


SOFTWARE

Open Access



ddSeeker: a tool for processing Bio-Rad ddSEQ single cell RNA-seq data

Dario Romagnoli^{1†}, Giulia Boccalini^{2†}, Martina Bonechi², Chiara Biagioni^{2,3}, Paola Fassan⁴, Roberto Bertorelli⁴, Veronica De Sanctis⁴, Angelo Di Leo³, Ilenia Migliaccio², Luca Malorni^{2,3} and Matteo Benelli^{1*} 

Abstract

Background: New single-cell isolation technologies are facilitating studies on the transcriptomics of individual cells. Bio-Rad ddSEQ is a droplet-based microfluidic system that, when coupled with downstream Illumina library preparation and sequencing, enables the monitoring of thousands of genes per cell. Sequenced reads show unique features that do not permit the use of freely available tools to perform single cell demultiplexing.

Results: We present ddSeeker, a tool to perform initial processing and quality metrics of reads generated through Bio-Rad ddSEQ/Illumina experiments. Its application to the Illumina test dataset demonstrates that ddSeeker performs better than Illumina BaseSpace software, enabling a higher recovery of valid reads. We also show its utility in the analysis of an in-house dataset including two read sets characterized by low and high sequencing quality. ddSeeker and its source code are available at <https://github.com/cgplab/ddSeeker>.

Conclusions: ddSeeker is a freely available tool to perform initial processing and quality metrics of reads generated through Bio-Rad ddSEQ/Illumina single cell transcriptomic experiments.

Keywords: Single-cell transcriptomics, scRNA-seq, Bioinformatics

Background

Recent advances in single-cell transcriptome profiling (single cell RNA-seq, scRNA-seq), are improving our understanding of different biological processes, with impact in many areas of research, including the immune system, brain and mammal development and cancer [1, 2]. scRNA-seq techniques are contributing to refine our knowledge of cell types and states [3, 4] and have been successfully used to characterize intratumoral heterogeneity in different tumor types, including glioblastoma [5], melanoma [6] and breast cancer [7]. Clinical application of scRNA-seq has been also investigated by various groups. Most of these studies have focused on dissecting the interplay between tumor cell biology and cancer treatment, with the ultimate goal of identifying new treatment hypotheses [6, 8, 9]. A recent example includes the application of scRNA-seq to triple negative breast

cancer patients in order to understand clonal evolution in response to chemotherapy [10].

A variety of computational tools have been designed to address the specific challenges of scRNA-seq data. These involve new normalization methods for dealing with the small number of reads per gene and cell [11, 12], imputation strategies to model the sparsity of the data [13, 14], statistical methods to perform differential expression analysis [15, 16] and clustering techniques to capture cell population heterogeneity [17] and cell population dynamics [18].

Several platforms enabling single-cell transcriptome profiling are based on droplet-microfluidic technology, wherein cells are encapsulated into nanoliter droplets then lysed and mRNA is barcoded [19]. Available commercial systems include 10x Genomics Chromium [20] and Bio-Rad ddSEQ Single Cell Isolator [21]. In particular, ddSEQ is a new platform capable of isolating and barcoding about 300 cells per well. Libraries are then prepared by a specific Illumina kit (SureCell WTA 3' [22]) and sequenced. Though in its infancy, the Bio-Rad ddSEQ platform has

*Correspondence: matteo.benelli@uslcentro.toscana.it

[†]Dario Romagnoli and Giulia Boccalini contributed equally to this work

¹Bioinformatics Unit, Hospital of Prato, Prato, Italy

Full list of author information is available at the end of the article



been successfully used to characterize the molecular heterogeneity of cystic precursor lesions (IPMNs) and its role towards progressive dysplastic changes [23]. ddSEQ data have been also exploited for validating new computational methods [16, 24].

Popular tools for the processing of scRNA-seq data include Drop-seq tools [25], dropSeqPipe [26], dropEst [27], scPipe [28] and zUMI [29]. The first essential step in scRNA-seq is the identification of cell-specific barcodes. In Bio-Rad ddSEQ/Illumina cell barcoding, cells are identified by the combination of three barcodes of 6 nucleotides in length. However, compared to the other available platforms, the position of barcodes within Read 1 is not fixed due to the presence of phase blocks (Fig. 1). This unique feature renders the use of currently freely available tools for these data impossible; to date, the only available and feasible tool is a commercial software integrated in the suite of Illumina BaseSpace tools [30].

