BMC Genomics

SOFTWARE Open Access

Millefy: visualizing cell-to-cell heterogeneity in read coverage of single-cell RNA sequencing datasets

Haruka Ozaki[1,2,3]*, Tetsutaro Hayashi[3], Mana Umeda[3] and Itoshi Nikaido[3,4]

Abstract

Background: Read coverage of RNA sequencing data reflects gene expression and RNA processing events. Single-cell RNA sequencing (scRNA-seq) methods, particularly "full-length" ones, provide read coverage of many individual cells and have the potential to reveal cellular heterogeneity in RNA transcription and processing. However, visualization tools suited to highlighting cell-to-cell heterogeneity in read coverage are still lacking.

Results: Here, we have developed Millefy, a tool for visualizing read coverage of scRNA-seq data in genomic contexts. Millefy is designed to show read coverage of all individual cells at once in genomic contexts and to highlight cell-to-cell heterogeneity in read coverage. By visualizing read coverage of all cells as a heat map and dynamically reordering cells based on diffusion maps, Millefy facilitates discovery of "local" region-specific, cell-to-cell heterogeneity in read coverage. We applied Millefy to scRNA-seq data sets of mouse embryonic stem cells and triple-negative breast cancers and showed variability of transcribed regions including antisense RNAs, 3′ UTR lengths, and enhancer RNA transcription.

Conclusions: Millefy simplifies the examination of cellular heterogeneity in RNA transcription and processing events using scRNA-seq data. Millefy is available as an R package (https://github.com/yuifu/millefy) and as a Docker image for use with Jupyter Notebook (https://hub.docker.com/r/yuifu/datascience-notebook-millefy).

Keywords: Single-cell RNA sequencing, Visualization, Read coverage

Background

Single-cell RNA sequencing (scRNA-seq) has been increasingly important in many areas, including developmental biology and cancer biology. In scRNA-seq data analyses, visualization is crucial for quality control (QC) as well as exploratory data analyses. For example, dimensionality reduction techniques such as principal component analysis (PCA) [1] or t-distributed stochastic neighbor embedding [2] are applied to a gene expression matrix to visualize individual cells as points in 2- or 3-dimensional space. Heat maps are also used to visualize gene expression matrices and highlight latent clusters of genes and cells. To date, many tools for visualizing gene expression matrices of scRNA-seq data have been proposed [3].

In contrast to visualization of gene expression matrices, visualization of read coverage, which is the distribution of mapped reads along genomic coordinates, helps reveal diverse aspects of RNA sequencing data and thus RNA biology and functional genomics. For example, read coverage reflects transcribed gene structures (e.g., exon-intron structures and transcript isoforms) [4], RNA processing events (e.g., normal and recursive splicing [5]), and transcription of intergenic and unannotated regions (e.g., enhancer RNAs [eRNAs]) [6]. Moreover, visual inspection of read coverage enables quality assessment of experimental methods (e.g., whether amplification is biased [7]) and bioinformatic methods (e.g., the accuracy of expression level estimation).

*Correspondence: haruka.ozaki@md.tsukuba.ac.jp
[1] Bioinformatics Laboratory, Faculty of Medicine, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, 305-8575, Ibaraki, Japan
[2] Center for Artificial Intelligence Research, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, 305-8577 Ibaraki, Japan
Full list of author information is available at the end of the article

Given that scRNA-seq has revealed cellular heterogeneity in gene [8] and splicing isoform expression [9, 10], visualization of read coverage of scRNA-seq data is expected to reveal cellular heterogeneity in read coverage, which can be interpreted as biological (e.g., transcription and RNA processing) and technical (e.g., amplification biases) heterogeneity. Read coverage is informative, especially for so-called "full-length" scRNA-seq methods such as Smart-seq2 [11] and RamDA-seq [12], compared with "3′-tag sequencing" scRNA-seq methods, which sequence only the 3′ ends of RNAs and cannot be used to extract rich information from read coverage [13, 14]. Despite their potential importance, however, tools specifically for the visualization of read coverage of scRNA-seq data are still lacking.

To explore cell-to-cell heterogeneity in read coverage, we propose several requirements of a tool for visualization of read coverage in scRNA-seq data (Table 1). First, the tool must be able to display read coverage of all individual cells in a scRNA-seq dataset at once. This is because scRNA-seq data consist of many (10^2–10^3) cells and frequently includes latent heterogeneity that is masked by the summation of expression across cells. Second, the tool must associate read coverage with genomic contexts, such as gene structures and epigenomic features, because read coverage data can be interpreted only when it is displayed simultaneously with their genomic contexts. Third, the tool must highlight the cell-to-cell heterogeneity of read coverage within focal regions. This is because there should be "local" region-specific cell-to-cell heterogeneity in read coverage at transcriptional (e.g., antisense RNAs and eRNAs) and post-transcriptional (e.g., alternative splicing) levels, and such heterogeneity is difficult to notice in advance by cell groupings defined according to global similarity among cells.

Genome browsers and heat maps are two major tools for read coverage visualization. However, they are insufficient for fulfilling the above requirements.

Genome browsers, such as IGV [15] and JBrowse [16], utilize "tracks" to display various types of biological data, including gene annotations, positions of regulatory elements, and read mapping of next-generation sequencing (NGS) data along genomic coordinates. By stacking tracks in genome browsers, read coverage can easily be compared with other features and be interpreted in genomic contexts like gene models and epigenomic signals, which helps to generate and validate biological hypotheses. However, existing genome browsers are not suited for the large numbers of samples (i.e., cells) in scRNA-seq experiments. Indeed, efforts to visualize read coverage of scRNA-seq data using genome browsers have been limited to displays of a few dozen cells without the need to scroll [17, 18]. Although IGV and JBrowse implement heat map representations of tracks to show many cells at once, they cannot dynamically reorder tracks to reveal local cell-to-cell heterogeneity in read coverage.

