

12 Grand Challenges in Single-Cell Data Science

David Lähnemann^{*1,2,3}, Johannes Köster^{*,+,1,4}, Ewa Szczurek^{*,5}, Davis J. McCarthy^{*,6,7}, Stephanie C. Hicks^{*,8}, Mark D. Robinson^{*,9}, Catalina A. Vallejos^{*,10,11}, Niko Beerenwinkel^{*,12,13}, Kieran R. Campbell^{*,15,16,17}, Ahmed Mahfouz^{*,18,19}, Luca Pinello^{*,20,21,22}, Pavel Skums^{*,23}, Alexandros Stamatakis^{*,24,25}, Camille Stephan-Otto Attolini^{*,26}, Samuel Aparicio^{16,27}, Jasmijn Baaijens²⁹, Marleen Balvert^{29,31}, Buys de Barbanson^{32,33,34}, Antonio Cappuccino³⁵, Giacomo Corleone³⁶, Bas Dutilh^{31,38}, Maria Florescu^{32,33,34}, Victor Gurjev⁴¹, Rens Holmer⁴², Katharina Jahn^{12,13}, Thamar Jessurun Lobo⁴¹, Emma M. Keizer⁴⁵, Indu Khatri⁴⁶, Szymon M. Kielbasa⁴⁷, Jan O. Korbel⁴⁸, Alexey M. Kozlov²⁴, Tzu-Hao Kuo³, Boudewijn P.F. Lelieveldt^{49,50}, Ion I. Mandoiu⁵¹, John C. Marioni^{52,53,54}, Tobias Marschall^{55,56}, Felix Mölder^{1,59}, Amir Niknejad^{60,61}, Łukasz Rączkowski⁵, Marcel Reinders^{18,19}, Jeroen de Ridder^{32,33}, Antoine-Emmanuel Saliba⁶², Antonios Somarakis⁵⁰, Oliver Stegle^{48,54,63}, Fabian J. Theis⁶⁷, Huan Yang⁶⁸, Alex Zelikovskiy^{69,70}, Alice C. McHardy^{+,3}, Benjamin J. Raphael^{+,71}, Sohrab P. Shah^{+,72}, and Alexander Schönhuth^{@,+,*,29,31}

^{*} Joint first authors, major contributions to manuscript.

- ¹Algorithms for Reproducible Bioinformatics, Genome Informatics, Institute of Human Genetics, University Hospital Essen, University of Duisburg-Essen, Germany
- ²Department of Paediatric Oncology, Haematology and Immunology, Medical Faculty, Heinrich Heine University, University Hospital, Düsseldorf, Germany
- ³Computational Biology of Infection Research Group, Helmholtz Centre for Infection Research, Braunschweig, Germany
- ⁴Medical Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, USA
- ⁵Institute of Informatics, Faculty of Mathematics, Informatics and Mechanics University of Warsaw, Poland
- ⁶Bioinformatics and Cellular Genomics, St Vincent's Institute of Medical Research, Fitzroy, Australia
- ⁷Melbourne Integrative Genomics, School of BioSciences — School of Mathematics & Statistics, Faculty of Science, University of Melbourne, Australia
- ⁸Department of Biostatistics, Johns Hopkins University, Baltimore, MD, USA
- ⁹Institute of Molecular Life Sciences and SIB Swiss Institute of Bioinformatics, University of Zurich, Switzerland
- ¹⁰MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, UK
- ¹¹The Alan Turing Institute, British Library, London, UK
- ¹²Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland
- ¹³SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland
- ¹⁵Department of Statistics, University of British Columbia, Vancouver, Canada
- ¹⁶Department of Molecular Oncology, BC Cancer Agency, Vancouver, Canada
- ¹⁷Data Science Institute, University of British Columbia, Vancouver, Canada
- ¹⁸Leiden Computational Biology Center, Leiden University Medical Center, The Netherlands
- ¹⁹Delft Bioinformatics Lab, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, The Netherlands
- ²⁰Molecular Pathology Unit and Center for Cancer Research, Massachusetts General Hospital Research Institute, Charlestown, USA
- ²¹Department of Pathology, Harvard Medical School, Boston, USA
- ²²Broad Institute of Harvard and MIT, Cambridge, MA, USA
- ²³Department of Computer Science, Georgia State University, Atlanta, USA
- ²⁴Computational Molecular Evolution Group, Heidelberg Institute for Theoretical Studies, Germany
- ²⁵Institute for Theoretical Informatics, Karlsruhe Institute of Technology, Germany
- ²⁶Institute for Research in Biomedicine, The Barcelona Institute of Science and Technology, Spain
- ²⁷Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, Canada
- ²⁹Life Sciences and Health, Centrum Wiskunde & Informatica, Amsterdam, The Netherlands
- ³¹Theoretical Biology and Bioinformatics, Science for Life, Utrecht University, The Netherlands
- ³²Center for Molecular Medicine, University Medical Center Utrecht, The Netherlands
- ³³Onco Institute, Utrecht, The Netherlands
- ³⁴Quantitative biology, Hubrecht Institute, Utrecht, The Netherlands
- ³⁵Institute for Advanced Study, University of Amsterdam, The Netherlands
- ³⁶Department of Surgery and Cancer, The Imperial Centre for Translational and Experimental Medicine, Imperial College London, UK
- ³⁸Centre for Molecular and Biomolecular Informatics, Radboud University Medical Center, Nijmegen, The Netherlands
- ⁴¹European Research Institute for the Biology of Ageing, University Medical Center Groningen, University of Groningen, The Netherlands
- ⁴²Bioinformatics Group, Wageningen University, The Netherlands
- ⁴⁵Biometris, Wageningen University & Research, The Netherlands
- ⁴⁶Department of Immunohematology and Blood Transfusion, Leiden University Medical Center, The Netherlands
- ⁴⁷Department of Biomedical Data Sciences, Leiden University Medical Center, The Netherlands
- ⁴⁸Genome Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany
- ⁴⁹PRB lab, Delft University of Technology, The Netherlands
- ⁵⁰Division of Image Processing, Department of Radiology, Leiden University Medical Center, The Netherlands
- ⁵¹Computer Science & Engineering Department, University of Connecticut, Storrs, USA
- ⁵²Cancer Research UK Cambridge Institute, Li Ka Shing Centre, University of Cambridge, UK
- ⁵³Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, UK
- ⁵⁴European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, UK
- ⁵⁵Center for Bioinformatics, Saarland University, Saarbrücken, Germany
- ⁵⁶Max Planck Institute for Informatics, Saarbrücken, Germany
- ⁵⁹Institute of Pathology, University Hospital Essen, University of Duisburg-Essen, Germany.
- ⁶⁰Computation molecular design, Zuse Institute Berlin, Germany
- ⁶¹Mathematics department, Mount Saint Vincent, New York, USA
- ⁶²Helmholtz Institute for RNA-based Infection Research, Helmholtz-Center for Infection Research, Würzburg, Germany
- ⁶³Division of Computational Genomics and Systems Genetics, German Cancer Research Center – DKFZ, Heidelberg, Germany
- ⁶⁷Institute of Computational Biology, Helmholtz Zentrum München – German Research Center for Environmental Health, Neuherberg, Germany
- ⁶⁸Division of Drug Discovery and Safety, Leiden Academic Center for Drug Research – LACDR — Leiden University, The Netherlands
- ⁶⁹Department of Computer Science, Georgia State University, Atlanta, USA
- ⁷⁰The Laboratory of Bioinformatics, I.M. Sechenov First Moscow State Medical University, Moscow, Russia
- ⁷¹Department of Computer Science, Princeton University, USA
- ⁷²Computational Oncology, Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, USA

