TSSs biological process. (a) The identification of
4 memory and non-memory ISGs and GO analysis for the identified memory and
5 nonmemory ISGs. (b) Comparison of alternative TSSs numbers per gene between naïve
6 1h and pretreated 1h MEFs, the paired wilcoxon signed-rank test are performed, (c)
7 Example of TSSs distribution for ISGs, *H2-K1*.

8

9