Here, we present ddSeeker, a new tool for the initial processing of data generated through Bio-Rad ddSEQ experiments that shows enhanced performance compared to the available commercial software. Our tool can be integrated in different scRNA-seq pipelines, allowing the users to take advantage of popular pipelines for the processing of scRNA-seq data. Additionally, ddSeeker provides a set of metrics which can assist the user to evaluate the quality of their own data.

Implementation

Our method is implemented following the Illumina recommendations for the analysis of Read 1 (R1) structure. As reported in Fig. 1, we expect R1 to contain the molecular tags to identify both single cells (cell barcodes) and

single transcript molecules (Unique Molecular Identifiers, UMI), while Read 2 (R2) contains the mRNA sequence. To identify reads with correct molecular tags (valid reads), our method implements the following steps:

1. the exact positions of the two linkers (L1 and L2) within the sequence of R1 are retrieved;
2. the distance between the starting positions of the two linkers is verified to be exactly 21 nt;
3. L1 is verified to start at least 7 nucleotides from the start of the sequence;
4. the sequences of the two trinucleotides flanking the UMI are verified to be ACG and GAC;
5. the sequences of the three barcode blocks (BC1, BC2, BC3) and the UMI are extracted based on their position relative to the linkers;
6. the sequences of the barcode blocks are compared with a predefined list of known barcode blocks, and retrieved only if a match is found.

In order to optimize the accuracy of our pipeline, we also considered insertions and deletions in the analysis of the sequences of linkers and barcodes (step 1 and 6) and at most one mismatch. Indeed, based on the analysis reported in Additional file 1: Figure S1 and Table S1, we observed that events with more than one mismatch, insertion or deletion in linkers represent a small fraction of total events (1.4% for L1 and 1.2% for L2). In step 4 only one mismatch and no indels are permitted due to the short length of sequences to be tested. For each valid read, the cell barcode is defined as the union of the three barcodes (BC1 + BC2 + BC3). Cell identifier and UMI are then associated to the corresponding R2. ddSeeker takes