⁺ Joint last authors, workshop organizers.

[@] Corresponding author: Alexander Schönhuth, as@cw.nl

The recent upswing of microfluidics and combinatorial indexing strategies, further enhanced by very low sequencing costs, have turned single cell sequencing into an empowering technology; analyzing thousands—or even millions—of cells per experimental run is becoming a routine assignment in laboratories worldwide. As a consequence, we are witnessing a data revolution in single cell biology. Although some issues are similar in spirit to those experienced in bulk sequencing, many of the emerging data science problems are unique to single cell analysis; together, they give rise to the new realm of 'Single Cell Data Science'.

Here, we outline twelve challenges that will be central in bringing this new field forward. For each challenge, the current state of the art in terms of prior work is reviewed, and open problems are formulated, with an emphasis on the research goals that motivate them.

This compendium is meant to serve as a guideline for established researchers, newcomers and students alike, highlighting interesting and rewarding problems in 'Single Cell Data Science' for the coming years.

Contents

1	Introduction	3
2	Single Cell Data Science: Themes and Categories	4
2.1	Varying levels of resolution . . .	5
2.2	Quantifying uncertainty of measurements and analysis results	5
2.3	Scaling to higher dimensionalities: more cells, more features, broader coverage	6
2.4	Challenge categories	7

3	Challenges in single-cell transcriptomics	7
3.1	Challenge I: Handling sparsity in single-cell RNA sequencing . . .	7
3.2	Challenge II: Defining flexible statistical frameworks for discovering complex differential patterns in gene expression . . .	13
3.3	Challenge III: Mapping single cells to a reference atlas	15
3.4	Challenge IV: Generalizing trajectory inference	16
3.5	Challenge V: Finding patterns in spatially resolved measurements	18
4	Challenges in single-cell genomics	19
4.1	Challenge VI: Improving single-cell DNA sequencing data quality and scaling to more cells	20
4.2	Challenge VII: Errors and missing data in the identification of features / variation from single-cell DNA sequencing data.	23
5	Challenges in single-cell phylogenomics	25
5.1	Challenge VIII: Scaling phylogenetic models to many cells and many sites	26
5.2	Challenge IX: Integrating multiple types of features / variation into phylogenetic models . . .	27
5.3	Challenge X: Inferring population genetic parameters of tumor heterogeneity by model integration	29

1	6 Overarching challenges	32
2	6.1 Challenge XI: Integration of	
3	single-cell data: across sam-	
4	ples, experiments and types of	
5	measurement	32
6	6.2 Challenge XII: Validating and	
7	benchmarking analysis tools	
8	for single-cell measurements . .	38
9	7 Acknowledgements	40

1 Introduction

Since being elevated to “Method of the Year” in 2013 [Nature Methods, 2013], sequencing of the genetic material of individual cells has become routine when investigating cell-to-cell heterogeneity. Single-cell measurements of both RNA and DNA, and more recently also of epigenetic marks and protein levels, can stratify cells at the finest resolution possible.

