PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

scAmpi—A versatile pipeline for single-cell RNA-seq analysis from basics to clinics

Anne Bertolini[1,2‡], **Michael Prummer**[1,2‡], **Mustafa Anil Tuncel**[3], **Ulrike Menzel**[3], **María Lourdes Rosano-González**[1,2], **Jack Kuipers**[2,3], **Daniel Johannes Stekhoven**[1,2], **Tumor Profiler consortium**[¶], **Niko Beerenwinkel**[2,3], **Franziska Singer**[1,2]*

1 ETH Zurich, NEXUS Personalized Health Technologies, Zurich, Switzerland, **2** SIB Swiss Institute of Bioinformatics, Zurich, Switzerland, **3** ETH Zurich, Department of Biosystems Science and Engineering, Basel, Switzerland

‡ These authors share first authorship on this work.
¶ Membership of Tumor Profiler consortium is provided in the Acknowledgments.
* singer@nexus.ethz.ch

Citation: Bertolini A, Prummer M, Tuncel MA, Menzel U, Rosano-González ML, Kuipers J, et al. (2022) scAmpi—A versatile pipeline for single-cell RNA-seq analysis from basics to clinics. PLoS Comput Biol 18(6): e1010097. https://doi.org/10.1371/journal.pcbi.1010097

Editor: Jason A. Papin, University of Virginia, UNITED STATES

Received: April 26, 2021

Accepted: April 12, 2022

Published: June 3, 2022

Data Availability Statement: The source code of scAmpi as well as usage examples are distributed open source on github at https://github.com/ETH-NEXUS/scAmpi_single_cell_RNA.

Funding: The study described in this paper is the result of a jointly-funded effort between several academic institutions (University of Zurich, University Hospital Zurich, Swiss Federal Institute of Technology in Zurich, University Hospital Basel), as well as F. Hoffmann-La Roche AG. They were involved in data collection. UM was supported by

Abstract

Single-cell RNA sequencing (scRNA-seq) has emerged as a powerful technique to decipher tissue composition at the single-cell level and to inform on disease mechanisms, tumor heterogeneity, and the state of the immune microenvironment. Although multiple methods for the computational analysis of scRNA-seq data exist, their application in a clinical setting demands standardized and reproducible workflows, targeted to extract, condense, and display the clinically relevant information. To this end, we designed scAmpi (**S**ingle **C**ell **A**nalysis **m**RNA **pi**peline), a workflow that facilitates scRNA-seq analysis from raw read processing to informing on sample composition, clinically relevant gene and pathway alterations, and *in silico* identification of personalized candidate drug treatments. We demonstrate the value of this workflow for clinical decision making in a molecular tumor board as part of a clinical study.

Author summary

Single-cell RNA sequencing (scRNA-seq) measures the expression levels across the genes expressed in each single cell. Thus, it is well suited to inform on the cell type composition and the function of cells in different tissues and diseases. However, it is challenging to correctly process and interpret the large amounts of data generated with scRNA-seq. To this end, we have developed an analysis workflow named scAmpi (Single Cell Analysis mRNA pipeline) that starts on the raw sequencing data and performs preprocessing, quality control, and subsequent analysis steps following state-of-the-art recommendations for scRNA-seq processing. The workflow removes low quality cells, assigns a cell type label to each cell, and visualizes the expression of individual genes of interest and functional pathways on the single cells. Moreover, in disease-related analyses scAmpi can link the observed gene expression to potential drug candidates that could be suited to treat the disease.

the ETH domain Personalized Health and Related Technologies (PHRT-510). The PHRT had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

In recent years, single-cell RNA sequencing (scRNA-seq) emerged as a high-throughput technology for uncovering gene expression at the single-cell level, which provides unprecedented insights into, e.g., cell differentiation, the immune compartment, and tumor heterogeneity [1,2]. Initially used to characterize PBMCs or differentiating stem cells, an increasing number of studies exploit scRNA-seq to investigate clinical samples such as tumor tissues [3,4]. There are multiple software suites available with extensive functionality for general scRNA-seq analysis, including the widely-used tools SEURAT [5] and ScanPy [6] or the web-based software suites CreSCENT [7] and ASAP [8]. However, they have some disadvantages: First, for non-bioinformaticians the usage can be difficult because setting all parameters and applying the different steps requires at least basic R or Python programming knowledge. Second, to the best of our knowledge, no software is available that facilitates *in-silico* drug candidate identifications based on single-cell data. Finally, existing software suites are not designed to manage large-scale data analysis in a highly reproducible, transparent, and auditable way, including error tracking and process documentation, and thus are not suitable to be employed for routine clinical use [9,10].