Table 1 Comparison of Millefy to other visualization tools

	Millefy	Genome browsers	Heatmaps
Visualize many cells at once	✓		✓
Associate read coverage with genomic contexts	✓	✓	
Highlight cell-to-cell heterogeneity in read coverage	✓		✓

Tools for heat maps combined with clustering algorithms have been used in the analysis of scRNA-seq data. Thus, heat maps can be used to visualize read coverage of all cells at once and reveal heterogeneity in read coverage. However, tools for generating heat maps are unsuited for visualizing read coverage of scRNA-seq data in genomic contexts, or they lack functionality to directly extract read coverage from standard NGS data formats.

Here, we have developed Millefy, which combines genome-browser-like visualization, heat maps, and dynamic reordering of single-cell read coverage and thus facilitates the examination of local heterogeneity within scRNA-seq data. Millefy extracts and organizes various types of useful information from read coverage of scRNA-seq data.

Implementation

Millefy visualizes read coverage from each individual cell as a heat map in which rows represent cells and columns represent genomic bins within a focal region. The heat map is aligned with tracks for gene annotations, genomic features, and bulk NGS data, enabling comparisons of single-cell read coverage with genomic contexts. To highlight latent cell-to-cell heterogeneity in read coverage, the heat map rows (i.e., cells) are automatically and dynamically reordered by ’local’ pseudo-time, which is calculated using diffusion maps [19], a nonlinear dimensionality reduction method. Specifically, diffusion maps are applied to matrices of single-cell read coverage, where rows are cells and columns are genomic bins, and the first diffusion component is used to dynamically reorder cells either in an“all cells” manner or in a "group-wise" manner when groupings of cells are provided by users. Alternatively, PCA can also be used to rearrange the order of cells.

Effective visualization requires iterative adjustment of various aspects of plots by visual inspection. Millefy supports iterative adjustment of plots by the `millefy_adjust()` function, which reuses the read coverage matrices of the last plot, enabling faster adjustment than simply replotting. For example, after viewing plots, users can easily adjust the maximal value of the color scale, which is essential in cases with exceptionally high read coverages in the focal region.

Millefy visualizations consist of five types of tracks (Fig. 1): (1) scRNA-seq tracks, which display scRNA-seq read coverage as a heat map with ordered cells, (2) mean scRNA-seq read coverage tracks, (3) bulk NGS data tracks, which display read coverage of other NGS data, (4) BED tracks, which display genomic intervals defined by BED files, and (5) gene annotation tracks. In scRNA-seq tracks and bulk NGS data tracks, read coverage is normalized by user-provided normalization factors to correct for differences in the number of mapped reads among samples. Using the above tracks, Millefy can simultaneously display read coverage of each cell and mean read coverage of cells in each user-defined cell group as well as align scRNA-seq data with genome annotation data and NGS data.

Millefy was implemented in R and can import scRNA-seq data without the need for format conversion. For scRNA-seq data, Millefy accepts BAM and BigWig formats, which are standard file formats for NGS data analysis. Millefy is dependent on the rtracklayer package [20]

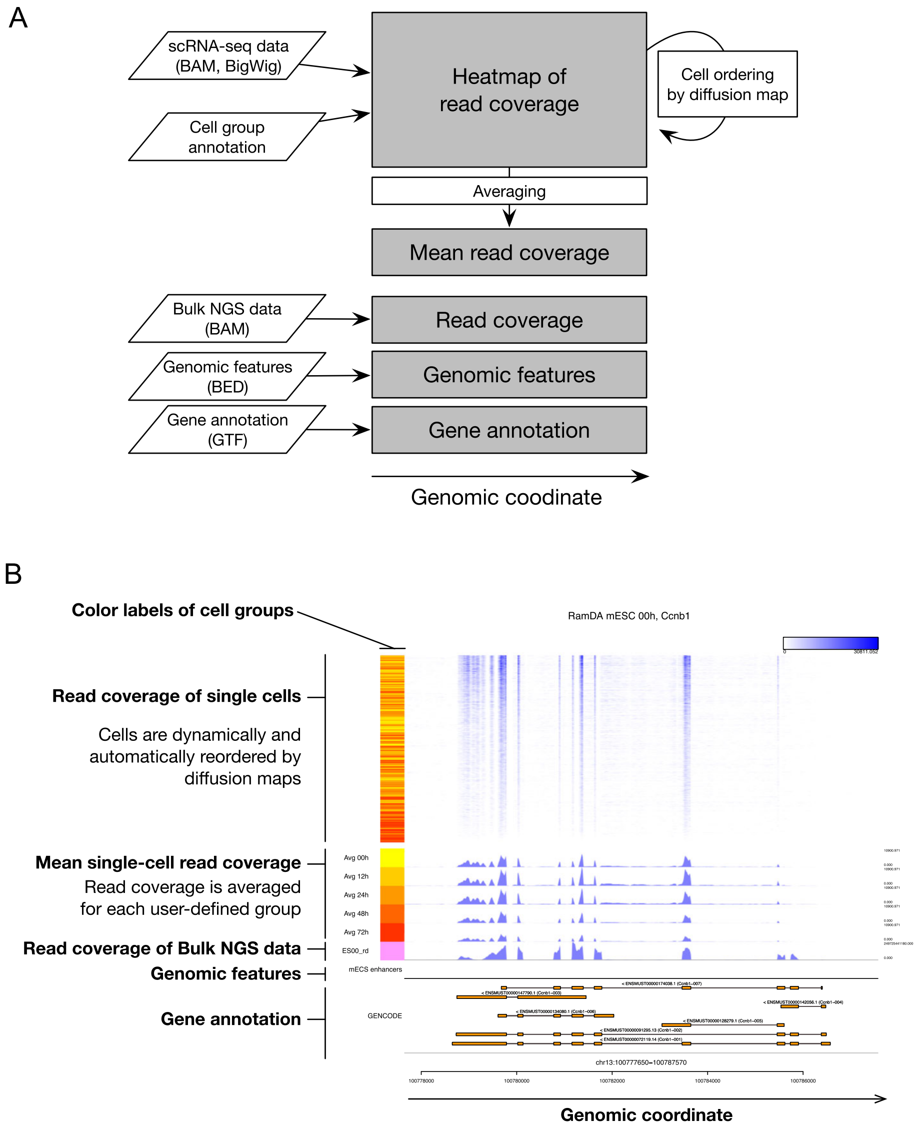
A
scRNA-seq data (BAM, BigWig)
Cell group annotation
Heatmap of read coverage
Cell ordering by diffusion map
Averaging
Mean read coverage
Bulk NGS data (BAM)
Read coverage
Genomic features (BED)
Genomic features
Gene annotation (GTF)
Gene annotation
Genomic coodinate
B
Color labels of cell groups
Read coverage of single cells
Cells are dynamically and automatically reordered by diffusion maps
Mean single-cell read coverage
Read coverage is averaged for each user-defined group
Read coverage of Bulk NGS data
Genomic features
Gene annotation
RamDA mESC 00h, Ccnb1
0
30811.052
Avg 00h
Avg 12h
Avg 24h
Avg 48h
Avg 72h
ES00_rd
10900.971
0.000
249725441180.000
mECS enhancers
GENCODE
< ENSMUST00000174038.1 (Ccnb1–007)
< ENSMUST00000147790.1 (Ccnb1–003)
< ENSMUST00000142056.1 (Ccnb1–004)
< ENSMUST00000134080.1 (Ccnb1–006)
< ENSMUST00000128279.1 (Ccnb1–005)
< ENSMUST00000091295.13 (Ccnb1–002)
< ENSMUST00000072119.14 (Ccnb1–001)
chr13:100777650–100787570
100778000
100780000
100782000
100784000
100786000
Genomic coordinate