Single-cell RNA sequencing (scRNA-seq) facilitates to distinguish cell states within coarser cell type clusters [for an early example, see Anchang et al., 2016], thereby arranging populations of cells according to novel types of hierarchies. It is also possible to identify cells in transition between states, so we get a much clearer view on the dynamics of tissue and organism development, and on structures within cell populations that had so far been perceived as homogeneous. Along a similar vein, analyses based on single-cell DNA sequencing (scDNA-seq) can highlight somatic clonal structures [e.g. in cancer, see Francis et al., 2014, Lawson et al., 2018] and are thus helpful for tracking the formation of certain cell lineages and to provide insight into evolutionary processes acting on somatic mutations.

The opportunities arising from single-cell sequencing (sc-seq) are enormous: only now is it possible to re-evaluate hypotheses about

differences between pre-defined sample groups at the single-cell level—no matter if such sample groups are disease subtypes, treatment groups or simply morphologically different cell types. It is therefore no surprise that the enthusiasm about the possibility to screen the genetic material of the basic units of life has been continuing to grow: a prominent example is the Human Cell Atlas [Regev et al., 2017], an initiative aiming to map the different types and states of cells that a human being is composed of, or Zhang and Liu [2019], as a most recent example of a list of single-cell analysis based opportunities in particular domains such as the blood, the brain and the lung.

Encouraged by the great potential of investigating DNA and RNA at the single-cell level, the development of the corresponding experimental technologies has experienced massive boosts. This upswing of high-throughput sc-seq technologies—most importantly in microfluidics techniques and combinatorial indexing strategies [Zilionis et al., 2017, Vitak et al., 2017, Svensson et al., 2018b, Luo et al., 2019, Gao et al., 2019]—means that tens or hundreds of thousands of cells, instead of just tens or hundreds, are routinely sequenced in one experiment; a development—further fueled by in the meantime low sequencing costs—that has recently even led to a publication on millions of cells in one experiment [Cao et al., 2019a]. As a consequence, primary and secondary sc-seq results of very large numbers of single cells are becoming available worldwide, constituting a data revolution for the field of single-cell analysis.

These vast amounts of data and the research hypotheses that motivate them, need to be handled in a computationally efficient and statistically sound manner. As these aspects clearly match a recent definition of “Data Science” [Hicks and Peng, 2019], we

1 posit that we have entered the era of Single
2 Cell Data Science (SCDS).

3 While SCDS faces many of the data sci-
4 ence issues arising in bulk sequencing, it also
5 substantially adds to them and further com-
6 pounds existing scientific challenges. Namely,
7 limited amounts of material available per cell
8 lead to exceptionally high levels of uncer-
9 tainty about (possibly missed) observations,
10 and where amplification is used to generate
11 more material, technical noise is added to the
12 resulting data. Further, a new level of resolu-
13 tion also means another—rapidly growing—
14 dimension in data matrices, thus requiring
15 scalable models and methods for data anal-
16 ysis. While the particular challenges can vary
17 greatly by research goal, tissue analyzed, ex-
18 perimental setup or—last but not least—just
19 by whether DNA or RNA is sequenced, fur-
20 ther factoring into various protocols, assaying
21 for example also the epigenome (bisulfite pro-
22 tocols), chromatin accessibility (e.g. ATAC-
23 seq) or protein levels (CITE-seq), the com-
24 mon denominator is that the challenges are
25 all rooted in data science, hence are compu-
26 tational or statistical in nature. Here, we pro-
27 pose the dozen data science challenges that we
28 believe to be most relevant for bringing SCDS
29 forward. We summarize and categorize them,
30 providing a thorough review of the status of
31 each challenge relative to existing approaches.
32 From this foundation, we point to possible di-
33 rections of research to tackle them. This cata-
34 logue of SCDS challenges aims at focusing the
35 development of data analysis methods and the
36 directions of research in this rapidly evolving
37 field—as a guideline for researchers looking
38 for rewarding problems that match their per-
39 sonal expertise and interests.

2 Single Cell Data Science: Themes and Categories

A number of challenging themes are common
to all single-cell analyses, regardless of the
particular assay or data modality generated.
We will start our review by broadly categoriz-
ing these aspects. Later, when discussing the
specific 12 challenges, we will refer to these
broader categories wherever appropriate and,
if this is sensible, lay out what these broader
theme issues mean in the particular context.
If challenges covered in later sections are par-
ticularly entangled with the broader themes
listed here, we will also refer to them from
within this section.