We therefore developed scAmpi, an end-to-end turn-key pipeline for scRNA-seq analysis from raw read processing to informing on sample composition, gene expression, and potential drug candidates. Utilizing the Snakemake workflow management system [11], scAmpi is easy to use and offers a high degree of flexibility in the choice of methods, while it can be employed in a highly standardized and reproducible fashion. This has led to the successful implementation of scAmpi for processing scRNA-seq data in the ongoing Tumor Profiler clinical study [12,13].

Design and implementation

Ethics statement. Ethics approval has been granted by the Kantonal Ethics Commission of Zürich with approval number BASEC-Nr.2018-02050.

In the following, we describe how scAmpi can be used for analyzing tumor scRNA-seq data from the 10x Genomics platform (Fig 1). While the initial installation of scAmpi and its dependencies demands basic IT know-how, running the pipeline only requires some familiarity with executing command-line code. Interpretation of the output tables and graphs is easily done by anyone with a general understanding of single cell transcriptomics analysis [14].

The default scRNA analysis workflow implemented in scAmpi follows state of the art recommendations [14] and the individual tools chosen for the different tasks have shown to produce high quality, reproducible results in our hands. However, the current choices may not be optimal for all possible situations and one or more tools may have to be exchanged with more suitable alternatives. For this reason, all workflow steps can be replaced with little effort and the workflow is directly applicable also to other tissue types.

A complete analysis of a single sample with approx. 4,000 cells and 50,000 reads per cell takes four to eight hours, depending on the available compute resources. The pipeline can easily scale to the parallel analysis of large cohorts of hundreds of samples, where each sample is processed independently in a single-sample analysis fashion. To ensure a thoroughly documented analysis, each workflow step is tracked with log files describing command, input, output, and resource requirements, as well as error documentation.

Read data processing and normalization. Using the Cellranger software, reads are assigned to their respective cells based on the 10x Genomics barcodes and simultaneously mapped to the reference genome to infer read counts per gene per cell. Subsequently, several

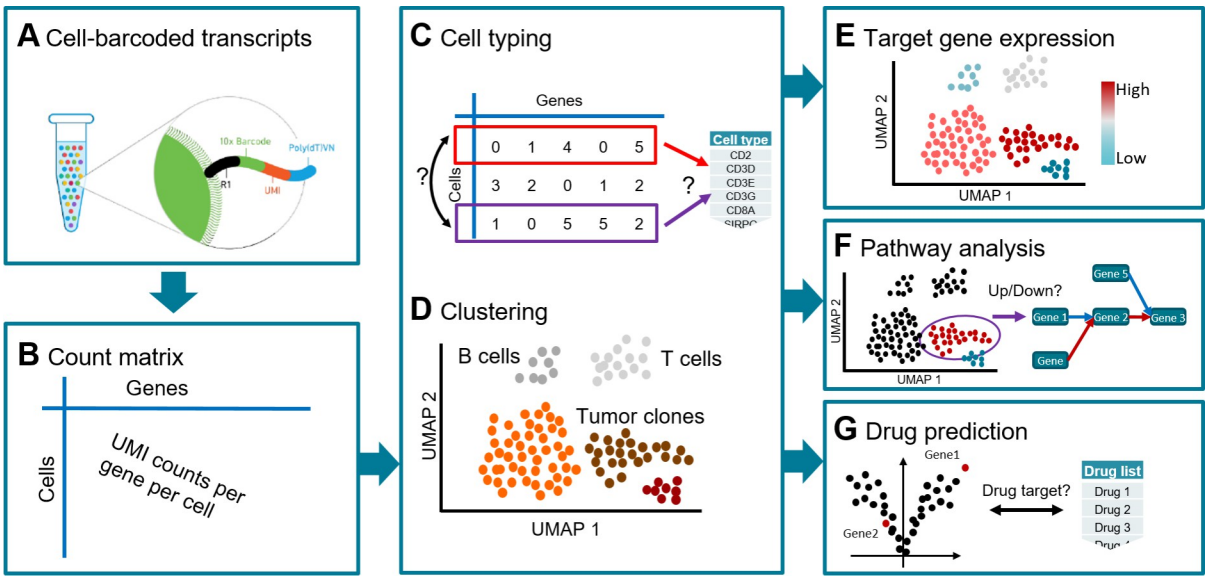

Fig 1. Overview of the workflow implemented in scAmpi, showing a tumor sample analysis as an example. Starting from droplet-based 10x Genomics raw data (A), genome-wide read counts for each cell are generated (B). This gene-by-cell count matrix is the basis for cell type prediction (C) and unsupervised clustering (D) to determine the cell type composition and tumor heterogeneity. Subsequent steps include gene expression (E) and gene set (F) analysis, and drug candidate identification (G).