Fig. 1 Overview of Millefy. **a** Millefy imports scRNA-seq data and visualizes read coverage of individual cells as a heat map. The rows (i.e., cells) of the heat map are dynamically reordered by diffusion maps. This automatic reordering highlights cell-to-cell heterogeneity in read coverage, which is hidden by mean read coverage data. Millefy associates genomic contexts, including bulk NGS data, genomic features, and gene annotations, thus facilitating the interpretation of single-cell read coverage. The parallelograms represent different types of input data. Gray boxes represent displayed tracks in Millefy. White boxes represent computation of a read coverage matrix. **b** An example plot in Millefy

and Rsamtools package [21] for importing BAM and BigWig files, respectively. For gene annotation data and genomic features, Millefy accepts GTF and BED formats, respectively. The data.table package [22] is used to import GTF and BED files. For performing diffusion maps on read coverage data, Millefy utilizes the destiny package [23].

We provide Millefy as an R package and as a Docker image based on the Jupyter Notebook Data Science stack (https://github.com/jupyter/docker-stacks) for use with Jupyter Notebook.

Results

Millefy highlights cellular heterogeneity in gene expression and transcribed gene structures

Researchers often merge read alignment files of single cells and visualize “synthetic bulk” data using standard genome browsers. However, in such cases, the merged (or averaged) read coverage cannot capture heterogeneity in read coverage. For example, a change in the merged read coverage cannot indicate whether the number of cells expressing a gene increased or the expression level of that gene increased across all cells. In contrast, Millefy visualizes read coverage of all individual cells in a scRNA-seq dataset as a heat map and thereby provides detailed information on cellular heterogeneity in read coverage.

To demonstrate the usefulness of Millefy’s ability to visualize read coverage in scRNA-seq data, we used a time-course RamDA-seq dataset derived from mouse embryonic stem cells (mESCs) upon induction of cell differentiation to primitive endoderm cells (at 0, 12, 24, 48, and 72 h) [12]. The dataset consists of 421 single cells.

Figure 2 shows the read coverage at *Sox17*, a differentiation marker gene. Cells were reordered according to the first diffusion component values calculated by a diffusion map of read coverage data for the locus, either within user-defined cell groups (Fig. 2a) or across all cells (Fig. 2b). While the height of the mean read coverage increased along the differentiation time course, the reordered heat map highlights the heterogeneity of read coverage among cells from the same time points (e.g., the 12 h group) (Fig. 2). Specifically, Millefy showed that the number of cells with *Sox17* expression increased, indicating asynchronous cell differentiation progression among cells.

Another example is *Zmynd8*, a transcriptional repressor. Figure 3 shows read coverage of 421 individual cells at the *Zmynd8* locus. The cells were dynamically reordered using diffusion maps based on the read coverage in the focal region. Expression of the *Zmynd8* short isoform is known to be associated with the expression of its antisense RNA *Zmynd8as* [24]. While *Zmynd8as* is unannotated in the current gene annotation, the heat map by Millefy clearly showed differential regulation of the long isoform of *Zmynd8* and *Zmynd8as*, facilitating visual inspection of the two separated transcription units (Figure 3). We note that the averaged read coverage for each time point cannot distinguish whether the long and short isoforms of *Zmynd8* and *Zmynd8as* are correlated or uncorrelated. These results demonstrate that Millefy’s functionality for displaying read coverage as a reordered heat map reveals cell-to-cell heterogeneity at the focal locus.

Millefy application on scRNA-seq data from triple-negative breast cancer patients

We also applied Millefy to a scRNA-seq dataset from triple-negative invasive cancer (TNBC) patients: the Smart-seq2 data of single sorted cells from six tumors collected from six women with primary, non-metastatic triple-negative invasive ductal carcinomas [25]. The dataset consists of B-cells (n=19), endothelial cells (n=14), epithelial cells (n=868), macrophages (n=64), stromal cells (n=94), and T-cells (n=53), according to the cell type annotation published by the authors. Epithelial cells were further classified into five clusters by the authors [25]: Clusters 1 (n=22), 2 (n=398), 3 (n=231), 4 (n=170), and 5 (n=47).