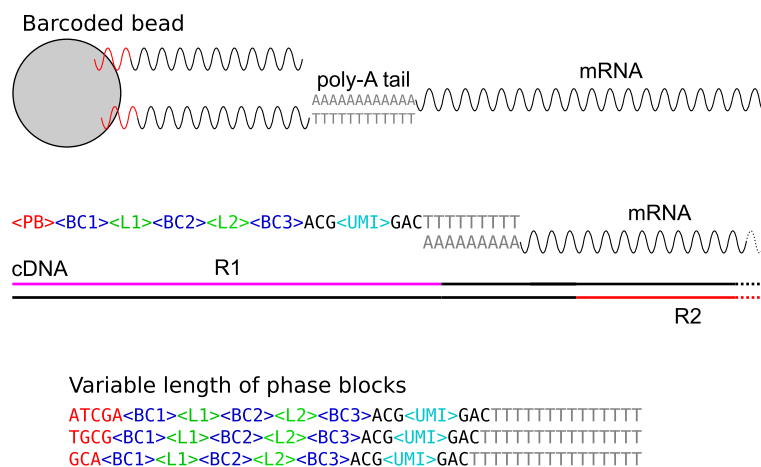


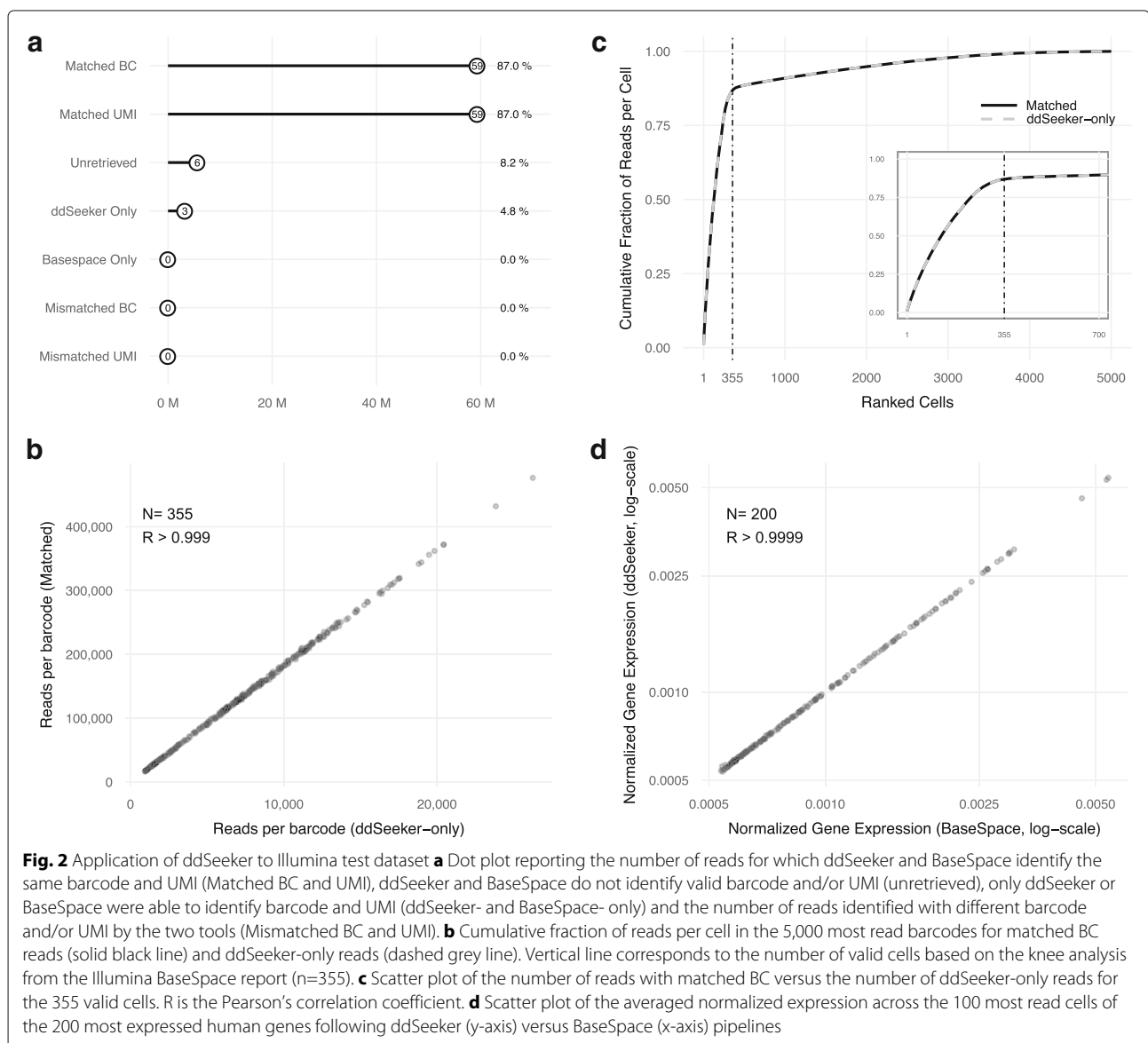
Fig. 1 Schematic of Bio-Rad ddSEQ/Illumina reads' structure (Top) In BioRad ddSEQ, barcoded beads capture mRNA molecules through hybridization with mRNA poly-A tails. Each single DNA strand is characterized by the following structure: a phase block (PB), three barcode blocks (BC1, BC2, BC3) interlinked by two different linkers (L1 and L2), and one UMI flanked by two trinucleotides (ACG and GAC). (Middle) Read 1 (R1) contains molecular tags while Read 2 (R2) contains the information of the mRNA sequence (R1 and R2 are not in scale). (Bottom) Separation of cDNA from the beads can occur at different nucleotides within the PB, thus making the position of the two linkers variable

R1 and R2 fastq files as input and outputs an unmapped BAM file containing R2 and corresponding cell barcode and UMI molecular tags. ddSeeker uses as default a Drop-seq tools -like tag scheme (XC for cell barcode and XM for UMI), but different tag schemes can be chosen by the user. For non-valid reads, ddSeeker reports the error identifier in a custom defined tag XE (see Additional file 1 for the description of possible errors). Our tool is written in Python3 using Biopython [31] and pysam modules [32]. ddSeeker can be integrated with existing scRNA-seq pipelines, including Drop-seq tools, dropEst and scPipe. A detailed description of the algorithm is reported in Additional file 1. Analyses were made using R custom scripts [33] and plots generated by the 'Tidyverse' packages [34].