These elementary themes may reflect issues
one also experiences when analyzing bulk se-
quencing data. However, even if not unique
to single-cell experiments, these issues may
become particularly dominant in the analysis
of sc-seq data and therefore require particu-
lar attention. The most driving of such el-
ementary themes, not necessarily unique to
sc-seq, are: (i) The need to quantify mea-
surement uncertainty (see challenges in sec-
tion 2.2) (ii) The need to benchmark methods
systematically, in a way that highlights the
metrics that are particularly critical in sc-seq
(section 6.2). The most driving themes spe-
cific to sc-seq, exacerbated by the rapid ad-
vances in terms of experimental technologies
supporting single-cell analyses, are: (i) The
need to scale to higher dimensional data, be
it more cells measured or more data mea-
sured per cell (section 2.3); this often arises
in combination with: (ii) The need to inte-
grate data across different types of single-
cell measurements (e.g. RNA, DNA, proteins,
methylation and so on) and across samples,
be they from different time points, treatment
groups or even organisms (section 6.1). Fi-
nally, the possibility to operate on the finest

levels of resolution casts an important, overarching question: (iii) Which exact level of resolution is appropriate relative to the particular research question one has in mind (section 2.1)? We will start by qualifying this last one.

2.1 Varying levels of resolution

Sc-seq allows for a fine-grained definition of cell types and states. Hence it allows for characterizations of cell populations that are significantly more detailed than characterizations supported by bulk sequencing experiments. However, even though sc-seq operates at the most basic level, mapping cell types and states at a particular level of resolution of interest may be challenging: Depending on whether the research question allows for a certain freedom in terms of resolution, and depending on the limits imposed by the particular experimental setup, achieving the targeted level of resolution or granularity for the intended map of cells may require substantial methodological efforts.

When drawing maps of cell types and states, it is important that they: (i) have a structure that recapitulates both tissue development and tissue organization; (ii) account for continuous cell states in addition to discrete cell types (i.e. reflecting cell state trajectories within cell types and smooth transitions between cell types, as observed in tissue generation); (iii) allow for choosing the level of resolution flexibly (i.e. the map should possibly support zoom type operations, to let the researcher choose the desired level of granularity with respect to cell types and states conveniently, ranging from whole organisms via tissues to cell populations and cellular subtypes); (iv) include biological and functional annotation wherever available and helpful in the intended functional context.

An exemplary illustration of how maps of

cell types and states can support different levels of resolution are the structure-rich topologies generated by PAGA based on scRNA-seq [Wolf et al., 2019], see Figure 1 for an illustration¹. At the highest levels of resolution, these topologies also reflect intermediate cell states and the developmental trajectories passing through them. A similar approach that also allows for consistently zooming into more detailed levels of resolution is provided by hierarchical stochastic neighbor embedding (HSNE, Pezzotti et al. [2016]), a method pioneered on mass cytometry data sets [Unen et al., 2017, Höllt et al., 2018]. In addition, manifold learning [Welch et al., 2017, Moon et al., 2018] and metric learning [Hoffer and Ailon, 2015, Bromley et al., 1993] may provide further theoretical support for even more accurate maps, because they provide sound theories about reasonable, continuous distance metrics, instead of just distinct, discrete clusters.

2.2 Quantifying uncertainty of measurements and analysis results

The amount of material sampled from single cells is considerably less in comparison with the amounts of material raised in bulk experiments, because the latter are based on examining the DNA or RNA of larger pools of cells together. Signals become more stable when individual signals are summarized (such as in a bulk experiment), thus the increase in resolution due to sc-seq also means a reduction of the stability of the supporting signals. The reduction in signal stability, in turn, implies that data becomes substantially more

¹Figure 1 was adapted from Wolf et al. [2019], Fig. 3, provided under Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>).