https://doi.org/10.1371/journal.pcbi.1010097.g001

filters are applied to remove contaminants, cell fragments, or dying cells. Doublet detection and removal is done using scDblFinder [15]. Per default, all non-protein-coding genes and genes coding for ribosomal proteins are removed. All cells exceeding a specific threshold in the number of reads mapping to mitochondrial genes are discarded, because they are likely broken cells [16]. This threshold can be either user-specified or estimated from the data. Further, all cells with too few expressed genes are discarded in order to remove low-quality cells or presumably empty droplets. The remaining counts are normalized for cell-cycle effect and library size using sctransform [17], which yields Pearson residuals as well as corrected counts per gene per cell for the subsequent analysis steps. Several types of plots, including scatter plots showing cells that are filtered out and box plots showing the most highly expressed genes in the sample, are provided to support quality control.

Sample composition. The two key analyses to inform on sample composition are cell type identification (Fig 1C) and unsupervised clustering (Fig 1D). Per default, clustering is performed using Phenograph [18]. The clustering compares expression profiles across cells and yields groups of highly similar cells. Per default, a minimum of 20 cells per group are required, in order to reach a group size suitable for subsequent differential gene expression analysis.

In contrast, automated cell type classification is applied to each cell individually [19]. Briefly, the expression profile of each cell is compared to *a priori* defined lists of cell type marker genes. Each cell type is represented by a list of genes that are known to be specific for and highly expressed in this cell type. The set of cell types used for classification is expected to reflect the cell types present in the analyzed tissue. The method accommodates for uncertainty in the typing as well as unknown cell types. If the expression profile of a cell does not reach a specified similarity threshold, it is labeled as 'unknown'. If a cell matches two or more cell types with high similarity (i.e., the best and second-best similarity scores are too similar), it is typed as 'uncertain'. Cell type lists can be derived from literature. For example, for melanoma biopsies, we based our typing on the markers published by Tirosh et al. [3]. Using a cell type list that was derived from data of another tissue is also possible, but should be done with care

due to tissue-dependent expression differences. The cell type analysis works in a two-step hierarchical fashion. In the first iteration the major cell type populations are identified, e.g., tumor cells are distinguished from T cells. In a second step, all cells belonging to a particular major cell type can be re-classified into subtypes. For instance, T cells can be sub-classified (among others) into gamma delta, memory resting, or regulatory T cells. scAmpi already offers predefined cell type lists for melanoma, AML, ovarian cancer, and PBMCs, but user-specified marker lists can be easily added.

The results of the sample composition analysis (unsupervised clustering and cell typing) are visualized in a low-dimensional representation using, e.g., Uniform Manifold Approximation and Projection (UMAP) [20].

Differential gene expression. Detecting differential gene expression (DE) is a major aspect of standard mRNA sequencing experiments. Here, we perform two main comparisons for scRNA-seq data using multiple linear regression: First, provided multiple tumor clusters are found, a DE analysis is performed that compares the expression phenotypes of the different tumor clusters and informs on the tumor heterogeneity. Second, provided malignant (tumor) cells as well as non-malignant cells are found, a DE analysis is performed that identifies genes with different expression levels in each tumor cluster compared to all non-malignant cells. Non-malignant cells can be any cell type present in the tissue, such as, immune cells, endothelial, or epithelial cells. The fold-change (FC) and FDR cutoffs applied to the DE analysis can be specified by the user (per default scAmpi applies $|\log FC| > 2$ and $FDR < 0.01$).

Gene expression and pathway analysis. The user can provide grouped lists of priority genes and pathways to be visualized (Fig 1E). Gene expression is visualized for each cell in a color-coded UMAP together with a violin plot that shows the expression distribution per cluster, separately for each group of genes. Further, for each cluster, various gene expression summary statistics are provided, such as the gene expression rank, the average expression, and the proportion of cells with non-zero expression.