We checked whether 3′ UTR shortening is observed in the endothelial cell clusters. The Neuroblastoma RAS viral (v-Ras) oncogene homolog (NRAS) and the Jun proto-oncogene (c-JUN) are known to show alternative polyadenylation (APA)-dependent 3′ UTR shortening in TNBC cells [26]. Millefy appears to show that there is cell-to-cell heterogeneity in the length of the 3′ UTRs of c-JUN and NRAS (Fig. 4). Specifically, for c-JUN, some cells showed short read coverage and others showed long read coverage (Fig. 4a). In the last exon of NRAS, many cells showed long 3′ UTR read coverage but some cells showed a shortened 3′ UTR read coverage (Fig. 4b). Such heterogeneity cannot be determined by the aggregated (averaged) read coverage alone (Fig. 4). The triple-negative breast tumors with the 3′ UTR shortening of c-JUN and/or NRAS are reported to be smaller and less proliferative but more invasive than those without the 3′ UTR shortening [26]. Therefore, such subpopulations may confer heterogeneous invasiveness even when the population size is small; thus, further investigation is warranted.

Millefy associates read coverage with genomic contexts to facilitate interpretation of read coverage

Genomic contexts are crucial for interpretation of read coverage in bulk and single-cell RNA sequencing data. For example, read coverage overlapped with gene annotations can confirm known and reveal novel exon-intron structures. Moreover, read coverage overlapped with enhancer annotations can be interpreted as eRNA expression [6]. Using Millefy, single-cell read coverage can be compared

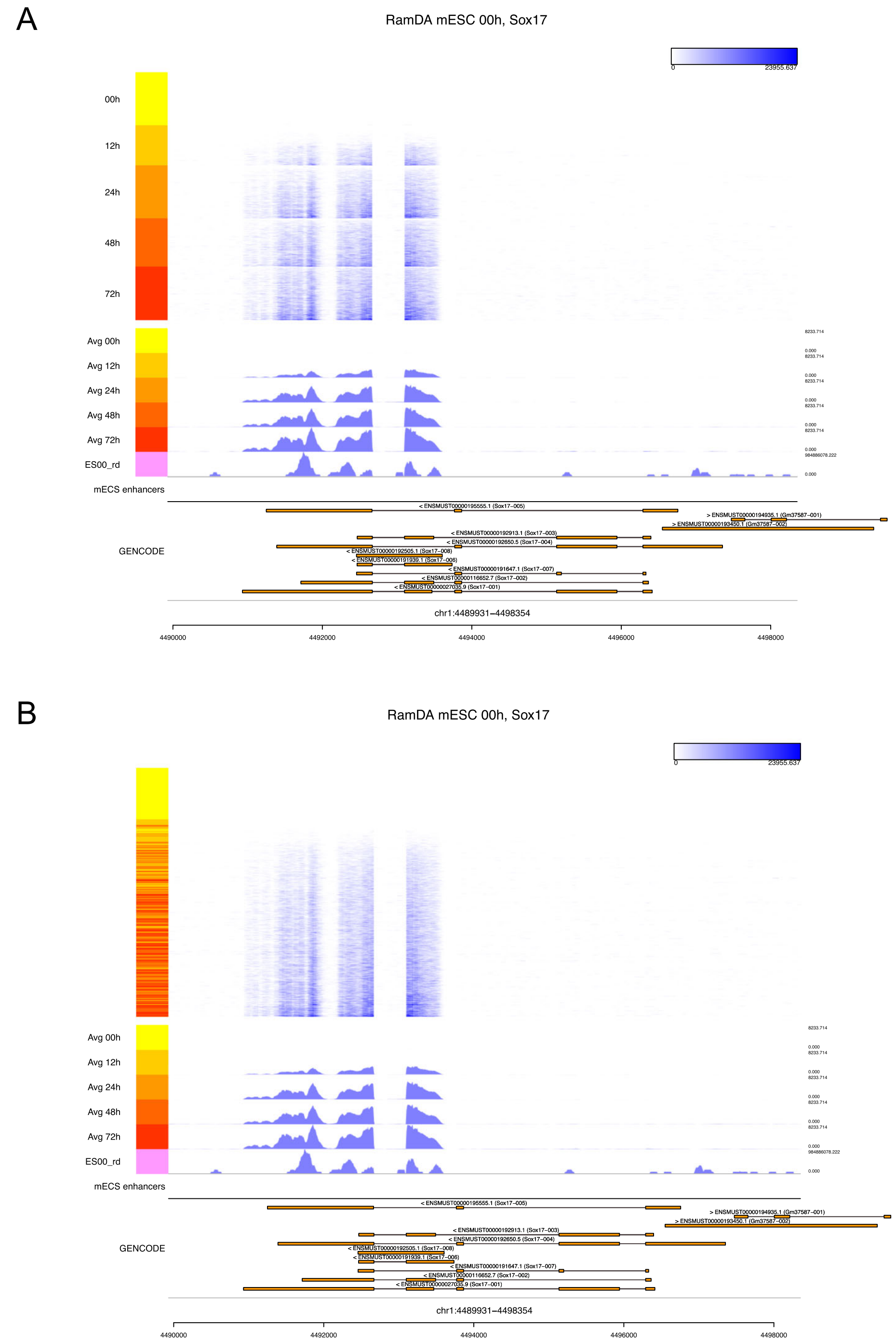

Fig. 2 Millefy visualization of read coverage at the *Sox17* locus. Millefy was applied to RamDA-seq data from mouse embryonic stem cells (mESCs) upon induction of cell differentiation to primitive endoderm cells. The top heat map shows single-cell read coverage. Color keys on the left side represent cells from different time points. The middle tracks show the averaged read coverage at different time points, the bulk RNA sequencing read coverage, and enhancer annotations. The bottom track shows the GENCODE reference gene annotation. Cells were reordered within **a** user-specified groups and **b** across all cells