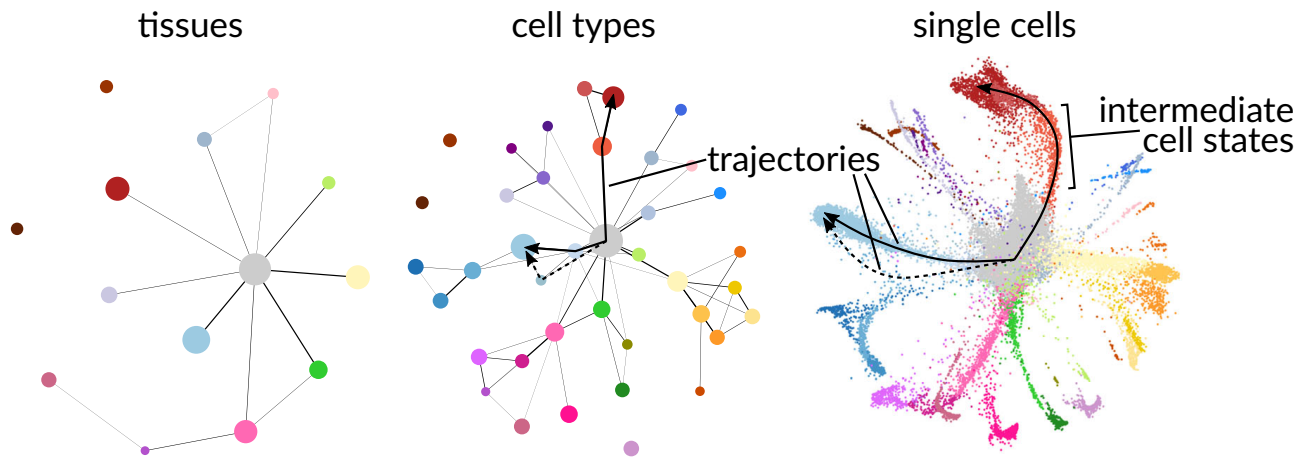


Figure 1: Different levels of resolution are of interest, depending on the research question and the data available. Thus, analysis tools and reference systems (such as cell atlases) will have to accommodate for multiple levels of resolution from whole organs and tissues over discrete cell types to continuously mappable intermediate cell states, indistinguishable even at the microscopic level. A graph abstraction that enables such multiple levels of focus is provided by PAGA [Wolf et al., 2019], a structure that allows for discretely grouping cells, as well as inferring trajectories as paths through a graph.

uncertain and tasks hitherto considered routine, such as single nucleotide variation (SNV) calling in bulk sequencing, require considerable methodological care to be resolved also for sc-seq.

These issues with data quality and in particular missing data pose challenges that are novel and unique to sc-seq, and are thus at the core of several challenges: regarding scDNA-seq data quality (see challenges in section 4.1) and especially regarding missing data in scDNA-seq (section 4.2) and scRNA-seq (section 3.1). In contrast, the non-negligible batch effects that scRNA-seq can suffer from reflect a common problem in high-throughput data analysis [Leek et al., 2010], and thus are not discussed here (although in certain protocols such effects can be alleviated by careful use of negative control data in the form of spike-in RNA of known content and concentration [Severson et al., 2018, BEARsc]).

Optimally, sc-seq analysis tools would accu-

rately quantify all uncertainties arising from experimental errors and biases. Thereby, these tools would prevent the uncertainties from propagating to the intended downstream analyses in an uncontrolled manner, and rather translate them into statistically sound and accurately quantified qualifiers of final results.

2.3 Scaling to higher dimensionalities: more cells, more features, broader coverage

The current blossoming of experimental methods poses considerable statistical challenges, and would do even if measurements were not affected by errors and biases.

The increase in the number of single cells analyzed per experiment translates into more data points being generated, requiring methods to scale rapidly. With scRNA-seq already

scaling to millions of cells, some of the respective methodology has picked up the thread [Sengupta et al., 2016, Sinha et al., 2018, Wolf et al., 2018, Iacono et al., 2018]. Of course, the respective issues have not yet been fully resolved; further improvements are conceivable. For scDNA-seq, experimental methodology has just been scaling up to more cells recently (see section 4.1 and section 5.1), making this a pressing challenge in the development of data analysis methods.

Beyond basic scRNA-seq and scDNA-seq experiments, various assays have been proposed to measure chromatin accessibility [Buenrostro et al., 2015, Cusanovich et al., 2015], DNA methylation [Karemaker and Vermeulen, 2018], protein levels [Virant-Klun et al., 2016], protein binding, and also for performing multiple simultaneous measurements [Clark et al., 2018, Cao et al., 2018] in single cells. The corresponding increase in experimental choices means another possible inflation of feature spaces.

In parallel to the increase in the number of cells queried and the number of different assays possible, the increase of the resolution per cell of specific measurement types causes a steady increase of the dimensionality of corresponding data spaces. For the field of SCDS this amounts to a severe and recurring case of the “curse of dimensionality” for all types of measurements. Here again, scRNA-seq based methods are in the lead when trying to deal with feature dimensionality, while scDNA-seq based methodology (which includes epigenome assays) has yet to catch up.

Finally, there are efforts to measure multiple feature types in parallel, e.g. from scDNA-seq (see section 5.2). Also, with spatial and temporal sampling becoming available (see section 3.5 and section 5.3), data integration methods need to scale to more and new types of context information for individual cells (see

section 6.1 for a comprehensive discussion of data integration approaches).

2.4 Challenge categories

All challenges we identified fall into at least one of three greater categories: transcriptomics (section 3), genomics (section 4) and phylogenomics (section 5). Here, the separation of phylogenomics from genomics is due to the distinct research goals the respective challenges address. Last but not least, two challenges are relevant to all of these categories, and are thus discussed as recapitulatory challenges at the end: the data integration challenge (section 6.1) draws on the types of measurements and experiments described in the category-specific challenges. The benchmarking challenge (presented in section 6.2), although being essential in many areas of data science, is worth highlighting here in particular, because benchmarking for SCDS is still in its infancy.