Results

Analysis of Illumina test dataset

To assess the performance of ddSeeker, we considered a test dataset provided by Illumina obtained from human embryonic kidney 293 (HEK 293) cells and NIH 3T3 mouse embryonic fibroblasts mixed at a 1:1 ratio (N=1400 cells in total). In this study, we considered one of the four replicates (sample A, N=350 cells), that includes 68.2 paired-end million reads. Figure 2 reports the results of the comparative analysis between ddSeeker and BaseSpace. Overall, ddSeeker and BaseSpace identified 62.6 (91.8%) and 59.4 (87.0%) million reads with valid barcodes, respectively (Fig. 2a). The distribution of ddSeeker's error tags is reported in Additional file 1: Table S2 and



shows that the most relevant errors were alignment error in linker 2 (L2, 1.94%), followed by error in barcodes (B, 1.86%) and alignment error in both linkers (LX, 1.54%). To evaluate the performance of ddSeeker in retrieving valid barcodes, we performed a read-by-read comparison of cell barcodes and UMI identified by ddSeeker and BaseSpace. Considering BaseSpace results as a reference, we computed the percentage of reads identified by ddSeeker with the same cell barcode and UMI than BaseSpace. We found that ddSeeker was able to correctly retrieve 100% of barcodes and UMI with no mis-identification. About 8% of reads were flagged as reads with no valid barcode or UMI by both the algorithms. Of note, we found that ddSeeker was able to retrieve 5% more reads with valid barcodes than BaseSpace (Fig. 2a). We verified that all of these reads showed insertions or deletions in the sequence of the linkers or barcodes.

Quality assessment of ddSeeker's additionally detected reads

To evaluate the quality of barcodes exclusively identified by our tool, we first studied the cumulative fraction of reads per cell in the matched (i.e., barcodes identified by both BaseSpace and ddSeeker) and ddSeeker-only barcodes. As reported in Fig. 1b, we found an exact overlap between the two curves, both showing the knee at about 350 cells, as expected based on Illumina BaseSpace report. We also investigated whether considering insertions and/or deletions in our pipeline could introduce biases in the quantification of valid barcodes (i.e., the presence of certain barcodes showing more insertions or deletions than expected). As reported in Fig. 1c, we observed that the number of ddSeeker-only reads linearly correlates with the number of valid barcode reads identified by both algorithms ($R > 0.999$).

We then studied how the ddSeeker pipeline impacts on downstream analysis, including read alignment and gene counting. First, we evaluated whether the additionally detected reads mapped equally well to the reference genome. To achieve that, we compared mapping quality values extracted from the Illumina bam file for matched and ddSeeker-only valid barcode reads. We found that 72% of ddSeeker-only valid reads has high mapping quality and, in general, the mapping quality distribution for matched and ddSeeker-only valid reads were markedly similar (Additional file 1: Figure S2). Secondly, we investigated the number of doublets detected using ddSeeker and BaseSpace and observed no difference ($n=13$ for both pipelines, see Additional file 1: Figure S3), demonstrating that additionally detected reads show high species-specificity. Lastly, we compared gene expression estimations following ddSeeker and BaseSpace pipelines. To calculate gene counts, we used the DigitalExpression tool included in Drop-seq tools. A library size normalization (i.e., gene counts per cell) was applied to quantify

gene expression levels. Figure 2d reports the mean gene expression level for the 200 most expressed human genes across the 100 most read cells using ddSeeker and BaseSpace (results for mouse genes are reported in Additional file 1: Figure S4). We obtained high correlation ($R > 0.999$) between ddSeeker and BaseSpace, demonstrating that overall gene expression estimation is not biased by additionally detected reads by ddSeeker.