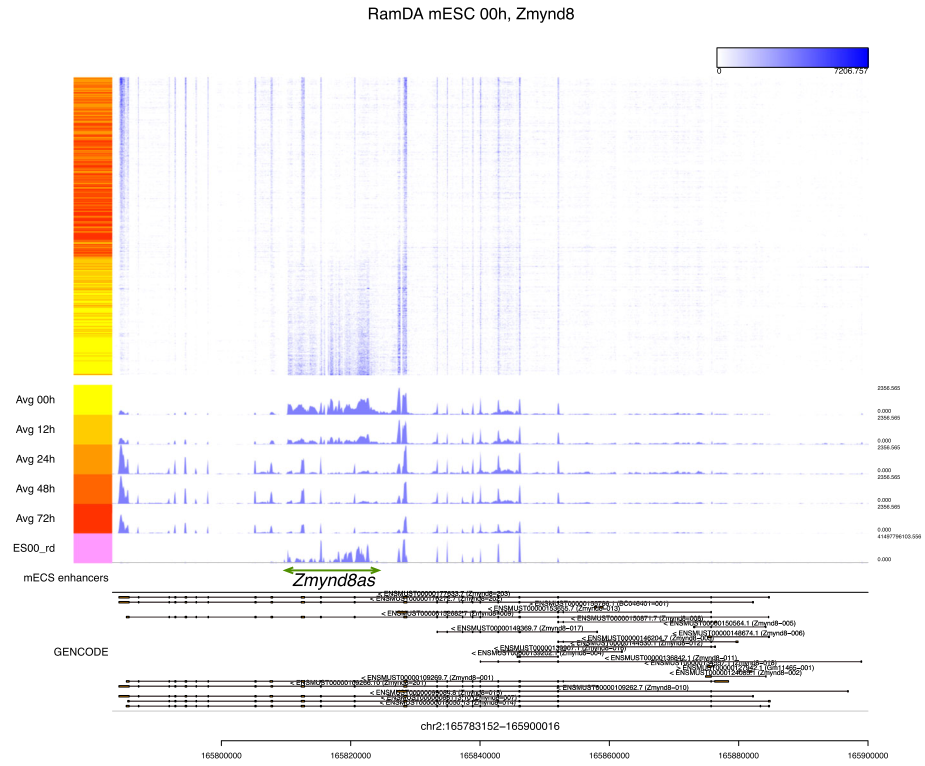

Fig. 3 Visualization of the *Zmynd8* locus by Millefy across all cells. Millefy was applied to RamDA-seq data from mouse embryonic stem cells (mESCs) upon induction of cell differentiation to primitive endoderm cells. All cells were reordered together without specified groupings. The green arrow has been added to indicate the position of *Zmynd8as*

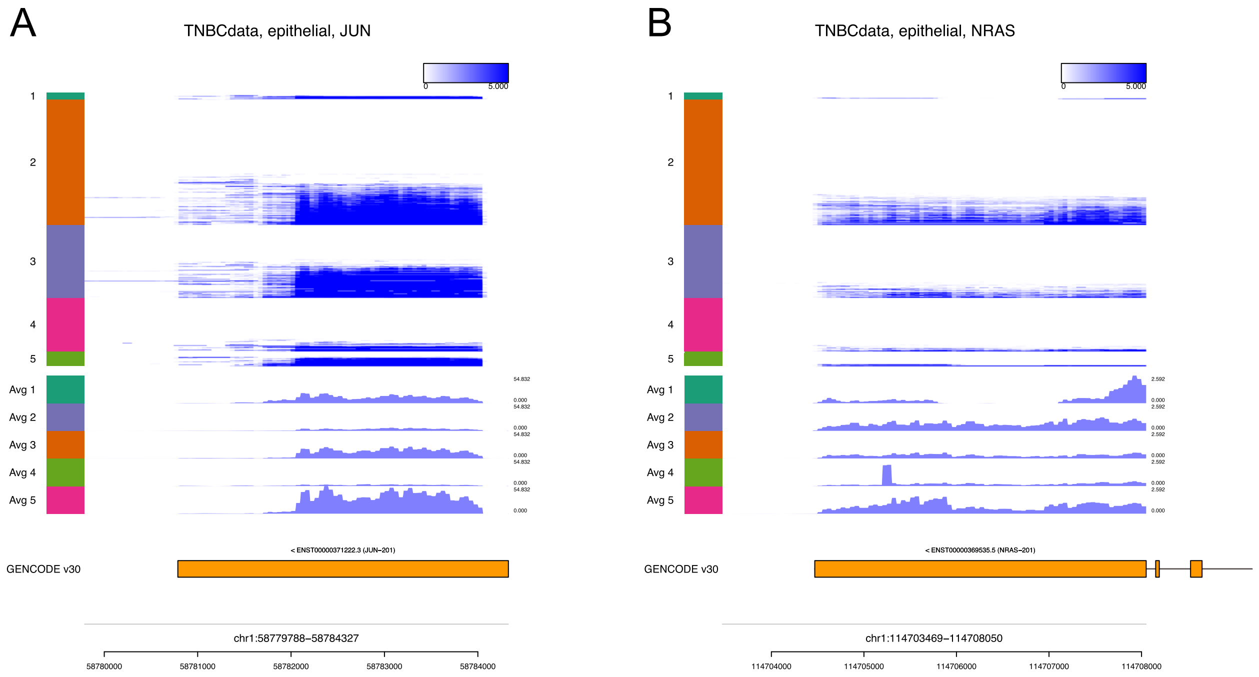

Fig. 4 3′ UTR shortening in *c-JUN* and *NRAS* in TNBC Smart-seq2 data. Millefy was applied to the Smart-seq2 data of epithelial cells from six women with primary, non-metastatic triple-negative invasive ductal carcinomas. Cells were reordered and visualized according to the cell cluster annotations

with genomic and epigenomic features like enhancer elements.

To demonstrate the usefulness of the simultaneous visualization of single-cell read coverage and genomic contexts, we compared read coverage of the RamDA-seq data from mESCs (0 h) with mESC enhancer regions. We downloaded H3K4me1 and H3K4me3 ChIP-seq peak regions for mESCs from the ENCODE project [27] and defined mESC enhancers as the H3K4me1 peaks that (1) did not overlap with the H3K4me3 peaks, (2) were at least 2 kbp away from the transcriptional start sites, and (3) were not included in the gene bodies of the GENCODE gene annotation (vM9) [28].

Figure 5 displays read coverage at the *Myc* locus, with the positions of enhancers active in mESCs. The *Myc* gene models and read coverage reveal that *Myc* was transcribed in mESCs. In addition, Millefy showed that some of the intergenic regions with transcribed RNA overlapped with the *Myc* downstream enhancer regions (Fig. 5). This is consistent with the previous report that RamDA-seq can detect eRNAs [12]. This result exemplifies how Millefy can help to interpret read coverage of scRNA-seq data in genomic contexts.