3 Challenges in single-cell transcriptomics

3.1 Challenge I: Handling sparsity in single-cell RNA sequencing

A comprehensive characterization of the transcriptional status of individual cells enables us to gain full insight into the interplay of transcripts within single cells. However, scRNA-seq measurements typically suffer from large fractions of observed zeros, where a given gene in a given cell has no unique molecule identifiers or reads mapping to it. These observed zero values can represent either missing data (i.e. a gene is expressed but not detected by the sequencing technology) or true absence of

expression. The proportion of zeros, or degree of sparsity, is thought to be due to imperfect reverse transcription and amplification, and other technical limitations (Hicks et al. [2018], Bacher and Kendzierski [2016]), and depends on the scRNA-seq platform used, the sequencing depth and the underlying expression level of the gene. The term “dropout” is often used to denote observed zero values in scRNA-seq data, but this term conflates zero values attributable to methodological noise and biologically-true zero expression, so we recommend against its use as a catch-all term for observed zeros.

Sparsity in scRNA-seq data can hinder downstream analyses, but it is challenging to model or handle it appropriately, and thus, there remains an ongoing need for improved methods. Sparsity pervades all aspects of scRNA-seq data analysis, but here we focus on the linked problems of learning latent spaces and “imputing” expression values from scRNA-seq data (Figure 2). Imputation, “data smoothing” and “data reconstruction” approaches are closely linked to the challenges of normalisation. But whereas normalisation generally aims to make expression values between cells more comparable to each other, imputation and data smoothing approaches aim to achieve adjusted data values that—it is hoped—better represent the true expression values. Imputation methods could therefore be used for normalisation, but do not entail all possible or useful approaches to normalisation.

3.1.1 Status

The imputation of missing values has been very successful for genotype data. Crucially, when imputing genotypes we often know which data are missing (e.g. when no genotype call is possible due to no coverage of a locus, although see section 4.2 for

the challenges with scDNA-seq data) and rich sources of external information are available (e.g. haplotype reference panels). Thus, genotype imputation is now highly accurate and a commonly-used step in data processing for genetic association studies [Das et al., 2018].

The situation is somewhat different for scRNA-seq data, as we do not routinely have external reference information to apply (see section 3.3). In addition, we can never be sure which observed zeros represent “missing data” and which accurately represent a true gene expression level in the cell [Hicks et al., 2018]. Observed zeros can either represent “biological” zeros, i.e. those present because the true expression level of a gene in a cell was zero. Or they are the result of methodological noise, which can arise when a gene has true non-zero expression in a cell, but no counts are observed due to failures at any point in the complicated process of processing mRNA transcripts in cells into mapped reads. Such noise can lead to artefactual zero that are either more systematic (e.g. sequence-specific mRNA degradation during cell lysis) or that occur by chance (e.g. barely expressed transcripts that at the same expression level will sometimes be detected and sometimes not, due to sampling variation, e.g in the sequencing). The high degree of sparsity in scRNA-seq data therefore arises from technical zeros and true biological zeros, which are difficult to distinguish from one another.

In general, two broad approaches can be applied to tackle this problem of sparsity: (i) use statistical models that inherently model the sparsity, sampling variation and noise modes of scRNA-seq data with an appropriate data generative model; or (ii) attempt to “impute” values for observed zeros (ideally the technical zeros; sometimes also non-zero values) that better approximate the true gene expression levels. We prefer to use the first option where possible, and for many single-cell data

analysis problems, statistical models appropriate for sparse count data exist and should be used (e.g. for differential expression analysis). However, there are many cases where the appropriate models are not available and accurate imputation of technical zeros would allow better results from downstream methods and algorithms that cannot handle sparse count data. For example, imputation could be particularly useful for many dimension reduction, visualisation and clustering applications. It is therefore desirable to improve both statistical methods that work on sparse count data directly and approaches for data imputation for scRNA-seq data, whether by refining existing techniques or developing new ones (see also section 2.2).

We define three broad (and sometimes overlapping) categories of methods that can be used to “impute” scRNA-seq data in the absence of an external reference: (i) *Model-based imputation methods of technical zeros* use probabilistic models to identify which observed zeros represent technical rather than biological zeros and aim to impute expression levels just for these technical zeros, leaving other observed expression levels untouched; or (ii) *Data-smoothing methods* define sets of “similar” cells (e.g. cells that are neighbours in a graph or occupy a small region in a latent space) and adjust expression values for each cell based on expression values in similar cells. These methods adjust all expression values, including technical zeros, biological zeros and observed non-zero values. (iii) *Data-reconstruction methods* typically aim to define a latent space representation of the cells. This is often done through matrix factorization (e.g. principal component analysis) or, increasingly, through machine learning approaches (e.g. variational autoencoders that exploit deep neural networks to capture non-linear relationships). Although a broad class of methods, both ma-

trix factorization methods and autoencoders (among others) are able to “reconstruct” the observed data matrix from low-rank or simplified representations. The reconstructed data matrix will typically no longer be sparse (with many zeros) and the implicitly “imputed” data can be used for downstream applications that cannot handle sparse count data.