Pathway analysis is performed in two independent approaches (Fig 1F). Based on the DE genes, a competitive gene set analysis is performed using the camera function from the limma R package [21]. Here, we output all pathways with an FDR below a user-defined cut-off that are up-regulated, down-regulated, or are categorized as mixed if both over- and under-expressed genes were identified in the respective pathway. Gene set enrichment based on DE analysis is very common, but has certain drawbacks for single-cell data, as these experiments often lack a proper reference, which can bias the pathway enrichment. Thus, we also perform a GSVA-based pathway analysis [22], in which gene sets are ranked relative to each other within each cell independent of all other cells. As this approach is comparing gene sets within a cell, it does not rely on the presence of a reference cell population.

In-silico drug candidate identification. Initially developed for bulk sequencing data [23], the *in-silico* drug candidate identification framework was refined and adapted to facilitate single-cell and expression data analysis (Fig 1G). For each tumor cell cluster, the differentially expressed genes resulting from the comparison of malignant versus non-malignant cells are used to query DGIdb [24] to obtain potential drug-gene interactions. These drug-gene interactions are undirected in the sense that they do not reveal whether the tumor might be sensitive or resistant to the identified drug. Thus, we further enrich the drug-gene interactions with information from clinicaltrials.gov and CIViC [25]. CIViC is a database of curated drug-gene interaction information providing information on the observed expression type, i.e., over-expression or under-expression. This directed *in-silico* drug candidate identification is also visualized on the sample composition UMAP.

Results

We showcase the readout and analyses possible with scAmpi for scRNA-seq data from a biopsy of a melanoma patient who was included in the Tumor Profiler clinical study [12]. The full analysis from raw fastq files to *in-silico* drug candidate identification is triggered with only two commands. For details on the default parameter settings, we refer to S1 Text. In the initial mapping step, Cellranger identifies 4193 cells. Subsequent filtering in scAmpi removes 10% (437) of cells due to low quality (Fig 2A). Fig 2B and 2C show examples of QC metrics on the UMAP representation of the cells. After normalization, the cell-cycle phase has no apparent effect on the embedding of the cells anymore. Instead, as shown in Fig 3A, the embedding is cleanly separated by cell type populations.

The cell type composition analysis identifies a melanocytic melanoma cell population that constitutes 34% of the sample. The tumor immune microenvironment is very diverse and shows a large group of T cells, mainly sub-classified as memory effector T cells, as well as macrophages, B cells, NK cells, and Endothelial cells (Fig 3A). This finding is in agreement with results from CyTOF experiments also performed on the case study presented in [12]. Further investigation of the immune microenvironment is facilitated by gene expression visualization and a population-based and ranked overview of the average gene expression and number of non-zero cells for each gene. For instance, as shown in Fig A in S2 Text, memory effector T cells express PDCD1 (PD-1), an immune checkpoint marker relevant for immunotherapy. Other immune checkpoint markers are also expressed, together with an observed MHC class I expression (HLA-A/B/C) on the tumor cells indicating that T cells would be able to recognize tumor cells. Taken together, this molecular phenotype suggests a potential suitability of anti-PD1 immunotherapy. This finding is also supported by other technologies presented in [12], such as CyTOF and imaging mass cytometry.

Unsupervised clustering (Fig 3A) reveals that the melanoma population groups into four clusters, indicating tumor heterogeneity. scAmpi offers multiple readouts to further investigate this heterogeneity, including, e.g., individual gene expression analysis, gene set enrichment analysis, and differential gene expression comparing the tumor clusters (see S3 Text for