Millefy facilitates quality control in full-length scRNA-seq methods

Millefy can also be used for QC in full-length scRNA-seq methods. For example, scRNA-seq read coverage of long transcripts indicates whether the method employed provided full-length transcript coverage. Full-length transcript coverage provides accurate information about isoform expression and gene structures and is a fundamental feature of full-length scRNA-seq methods [29].

We applied Millefy to C1-RamDA-seq data ($n = 96$) and C1-SMART-Seq V4 ($n = 95$) data from a dilution of 10 pg of mESC RNA. Figure 6 shows the read coverage at *Mdn1*, a gene with a long transcript (17,970 bp) consisting of 102 exons. C1-RamDA-seq detected all exons in most samples, while C1-SMART-seq V4 failed to detect a fraction of known exons. Interestingly, the patterns of missing exons in C1-SMART-seq V4 seemed to vary among the samples. The lower reproducibility in read coverage of C1-SMART-seq V4 relative to C1-RamDA-seq is likely owing to technical noise because the samples were prepared not from living cells but from a dilution of 10 pg of RNA. We note that mean read coverage cannot provide such detailed information on reproducibility in read coverage

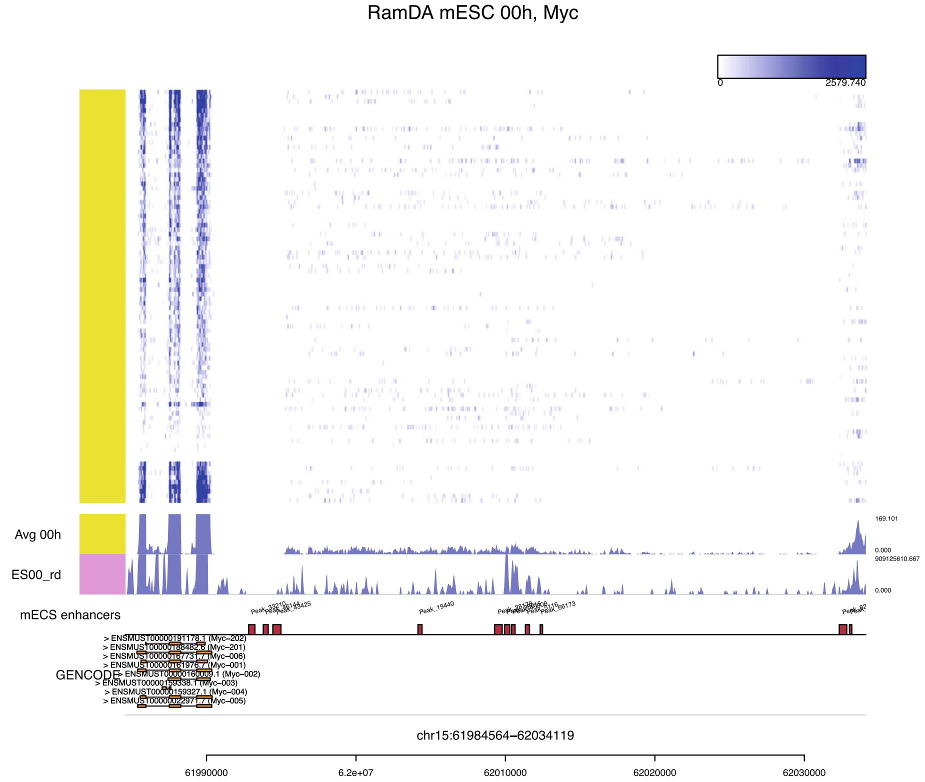

Fig. 5 Millefy visualization of read coverage and enhancer regions around the *Myc* locus. Millefy was applied to RamDA-seq data from mouse embryonic stem cells (mESCs). All cells were reordered together without specified groupings. RNA transcription was observed for the enhancers on the left