The first category of methods generally seeks to infer a probabilistic model that captures the data generation mechanism. Such generative models can be used to identify, probabilistically, which observed zeros correspond to technical zeros (to be imputed) and which correspond to biological zeros (to be left alone). There are many model-based imputation methods already available that use ideas from clustering (e.g. k-means), dimension reduction, regression and other techniques to impute technical zeros, oftentimes combining ideas from several of these approaches. These include SAVER [Huang et al., 2018], ScImpute [Li and Li, 2018], bayNorm [Tang et al., 2018], scRecover [Miao et al., 2019], and VIPER [Chen and Zhou, 2018]. Clustering methods that implicitly impute values, such as CIDR [Lin et al., 2017b] and BISCUIT [Azizi et al., 2017], are closely related to this class of imputation methods.

Data-smoothing methods, which adjust all gene expression levels based on expression levels in “similar” cells, have also been proposed to handle imputation problems. We might regard these approaches as “denoising” methods. To take a simplified example (Figure 2), we might imagine that single cells originally refer to points in two-dimensional space, but are likely to describe a one-dimensional curve; projecting data points onto that curve eventually allows imputation of the “missing” values (but all points are adjusted, or smoothed, not just true technical zeros). Prominent data-smoothing ap-

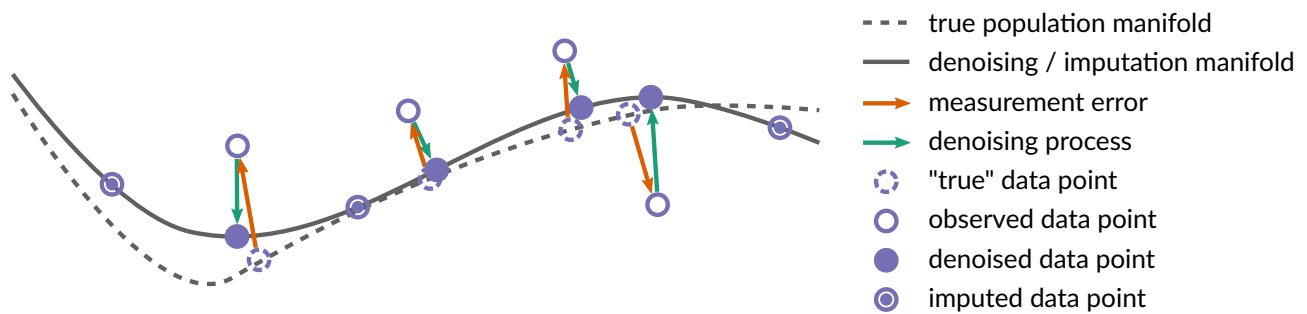


Figure 2: Measurement error requires denoising methods or approaches that quantify uncertainty and propagate it down analysis pipelines. Also, whenever methods cannot deal with the abundant missing values, imputation approaches are necessary. Whereas the true population manifold that generated data is never known, one can usually obtain some estimation of it that can be used for both denoising and imputation.

- 1 approaches to handling sparse counts include:
- 2 • diffusion-based MAGIC [Dijk et al., 2018]
 - 3
 - 4 • k-nearest neighbor-based knn-smooth
 - 5 [Wagner et al., 2018b]
 - 6 • network diffusion-based netSmooth
 - 7 [Jonathan Ronen, 2018]
 - 8 • clustering-based DrImpute [Gong et al., 2018]
 - 9
 - 10 • locality sensitive imputation in LSIIm-
 - 11 pute [Moussa and Măndoiu, 2019]

12 A major task in the analysis of high-
 13 dimensional single-cell data is to find low-
 14 dimensional representations of the data that
 15 capture the salient biological signals and ren-
 16 der the data more interpretable and amenable
 17 to further analyses. As it happens, the ma-
 18 trix factorization and latent-space learning
 19 methods used for that task also provide an-
 20 other route for imputation through their abil-
 21 ity to *reconstruct* the observed data matrix
 22 from simplified representations of it. Prin-
 23 cipal component analysis (PCA) is one such
 24 standard matrix factorization method that
 25 can be applied to scRNA-seq data (preferably

after suitable data normalisation) as are other
 widely-used general statistical methods like
 independent component analysis (ICA) and
 non-negative matrix factorization (NMF). As
 (linear) matrix factorization methods, PCA,
 ICA and NMF decompose the observed data
 matrix into a “small” number of factors in two
 low-rank matrices, one representing cell-by-
 factor weights and one gene-by-factor load-
 ings. Many matrix factorization methods
 with tweaks for single-cell data have been pro-
 posed in recent years, including:

- ZIFA, a zero-inflated factor analysis [Pierson and Yau, 2015]
- f-scLVM, a sparse Bayesian latent variable model [Buettner et al., 2017]
- GPLVM, a Gaussian process latent variable model [Verma and Engelhardt, 2018]
- ZINB-WaVE, a zero-inflated negative binomial factor model [Risso et al., 2018]
- scCoGAPS, an extension of NMF [Stein-O’Brien et al., 2019]
- consensus NMF, a meta-analysis approach to NMF [Kotliar et al., 2019]