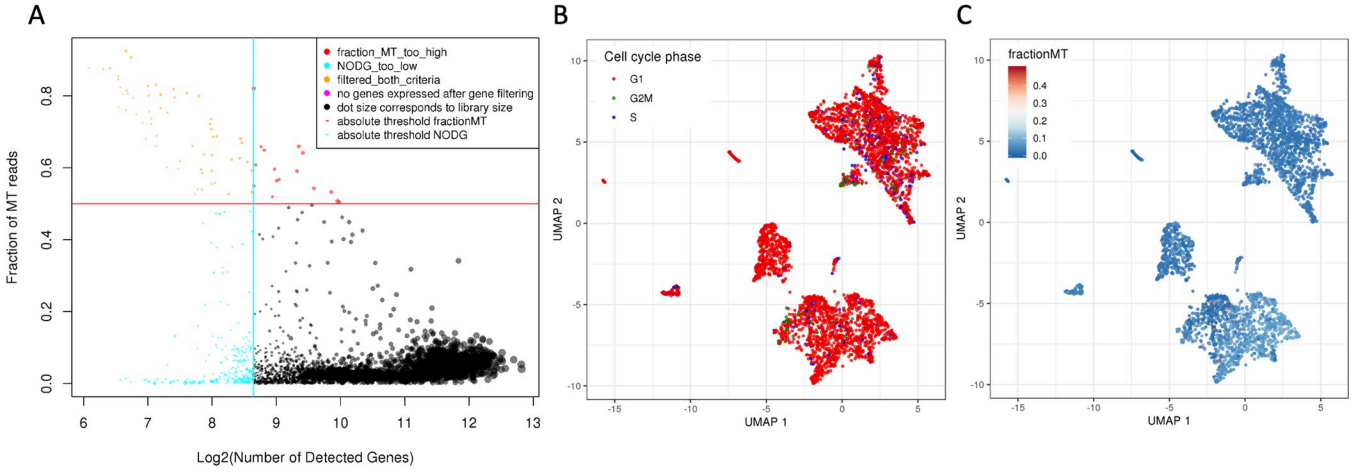

Fig 2. Examples of scAmpi's basic scRNA-seq quality control plots of a melanoma sample. The scatter plot in (A) shows cells colored by their respective category of applied filters. The vertical and horizontal lines indicate the chosen thresholds applied for the minimum number of genes (x-axis) and maximum fraction of reads mapping to mitochondrial genes per cell (y-axis), respectively. In (B), the UMAP embedding (after normalization) of all cells is shown, with cells colored by estimated cell-cycle phase. In (C), the same UMAP is shown, this time with cells colored by the fraction of reads mapping to mitochondrial genes.