Application of ddSeeker to in-house dataset

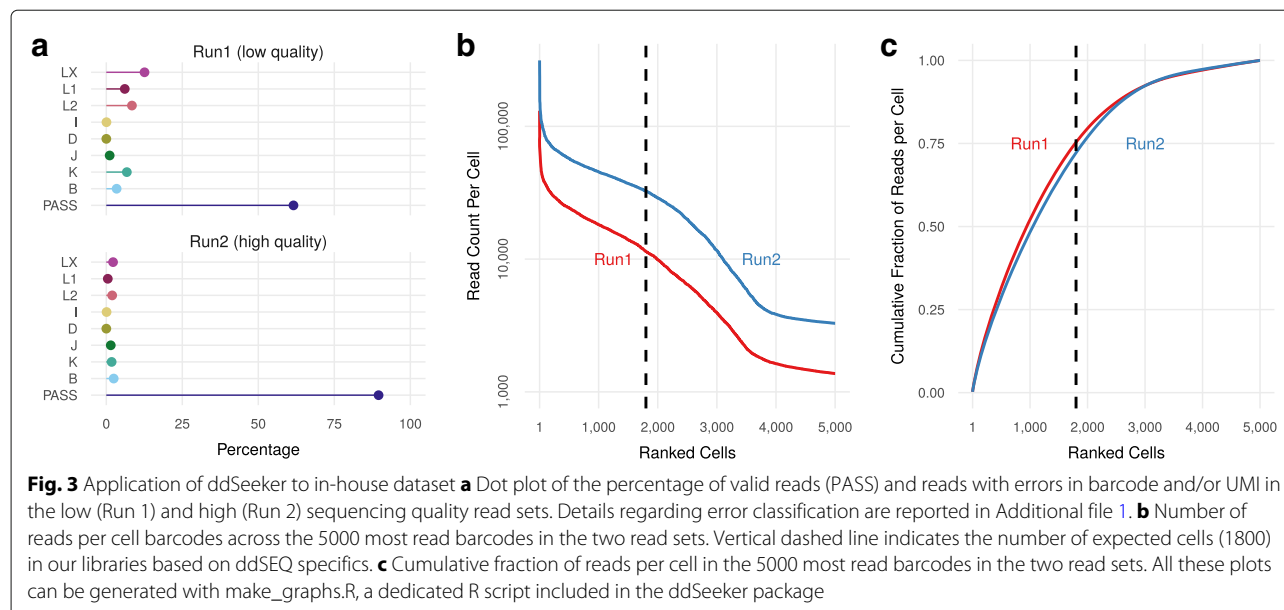
We tested ddSeeker further, in the analysis of an in-house dataset that includes 6 scRNA-seq libraries of the MDA-MB-361 breast cancer cell line. Details about cell culture, library preparation and sequencing are reported in the Additional file 1. Our dataset was generated by two Illumina runs characterized by low (cluster saturation) and high sequencing quality (Table 1). FastQC analysis [35] of a subset of R1 ($N = 10M$ reads for both low and high quality read sets) is reported in Additional file 1: Figure S5.

Using a 40-core machine, ddSeeker takes approximately 7.0 h to complete the analysis on 310,594,139 million reads from both runs, corresponding to about 1.1 million reads/cpu processed per hour. About 63% and 91% of the total reads were identified as valid in Run 1 (low quality) and Run 2 (high quality), respectively. Figure 3a and Additional file 1: Table S3 report the classification of the errors found by ddSeeker in the R1 of the two runs, and show that error tag distribution in Run 2 was comparable with that one obtained in the Illumina test dataset. We also found that the error tag distributions were the same across the different scRNA-seq libraries (Additional file 1: Figure S6). ddSeeker can output a text file reporting the number of valid reads per cell which can be useful to preliminary estimate the number of sequenced cells before computationally intensive steps such as read alignment, processing and gene counting (Fig. 3b). The number of reads obtained in the 5000 most read barcodes for the two runs is reported in Fig. 3c. Despite the difference between

Table 1 Sequencing run summary of the in-house dataset

	Run 1	Run 2
Quality	Low	High
# Libraries	6	6
Clusters PF (%)	33.05	93.9
Q30 (%)	70.9	82.6
Total Reads	112993193	197600946
Valid Reads	70826303 (63%)	180526539 (91%)
# Cells (expected)	1800	1800
# Cells (ddSeeker)	≈ 3000	≈ 3000

Valid reads are reads with valid barcodes and UMI. Expected number of cells is based on cell capture efficiency, as declared by Bio-Rad



the two read sets in terms of sequence quality, the curves showed a similar trend.