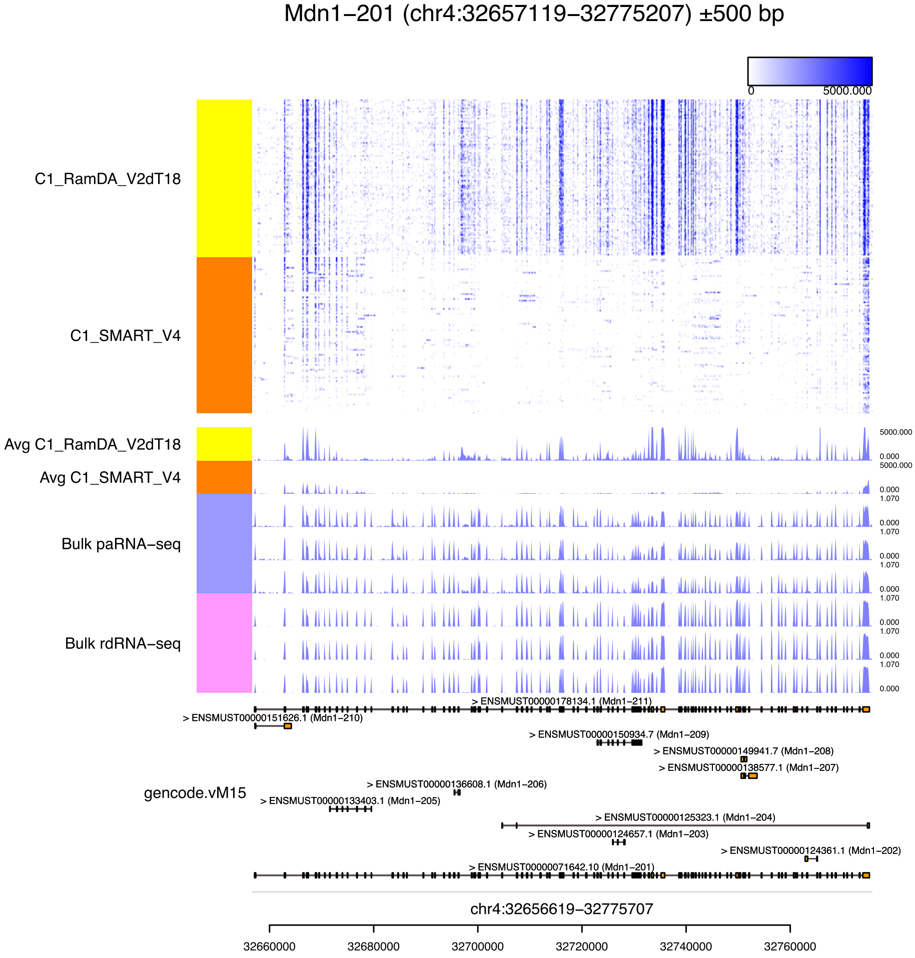

Fig. 6 Example of quality control of scRNA-seq methods. Visualization of read coverage from C1-RamDA-seq ($n = 95$) and C1-SMART-Seq V4 ($n = 96$) data from a dilution of 10 pg of RNA at the *Mdn1* locus by Millefy. The samples were reordered within user-specified groups

(Fig. 6). This result exemplifies how Millefy can be used for quality control of scRNA-seq methods.

Computational time

We measured the computational time of Millefy for visualizing whole gene bodies using RamDA-seq data with 793 samples from mESCs [12]. For 1000 randomly selected gene loci (of expressed genes with average TPM>5), Millefy processed BigWig and BAM files in 39.3 and 138.9 s, respectively, on average.

Discussion

In this paper, we proposed Millefy, a tool for visualizing cell-to-cell heterogeneity in read coverage in scRNA-seq data. Millefy combines genome-browser-like visualization, heat maps, and dynamic reordering of single-cell read coverage. Thereby, Millefy can display read coverage of all cells at once, associate read coverage with genomic contexts, and highlight the cell-to-cell heterogeneity of read coverage within focal regions (Table 1).

Using scRNA-seq data of mESCs and TNBC, we demonstrate the effectiveness of Millefy to reveal local heterogeneity in read coverage within scRNA-seq data. First, Millefy showed cellular heterogeneity in gene expression and transcribed gene structures through the cell differentiation time course of mESCs (Figs. 2 and 3). Second, we found cellular heterogeneity in the 3′ UTR shortening of c-JUN and/or NRAS in TNBC data (Fig. 4), which was not mentioned in the original paper [25]. Third, by associating read coverage with enhancer annotations, Millefy helped to interpret RNA transcription events in non-coding regions (Fig. 5). Collectively, these results indicate that Millefy enables the exploration of cellular heterogeneity of various biological events from read coverage of scRNA-seq data, which could be missed by conventional visualization tools.

Quality control is essential for developing new experimental and computational tools. Using scRNA-seq data of diluted RNA with different full-length scRNA-seq methods (Fig. 6), we demonstrate that Millefy visualizes read coverage in scRNA-seq as a QC measure and complements existing scRNA-seq QC pipelines based primarily on gene expression matrices [3]. In the development of bioinformatics methods using rule-based and machine learning approaches for profiling alternative splicing or novel RNAs by scRNA-seq, read coverage visualization tools like Millefy will become more important for evaluating and representing the predictions of algorithms. Indeed, Millefy was recently used for evaluating and visualizing the results of software developed to discover differentially expressed (DE) gene regions [30].

Conclusions

Millefy, which is integrated with Jupyter Notebook and provided as a Docker image, can easily be utilized in exploratory analyses of scRNA-seq data. In conclusion, Millefy will provide new opportunities to analyze scRNA-seq data from the point of view of cell-to-cell heterogeneity in read coverage, and help researchers assess cellular heterogeneity and RNA biology using scRNA-seq data.

Availability and requirements

Project name: Millefy
Project home page: https://github.com/yuifu/millefy (R package), https://hub.docker.com/r/yuifu/datascience-notebook-millefy (Docker image)
Archived version: DOI:10.5281/zenodo.3591109
Operating system(s): Platform independent
Programming language: R
Other requirements: R version 3.2.2 or higher
License: MIT
Any restrictions to use by non-academics: No

Abbreviations

eRNAs: enhancer RNAs; mESCs: mouse embryonic stem cells; PCA: principal component analysis; QC: quality control; scRNA-seq: single-cell RNA sequencing; TNBC: triple negative breast cancer

Acknowledgements

We thank the members of the Bioinformatics Research Unit, particularly Hirotaka Matsumoto and Mika Yoshimura, for discussion of data analyses and Manabu Ishii and Akihiro Matsushima for IT infrastructure management.

Authors' contributions

HO designed and implemented the software, performed the data analyses, and wrote the manuscript. TH and MU performed the C1-RamDA-seq and C1-SMART-Seq V4 experiments. IN contributed to the design of the data analyses. All authors read and approved the final manuscript.

Funding

HO was supported by the Special Postdoctoral Researchers Program from RIKEN. This work was supported by the Projects for Technological Development, Research Center Network for Realization of Regenerative Medicine by Japan (18bm0404024h0001), the Japan Agency for Medical Research and Development (AMED), and JST CREST grant number JPMJCR16G3, Japan to I.N. The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

The datasets supporting the conclusions of this article are available in the Gene Expression Omnibus repository, GSE98664 and GSE118390.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1] Bioinformatics Laboratory, Faculty of Medicine, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, 305-8575, Ibaraki, Japan. [2] Center for Artificial Intelligence Research, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, 305-8577 Ibaraki, Japan. [3] Laboratory for Bioinformatics Research, RIKEN Center for Biosystems Dynamics Research, 2-1 Hirosawa, Wako, 351-0198 Saitama, Japan. [4] Bioinformatics Course, Master's/Doctoral Program in Life Science Innovation, School of Integrative and Global Majors, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, 305-8575 Ibaraki, Japan.