https://doi.org/10.1371/journal.pcbi.1010097.g002

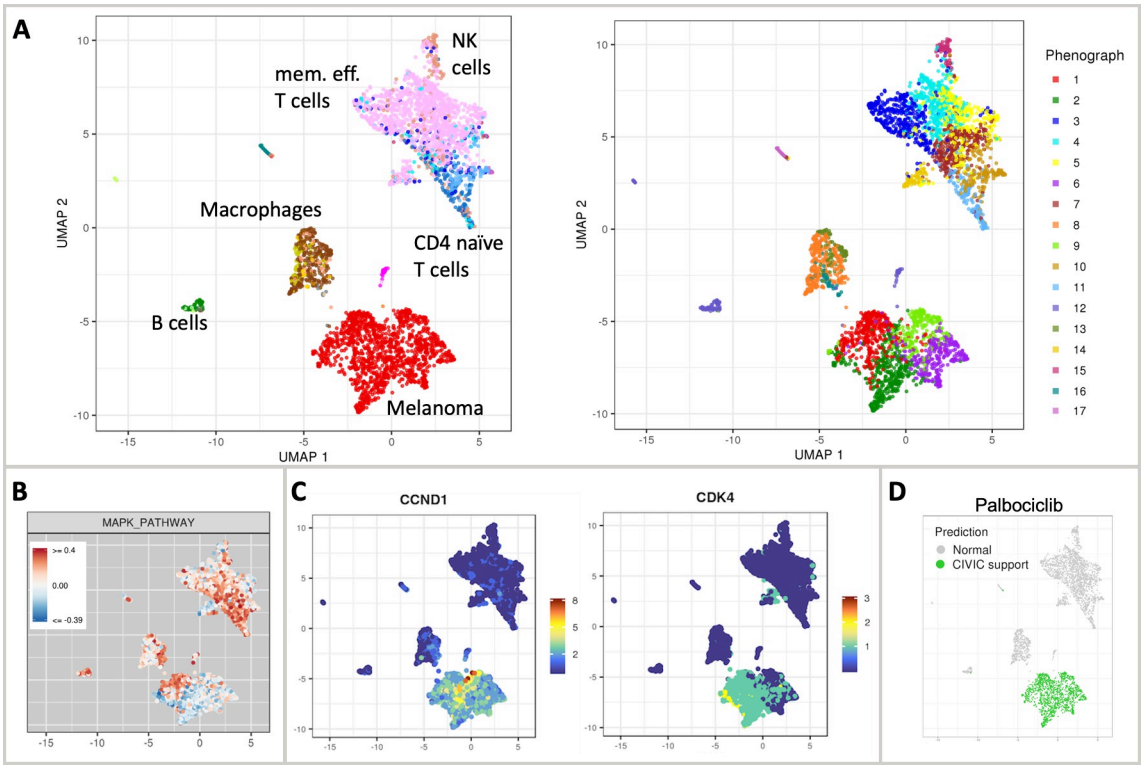

Fig 3. Sample composition and interpretation of a melanoma sample. In (A) the UMAP embedding is colored by cell type label (left) and cluster (right), with major cell type populations highlighted in the figure. For a complete overview of cell types, see Fig B in S2 Text. In (B), the enrichment of the MAPK pathway is exemplified. In (C), UMAPs showing the gene expression of CCND1 and CDK4 are shown as selected examples of individual gene expression plots. The UMAP in (D) shows the drug candidate identification result for the drug palbociclib.

https://doi.org/10.1371/journal.pcbi.1010097.g003

details). As shown in Fig 3B, three of the four tumor clusters display down-regulation of the MAPK pathway (gene set taken from the Hallmark MSigDB [26]), precluding the use of BRAF/MEK inhibitor treatment. In contrast, the *in-silico* drug candidate identification of scAmpi marked the complete tumor population to be potentially sensitive to palbociclib treatment, based on the over-expression of CCND1 and further supported by the expression of CDK4 (Fig 3C and 3D). This finding is observed across other technologies described in [12], such as drug response testing (referred to as Pharmacoscopy).

Taken together, scAmpi provides not only insights into the general sample composition and gene and pathway expression, but also enables downstream data interpretation to support clinical decision making.

Availability and future directions

The source code of scAmpi is available on github at https://github.com/ETH-NEXUS/scAmpi_single_cell_RNA. scAmpi offers comprehensive functionality for the analysis of scRNA-seq data. Key aspects are on the one hand its flexibility and ease of use, which allows the application to various tissues and disease types. On the other hand, it provides a standardized and reproducible workflow that is suited for application in clinical settings and was already utilized in a clinical study [8,21]. Moreover, scAmpi facilitates *in-silico* drug candidate identification on the single-cell level, thereby directly accounting for disease heterogeneity in the design of optimal drug treatment. Finally, because of the modular Snakemake framework,

we foresee a continued extension and refinement of the pipeline and its open source code, also by the single-cell community.

Supporting information

S1 Text. Parameter setting and analysis call.
(DOCX)

S2 Text. Cell types and immune gene expression.
(DOCX)

S3 Text. Tumor heterogeneity.
(DOCX)

Acknowledgments

Membership of the Tumor Profiler consortium:

Rudolf Aebersold, Melike Ak, Faisal S Al-Quaddoomi, Jonas Albinus, Ilaria Alborelli, Sonali Andani, Per-Olof Attinger, Marina Bacac, Daniel Baumhoer, Beatrice Beck-Schimmer, Niko Beerenwinkel, Christian Beisel, Lara Bernasconi, Anne Bertolini, Bernd Bodenmiller, Ximena Bonilla, Lars Bosshard, Byron Calgua, Ruben Casanova, Stéphane Chevrier, Natalia Chicherova, Maya D'Costa, Esther Danenberg, Natalie Davidson, Monica-Andreea Drăgan, Reinhard Dummer, Stefanie Engler, Martin Erkens, Katja Eschbach, Cinzia Esposito, André Fedier, Pedro Ferreira, Joanna Ficek, Anja L Frei, Bruno Frey, Sandra Goetze, Linda Grob, Gabriele Gut, Detlef Günther, Martina Haberecker, Pirmin Haeuptle, Viola Heinzelmann-Schwarz, Sylvia Herter, Rene Holtackers, Tamara Huesser, Anja Irmisch, Francis Jacob, Andrea Jacobs, Tim M Jaeger, Katharina Jahn, Alva R James, Philip M Jermann, André Kahles, Abdullah Kahraman, Viktor H Koelzer, Werner Kuebler, Jack Kuipers, Christian P Kunze, Christian Kurzeder, Kjong-Van Lehmann, Mitchell Levesque, Sebastian Lugert, Gerd Maass, Markus G Manz, Philipp Markolin, Julien Mena, Ulrike Menzel, Julian M Metzler, Nicola Miglino, Emanuela S Milani, Holger Moch, Simone Muenst, Riccardo Murri, Charlotte KY Ng, Stefan Nicolet, Marta Nowak, Patrick GA Pedrioli, Lucas Pelkmans, Salvatore Piscuoglio, Michael Prummer, Mathilde Ritter, Christian Rommel, María L Rosano-González, Gunnar Rätsch, Natascha Santacroce, Jacobo Sarabia del Castillo, Ramona Schlenker, Petra C Schwalie, Severin Schwan, Tobias Schär, Gabriela Senti, Franziska Singer, Sujana Sivapatham, Berend Snijder, Bettina Sobottka, Vipin T Sreedharan, Stefan Stark, Daniel J Stekhoven, Alexandre PA Theocharides, Tinu M Thomas, Markus Tolnay, Vinko Tosevski, Nora C Toussaint, Mustafa A Tuncel, Marina Tusup, Audrey Van Drogen, Marcus Vetter, Tatjana Vlajnic, Sandra Weber, Walter P Weber, Rebekka Wegmann, Michael Weller, Fabian Wendt, Norbert Wey, Andreas Wicki, Mattheus HE Wildschut, Bernd Wollscheid, Shuqing Yu, Johanna Ziegler, Marc Zimmermann, Martin Zoche, Gregor Zuend.