Conclusions

ddSeeker is, to our knowledge, the first freely-available tool to perform initial processing and quality metrics of reads generated through Bio-Rad ddSEQ/Illumina experiments. We performed a comparative study of ddSeeker and BaseSpace in the analysis of an Illumina test dataset. We showed that ddSeeker was able to identify 5% more reads with valid barcodes and UMI than the Illumina BaseSpace tool. The enhanced ability of ddSeeker in identifying reads with valid barcodes derives from its exclusive feature that implements the analysis of insertions or deletions events in the sequences of linkers and barcodes. Additionally, we demonstrated the reliability of ddSeeker's additionally detected reads. Our analyses show that downstream analysis is not biased in terms of mapping quality, presence of doublets and gene expression quantification. Finally, we showed the utility of ddSeeker in the analysis of an in-house dataset that includes two different read sets characterized by low and high sequencing quality. To conclude, our analyses suggest that ddSeeker is a valuable tool to perform quality control of Bio-Rad ddSEQ data, and to identify valid reads for downstream scRNA-seq data analysis.

Availability and requirements

Project name: ddSeeker

Project home page: <https://github.com/cgplab/ddSeeker>

Operating system: GNU/Linux

Programming language: Python 3

License: GPL-3.0

Abbreviations

BC1: Barcode block 1; BC2: Barcode block 2; BC3: Barcode block 3; UMI: Unique molecular identifier; scRNA-seq: Single cell RNA sequencing; uBAM: Unmapped BAM

Additional file

Additional file 1: Supplementary text, figures and tables. Supplementary text 1. Classification of error tags. Supplementary text 2. Detailed algorithm description. Supplementary figure 1. Linker alignment analysis. Supplementary figure 2. Error tags distribution for the in-house dataset. Supplementary table 1. Linker alignment analysis. Supplementary table 2. Error tags distribution for the Illumina dataset. Supplementary table 3. Error tags distribution for the in-house dataset. (PDF 407 kb)

Acknowledgements

Not applicable.

Funding

This work has been supported by Associazione Italiana per la Ricerca sul Cancro (MFAG n. 14371 to IM and 18880 to LM) for the development of the cell line model, and Fondazione Sandro Pitigliani per la lotta contro i tumori ONLUS and Fondazione Cassa di Risparmio di Firenze (MBE) for generation, sequencing and analysis of single cell RNA-seq experiments.

Availability of data and materials

The Illumina test dataset is available through the Illumina BaseSpace Sequence Hub [30]. The entire in-house dataset is available from the corresponding author upon request. A subset of the dataset is available through a dedicated GitHub repository [36]. The code used to conduct the data analyses of this manuscript is available from the corresponding author upon request.

Authors' contributions

MBE supervised development of work. MBE, DR and GB wrote the manuscript. DR designed and implemented the software. GB performed single cell experiments. MBO assisted in preparing the single cell libraries. PF, VDS and RB performed sequencing and helped to evaluate the data. ADL, IM and LM helped to evaluate the data and edit the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

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

Author details

¹Bioinformatics Unit, Hospital of Prato, Prato, Italy. ²Sandro Pitigliani Translational Research Unit, Hospital of Prato, Prato, Italy. ³Sandro Pitigliani Medical Oncology Department, Hospital of Prato, Prato, Italy. ⁴NGS Core Facility, Centre for Integrative Biology (CIBIO), University of Trento, Trento, Italy.