Received: 2 February 2019 Accepted: 29 January 2020
Published online: 03 March 2020

References

1. Hotelling H. Analysis of a complex of statistical variables into principal components. J Educ Psychol. 1933;24(6):417.
2. Maaten Lvd, Hinton G. Visualizing data using t-sne. J Mach Learn Res. 2008;9(Nov):2579–605.
3. Zappia L, Phipson B, Oshlack A. Exploring the single-cell rna-seq analysis landscape with the scrna-tools database. PLoS Comput Biol. 2018;14(6):1006245.
4. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008;456(7221):470.
5. Sibley CR, Emmett W, Blazquez L, Faro A, Haberman N, Briese M, Trabzuni D, Ryten M, Weale ME, Hardy J, et al. Recursive splicing in long vertebrate genes. Nature. 2015;521(7552):371.
6. Wu H, Nord AS, Akiyama JA, Shoukry M, Afzal V, Rubin EM, Pennacchio LA, Visel A. Tissue-specific rna expression marks distant-acting developmental enhancers. PLoS Genet. 2014;10(9):1004610.
7. Archer N, Walsh MD, Shahrezaei V, Hebenstreit D. Modeling enzyme processivity reveals that rna-seq libraries are biased in characteristic and correctable ways. Cell Syst. 2016;3(5):467–79.
8. Saliba A.-E., Westermann AJ, Gorski SA, Vogel J. Single-cell rna-seq: advances and future challenges. Nucleic Acids Res. 2014;42(14):8845–60.
9. Marinov GK, Williams BA, McCue K, Schroth GP, Gertz J, Myers RM, Wold BJ. From single-cell to cell-pool transcriptomes: stochasticity in gene expression and rna splicing. Genome Res. 2014;24(3):496–510.
10. Song Y, Botvinnik OB, Lovci MT, Kakaradov B, Liu P, Xu JL, Yeo GW. Single-cell alternative splicing analysis with expedition reveals splicing dynamics during neuron differentiation. Mol Cell. 2017;67(1):148–161.
11. Picelli S, Faridani OR, Björklund ÅK, Winberg G, Sagasser S, Sandberg R. Full-length rna-seq from single cells using smart-seq2. Nat Protoc. 2014;9(1):171.
12. Hayashi T, Ozaki H, Sasagawa Y, Umeda M, Danno H, Nikaido I. Single-cell full-length total rna sequencing uncovers dynamics of recursive splicing and enhancer rnas. Nat Commun. 2018;9(1):619.
13. Baran-Gale J, Chandra T, Kirschner K. Experimental design for single-cell rna sequencing. Brief Funct Genomics. 2017. https://doi.org/10.1093/bfgp/elx035.
14. Papalexi E, Satija R. Single-cell rna sequencing to explore immune cell heterogeneity. Nat Rev Immunol. 2018;18(1):35.
15. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. Integrative genomics viewer. Nat Biotechnol. 2011;29(1):24.

16. Buels R, Yao E, Diesh CM, Hayes RD, Munoz-Torres M, Helt G, Goodstein DM, Elsik CG, Lewis SE, Stein L, et al. Jbrowse: a dynamic web platform for genome visualization and analysis. Genome Biol. 2016;17(1):66.
17. Biase F, Cao X, Zhong S. Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell rna sequencing. Genome Res. 2014;177725:. https://doi.org/10.1101/gr.177725.114.
18. Ner-Gaon H, Melchior A, Golan N, Ben-Haim Y, Shay T. Jinglebells: a repository of immune-related single-cell rna–sequencing datasets. J Immunol. 2017;198(9):3375–9.
19. Coifman RR, Lafon S, Lee AB, Maggioni M, Nadler B, Warner F, Zucker SW. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. Proc Natl Acad Sci. 2005;102(21): 7426–31.
20. Lawrence M, Gentleman R, Carey V. rtracklayer: an r package for interfacing with genome browsers. Bioinformatics. 2009;25(14):1841–2.
21. Morgan M, Pages H, Obenchain V, Hayden N. Rsamtools: Binary alignment (bam), fasta, variant call (bcf), and tabix file import. R package version. 2016;1(0):.
22. Dowle M, Srinivasan A, Short T, Lianoglou S, Saporta R, Antonyan E. data.table: Extension of Data. frame. R package version 1.9. 6. 2015. https://www.r-bloggers.com/citing-r-packages/.
23. Angerer P, Haghverdi L, Büttner M, Theis FJ, Marr C, Buettner F. destiny: diffusion maps for large-scale single-cell data in r. Bioinformatics. 2015;32(8):1241–3.
24. Onodera CS, Underwood JG, Katzman S, Jacobs F, Greenberg D, Salama SR, Haussler D. Gene isoform specificity through enhancer-associated antisense transcription. PLoS ONE. 2012;7(8):43511.
25. Karaayvaz M, Cristea S, Gillespie SM, Patel AP, Mylvaganam R, Luo CC, Specht MC, Bernstein BE, Michor F, Ellisen LW. Unravelling subclonal heterogeneity and aggressive disease states in tnbc through single-cell rna-seq. Nat Commun. 2018;9(1):3588.
26. Miles WO, Lembo A, Volorio A, Brachtel E, Tian B, Sgroi D, Provero P, Dyson N. Alternative polyadenylation in triple-negative breast tumors allows nras and c-jun to bypass pumilio posttranscriptional regulation. Cancer Res. 2016;76(24):7231–41.
27. Consortium EP, et al. An integrated encyclopedia of dna elements in the human genome. Nature. 2012;489(7414):57.
28. Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu C, Wright J, Armstrong J, et al. Gencode reference annotation for the human and mouse genomes. Nucleic Acids Res. 2018;47(D1):766–73.
29. Ramsköld D, Luo S, Wang Y-C, Li R, Deng Q, Faridani OR, Daniels GA, Khrebtukova I, Loring JF, Laurent LC, et al. Full-length mrna-seq from single-cell levels of rna and individual circulating tumor cells. Nat Biotechnol. 2012;30(8):777.
30. Matsumoto H, Hayashi T, Ozaki H, Tsuyuzaki K, Umeda M, Iida T, Nakamura M, Okano H, Nikaido I. A nmf-based approach to discover overlooked differentially expressed gene regions from single-cell rna-seq data. NAR Genomics Bioinforma. 2020;2(1):020.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.