Author Contributions

Conceptualization: Anne Bertolini, Michael Prummer, Jack Kuipers, Niko Beerenwinkel, Franziska Singer.

Funding acquisition: Daniel Johannes Stekhoven, Niko Beerenwinkel.

Methodology: Anne Bertolini, Michael Prummer, Ulrike Menzel, Jack Kuipers, Franziska Singer.

Project administration: Niko Beerenwinkel, Franziska Singer.

Resources: Daniel Johannes Stekhoven, Niko Beerenwinkel.

Software: Anne Bertolini, Michael Prummer, Mustafa Anil Tuncel, María Lourdes Rosano-González, Franziska Singer.

Supervision: Jack Kuipers, Franziska Singer.

Visualization: Anne Bertolini, Michael Prummer, Franziska Singer.

Writing – original draft: Anne Bertolini, Michael Prummer, Franziska Singer.

Writing – review & editing: Anne Bertolini, Michael Prummer, Mustafa Anil Tuncel, Ulrike Menzel, María Lourdes Rosano-González, Jack Kuipers, Daniel Johannes Stekhoven, Niko Beerenwinkel, Franziska Singer.

References

1. Kulkarni A, Anderson AG, Merullo DP, Konopka G. Beyond bulk: a review of single cell transcriptomics methodologies and applications [Internet]. Vol. 58, Current Opinion in Biotechnology. 2019. p. 129–36. Available from: https://doi.org/10.1016/j.copbio.2019.03.001 PMID: 30978643
2. Zhu S, Qing T, Zheng Y, Jin L, Shi L. Advances in single-cell RNA sequencing and its applications in cancer research. Oncotarget. 2017 Aug 8; 8(32):53763–79. https://doi.org/10.18632/oncotarget.17893 PMID: 28881849
3. Tirosh I, Izar B, Prakadan SM, Wadsworth MH 2nd, Treacy D, Trombetta JJ, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science. 2016 Apr 8; 352 (6282):189–96. https://doi.org/10.1126/science.aad0501 PMID: 27124452
4. Shih AJ, Menzin A, Whyte J, Lovecchio J, Liew A, Khalili H, et al. Identification of grade and origin specific cell populations in serous epithelial ovarian cancer by single cell RNA-seq. PLoS One. 2018 Nov 1; 13(11):e0206785. https://doi.org/10.1371/journal.pone.0206785 PMID: 30383866
5. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018 Jun; 36(5):411–20. https://doi.org/10.1038/nbt.4096 PMID: 29608179
6. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 2018 Feb 6; 19(1):15. https://doi.org/10.1186/s13059-017-1382-0 PMID: 29409532
7. Mohanraj S, Díaz-Mejía JJ, Pham MD, Elrick H, Husić M, Rashid S, et al. CReSCENT: CanceR Single Cell ExpressioN Toolkit. Nucleic Acids Res. 2020 Jul 2; 48(W1):W372–9. https://doi.org/10.1093/nar/gkaa437 PMID: 32479601
8. Gardeux V, David FPA, Shajkofci A, Schwalie PC, Deplancke B. ASAP: a web-based platform for the analysis and interactive visualization of single-cell RNA-seq data [Internet]. Vol. 33, Bioinformatics. 2017. p. 3123–5. Available from: https://doi.org/10.1093/bioinformatics/btx337 PMID: 28541377
9. Peng RD. Reproducible research and Biostatistics. Biostatistics. 2009 Jul; 10(3):405–8. https://doi.org/10.1093/biostatistics/kxp014 PMID: 19535325
10. Moher D, Avey M, Antes G, Altman DG. The National Institutes of Health and guidance for reporting preclinical research. BMC Med. 2015 Feb 17; 13:34. https://doi.org/10.1186/s12916-015-0284-9 PMID: 25775278
11. Köster J, Rahmann S. Snakemake-a scalable bioinformatics workflow engine. Bioinformatics. 2018 Oct 15; 34(20):3600. https://doi.org/10.1093/bioinformatics/bty350 PMID: 29788404
12. Irmisch Anja, et al. "The Tumor Profiler Study: Integrated, multi-omic, functional tumor profiling for clinical decision support." *medRxiv* (2020), https://doi.org/10.1101/2020.02.13.20017921.
13. Irmisch A, Bonilla X, Chevrier S, Lehmann K-V, Singer F, Toussaint NC, et al. The Tumor Profiler Study: integrated, multi-omic, functional tumor profiling for clinical decision support. Cancer Cell. 2021 Mar 8; 39(3):288–93. https://doi.org/10.1016/j.ccell.2021.01.004 PMID: 33482122
14. Amezquita RA, Lun ATL, Becht E, Carey VJ, Carpp LN, Geistlinger L, et al. Orchestrating single-cell analysis with Bioconductor. Nat Methods. 2020 Feb; 17(2):137–45. https://doi.org/10.1038/s41592-019-0654-x PMID: 31792435
15. Germain P-L, Lun A, Macnair W, Robinson MD. Doublet identification in single-cell sequencing data using scDblFinder [Internet]. Vol. 10, F1000Research. 2021. p. 979. Available from: http://dx.doi.org/10.12688/f1000research.73600.1.