Received: 6 July 2018 Accepted: 14 November 2018

Published online: 24 December 2018

References

- Tang F, Lao K, Surani MA. Development and applications of single-cell transcriptome analysis. *Nat Methods*. 2011;8(4 Suppl):6–11. <https://doi.org/10.1038/nmeth.1557>.
- Shapiro E, Biezuner T, Linnarsson S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat Rev Genet*. 2013;14:618.
- Trapnell C. Defining cell types and states with single-cell genomics. *Genome Res*. 2015;25(10):1491–8. <https://doi.org/10.1101/gr.190595.115>. <http://genome.cshlp.org/content/25/10/1491.full.pdf+html>.
- Saunders A, Macosko E, Wysoker A, Goldman M, Krienen F, Bien E, Baum M, Wang S, Goeva A, Nemesh J, Kamitaki N, Brumbaugh S, Kulp D, McCarroll SA. A single-cell atlas of cell types, states, and other transcriptional patterns from nine regions of the adult mouse brain. *bioRxiv*. 2018. <https://doi.org/10.1101/299081>. <http://arxiv.org/abs/https://www.biorxiv.org/content/early/2018/04/10/299081.full.pdf>.
- Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, Cahill DP, Nahed BV, Curry WT, Martuza RL, Louis DN, Rozenblatt-Rosen O, Suvà ML, Regev A, Bernstein BE. Single-cell rna-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*. 2014;344(6190):1396–401. <https://doi.org/10.1126/science.1254257>. <http://science.sciencemag.org/content/344/6190/1396.full.pdf>.
- Tirosh I, Izar B, Prakadan SM, Wadsworth MH, Treacy D, Trombetta JJ, Rotem A, Rodman C, Lian C, Murphy G, Fallahi-Sichani M, Dutton-Regester K, Lin J-R, Cohen O, Shah P, Lu D, Genshaft AS, Hughes TK, Ziegler CGK, Kazer SW, Gaillard A, Kolb KE, Villani A-C, Johannessen CM, Andreev AY, Van Allen EM, Bertagnoli M, Sorger PK, Sullivan RJ, Flaherty KT, Frederick DT, Jané-Valbuena J, Yoon CH, Rozenblatt-Rosen O, Shalek AK, Regev A, Garraway LA. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell rna-seq. *Science*. 2016;352(6282):189–96. <https://doi.org/10.1126/science.aad0501>. <http://science.sciencemag.org/content/352/6282/189.full.pdf>.
- Chung W, Eum HH, Lee H-O, Lee K-M, Lee H-B, Kim K-T, Ryu HS, Kim S, Lee JE, Park YH, Kan Z, Han W, Park W-Y. Single-cell rna-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nat Commun*. 2017;8:15081.
- Tirosh I, Venteicher AS, Hebert C, Escalante LE, Patel AP, Yizhak K, Fisher JM, Rodman C, Mount C, Filbin MG, Neftel C, Desai N, Nyman J, Izar B, Luo CC, Francis JM, Patel AA, Onozato ML, Riggi N, Livak KJ, Gennert D, Satija R, Nahed BV, Curry WT, Martuza RL, Mylvaganam R, Iafate AJ, Frosch MP, Golub TR, Rivera MN, Getz G, Rozenblatt-Rosen O, Cahill DP, Monje M, Bernstein BE, Louis DN, Regev A, Suvà ML. Single-cell rna-seq supports a developmental hierarchy in human oligodendroglioma. *Nature*. 2016;539:309.
- Kim K-T, Lee HW, Lee H-O, Song HJ, Jeong DE, Shin S, Kim H, Shin Y, Nam D-H, Jeong BC, Kirsch DG, Joo KM, Park W-Y. Application of single-cell rna sequencing in optimizing a combinatorial therapeutic strategy in metastatic renal cell carcinoma. *Genome Biol*. 2016;17(1):80. <https://doi.org/10.1186/s13059-016-0945-9>.
- Kim C, Gao R, Sei E, Brandt R, Hartman J, Hatschek T, Crosetto N, Foukakis T, Navin NE. Chemoresistance evolution in triple-negative breast cancer delineated by single-cell sequencing. *Cell*. 2018;173(4):879–89313. <https://doi.org/10.1016/j.cell.2018.03.041>.
- Bacher R, Chu L-F, Leng N, Gasch AP, Thomson JA, Stewart RM, Newton M, Kendziorski C. Scnorm: robust normalization of single-cell rna-seq data. *Nat Methods*. 2017;14:584.
- Lun AT, Bach K, Marioni JC. Pooling across cells to normalize single-cell rna sequencing data with many zero counts. *Genome Biol*. 2016;17(1):75. <https://doi.org/10.1186/s13059-016-0947-7>.
- Li WW, Li JJ. An accurate and robust imputation method scimpute for single-cell rna-seq data. *Nat Commun*. 2018;9(1):997. <https://doi.org/10.1038/s41467-018-03405-7>.