16. Ilicic T, Kim JK, Kolodziejczyk AA, Bagger FO, McCarthy DJ, Marioni JC, et al. Classification of low quality cells from single-cell RNA-seq data. Genome Biol. 2016 Feb 17; 17:29. https://doi.org/10.1186/s13059-016-0888-1 PMID: 26887813

17. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol. 2019 Dec 23; 20(1):296. https://doi.org/10.1186/s13059-019-1874-1 PMID: 31870423

18. Levine JH, Simonds EF, Bendall SC, Davis KL, Amir E-AD, Tadmor MD, et al. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. Cell. 2015 Jul 2; 162 (1):184–97. https://doi.org/10.1016/j.cell.2015.05.047 PMID: 26095251

19. Prummer M, Bertolini A, Bosshard L, Barkmann F, Yates J, Boeva V, et al. scROSHI—robust supervised hierarchical identification of single cells [Internet]. Available from: http://dx.doi.org/10.1101/2022.04.05.487176

20. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. 2018 Feb 9 [cited 2022 Apr 28]; Available from: http://dx.doi.org/10.48550/arXiv.1802.03426.

21. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015 Apr 20; 43(7):e47. https://doi.org/10.1093/nar/gkv007 PMID: 25605792

22. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics. 2013 Jan 16; 14:7. https://doi.org/10.1186/1471-2105-14-7 PMID: 23323831

23. Singer F, Irmisch A, Toussaint NC, Grob L, Singer J, Thurnherr T, et al. SwissMTB: establishing comprehensive molecular cancer diagnostics in Swiss clinics. BMC Med Inform Decis Mak. 2018 Oct 29; 18 (1):89. https://doi.org/10.1186/s12911-018-0680-0 PMID: 30373609

24. Cotto KC, Wagner AH, Feng Y-Y, Kiwala S, Coffman AC, Spies G, et al. DGIdb 3.0: a redesign and expansion of the drug-gene interaction database. Nucleic Acids Res. 2018 Jan 4; 46(D1):D1068–73. https://doi.org/10.1093/nar/gkx1143 PMID: 29156001

25. Griffith M, Spies NC, Krysiak K, McMichael JF, Coffman AC, Danos AM, et al. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. Nat Genet. 2017 Jan 31; 49(2):170–4. https://doi.org/10.1038/ng.3774 PMID: 28138153

26. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. 2015 Dec 23; 1(6):417–25. https://doi.org/10.1016/j.cels.2015.12.004 PMID: 26771021