- van Dijk D, Sharma R, Nainys J, Yim K, Kathail P, Carr AJ, Burdziak C, Moon KR, Chaffer CL, Pattabiraman D, Bieri B, Mazutis L, Wolf G, Krishnaswamy S, Pe'er D. Recovering gene interactions from single-cell data using data diffusion. *Cell*. 2018;174(3):716–72927. <https://doi.org/10.1016/j.cell.2018.05.061>.
- Qiu X, Hill A, Packer J, Lin D, Ma Y-A, Trapnell C. Single-cell mrna quantification and differential analysis with census. *Nat Methods*. 2017;14:309.
- Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol*. 2018;36:411.
- Rodriguez A, Laio A. Clustering by fast search and find of density peaks. *Science*. 2014;344(6191):1492–6. <https://doi.org/10.1126/science.1242072>. <http://science.sciencemag.org/content/344/6191/1492.full.pdf>.
- Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, Lennon NJ, Livak KJ, Mikkelsen TS, Rinn JL. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol*. 2014;32:381.
- Prakadan SM, Shalek AK, Weitz DA. Scaling by shrinking: empowering single-cell 'omics' with microfluidic devices. *Nat Rev Genet*. 2017;18(6):345–61. <https://doi.org/10.1038/nrg.2017.15>.
- 10x Genomics. <https://www.10xgenomics.com>. Accessed Oct 2018.
- Bio-Rad. <https://www.bio-rad.com>. Accessed Oct 2018.
- Illumina. <http://www.illumina.com>. Accessed Oct 2018.
- Bernard V, Semaan A, Huang J, San Lucas FA, Mulu FC, Stephens BM, Guerrero PA, Huang Y, Zhao J, Kamyabi N, Sen S, Scheet PA, Taniguchi CM, Kim MP, Tzeng C-W, Katz MH, Singhi AD, Maitra A, Alvarez HA. Single cell transcriptomics of pancreatic cancer precursors demonstrates epithelial and microenvironmental heterogeneity as an early event in neoplastic progression. *bioRxiv*. 2018. <https://doi.org/10.1101/306134>. <https://www.biorxiv.org/content/early/2018/04/26/306134.full.pdf>.
- Monocle. <http://cole-trapnell-lab.github.io/monocle-release/docs/>. Accessed Oct 2018.
- Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, Trombetta JJ, Weitz DA, Sanes JR, Shalek AK, Regev A, McCarroll SA. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*. 2015;161(5):1202–14. <https://doi.org/10.1016/j.cell.2015.05.002>.
- dropSeqPipe. <http://github.com/Hoohm/dropSeqPipe>. Accessed Oct 2018.
- Petukhov V, Guo J, Baryawno N, Severe N, Scadden DT, Samsonova MG, Kharchenko PV. dropest: pipeline for accurate estimation of molecular counts in droplet-based single-cell rna-seq experiments. *Genome Biol*. 2018;19(1):78. <https://doi.org/10.1186/s13059-018-1449-6>. Accessed Oct 2018.
- Tian L, Su S, Dong X, Amann-Zalcenstein D, Biben C, Seidi A, Hilton DJ, Naik SH, Ritchie ME. scpipe: A flexible r/bioconductor preprocessing pipeline for single-cell rna-sequencing data. *PLoS Comput Biol*. 2018;14(8):1–15. <https://doi.org/10.1371/journal.pcbi.1006361>.
- Parekh S, Ziegenhain C, Vieth B, Enard W, Hellmann I. zumis a fast and flexible pipeline to process rna sequencing data with umis. *GigaScience*. 2018;7(6):059. <https://doi.org/10.1093/gigascience/giy059>.
- Illumina BaseSpace. <https://basespace.illumina.com>. Accessed Oct 2018.
- Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, de Hoon MJL. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*. 2009;25(11):1422–3. <https://doi.org/10.1093/bioinformatics/btp163>.

32. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, and RD. The sequence alignment/map format and SAMtools. *Bioinformatics*. 2009;25(16):2078–9. <https://doi.org/10.1093/bioinformatics/btp352>.
33. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2018. R Foundation for Statistical Computing. <https://www.R-project.org/>.
34. Wickham H. Tidyverse: Easily Install and Load the 'Tidyverse'. 2017. R package version 1.2.1. <https://CRAN.R-project.org/package=tidyverse>. Accessed Oct 2018.
35. FastQC. <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>. Accessed Oct 2018.
36. Subset of the In-house Dataset. https://github.com/cgplab/ddSeeker_example_dataset. Accessed Oct 2018.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