1	• pCMF, probabilistic count matrix factor-	39
2	ization with a Poisson model [Durif et al.,	40
3	2019]	41
4	• SDA, sparse decomposition of arrays;	
5	another sparse Bayesian method [Jung	42
6	et al., 2019].	
7	Some data reconstruction approaches have	
8	been specifically proposed for imputation, in-	
9	cluding:	
10	• ENHANCE, denoising PCA with an ag-	
11	gregation step [Wagner et al., 2019]	43
12	• ALRA, SVD with adaptive thresholding	
13	[Linderman et al., 2018]	44
14	• scRMD, robust matrix decomposition	
15	[Chen et al., 2018]	45
16	Recently, machine learning methods have	
17	emerged that apply autoencoders [AutoIm-	
18	pute, Talwar et al., 2018] and deep neu-	
19	ral networks [DeepImpute, Arisdakessian	
20	et al., 2018]) or ensemble learning [EnImpute,	
21	Zhang et al., 2019c]) to impute expression val-	
22	ues.	
23	Additionally, many deep learning methods	
24	have been proposed for single-cell data anal-	
25	ysis that can, but need not, use probabilis-	
26	tic data generative processes to capture low-	
27	dimensional or latent space representations of	
28	a dataset. Even if imputation is not a main	
29	focus, such methods can generate “imputed”	
30	expression values as an upshot of a model pri-	
31	marily focused on other tasks like learning la-	
32	tent spaces, clustering, batch correction, or	
33	visualization (and often several of these tasks	
34	simultaneously). The latter set includes tools	
35	such as:	
36	• DCA, an autoencoder with a zero-	
37	inflated negative binomial distribution	
38	[Eraslan et al., 2019]	46
	• scVI, a variational autoencoder with a	47
	zero-inflated negative binomial model	48
	[Lopez et al., 2018]	49
	• LATE [Badsha et al., 2018]	50
	• VASC [Wang and Gu, 2018]	51
	• compscVAE [Grønbech et al., 2018]	52
	• scScope [Deng et al., 2019]	53
	• Tybalt [Way and Greene, 2018]	54
	• SAUCIE [Amodio et al., 2019]	55
	• scvis [Ding et al., 2018]	56
	• net-SNE [Cho et al., 2018]	57
	• BERMUDA, focused on batch correction	58
	[Wang et al., 2019]	59
	• DUSC [Srinivasan et al., 2019]	60
	• Expression Saliency [Kinalis et al., 2019]	61
	• others [Lin et al., 2017a, Zhang, 2019]	62
	Besides the three categories described	63
	above, a small number of scRNA-seq impu-	64
	tation methods have been developed to in-	65
	corporate information external to the cur-	66
	rent dataset for imputation. These include:	67
	ADImpute [Leote et al., 2019], which uses	68
	gene regulatory network information from	
	external sources; SAVER-X [Wang et al.,	
	2018], a transfer learning method for denois-	
	ing and imputation that can use informa-	
	tion from atlas-type resources; and methods	
	that borrow information from matched bulk	
	RNAseq data like URSM [Zhu et al., 2018]	
	and SCRABBLE [Peng et al., 2019].	

3.1.2 Open problems

A major challenge in this context is the circularity that arises when imputation solely relies on information that is internal to the imputed dataset. This circularity can artificially amplify the signal contained in the data, leading to inflated correlations between genes and/or cells. In turn, this can introduce false positives in downstream analyses such as differential expression testing and gene network inference [Andrews and Hemberg, 2019]. Handling batch effects and potential confounders requires further work to ensure that imputation methods do not mistake unwanted variation from technical sources for biological signal. In a similar vein, single-cell experiments are affected by various uncertainties (see section 2.2). Approaches that allow quantification and propagation of the uncertainties associated with expression measurements (section 2.2), may help to avoid problems associated with ‘overimputation’ and the introduction of spurious signals noted by Andrews and Hemberg [2019].

To avoid this circularity, it is important to identify reliable external sources of information that can inform the imputation process. One possibility is to exploit external reference panels (like in the context of genetic association studies). Such panels are not generally available for scRNA-seq data, but ongoing efforts to develop large scale cell atlases [e.g. Regev et al., 2017, see also section 3.3] could provide a valuable resource for this purpose. Systematic integration of known biological network structures is desirable and may also help to avoid circularity. A possible approach is to encode network structure knowledge as prior information, as attempted in netSmooth and ADImpute. Another alternative solution is to explore complementary types of data that can inform scRNA-seq imputation. This idea was

adopted in SCRABBLE and URSM, where an external reference is defined by bulk expression measurements from the same population of cells for which imputation is performed. Yet another possibility could be to incorporate orthogonal information provided by different types of molecular measurements (see section 6.1). Methods designed to integrate multi-omics data could then be extended to enable scRNA-seq imputation, e.g. through generative models that explicitly link scRNA-seq with other data types [e.g. clonealign, Campbell et al., 2019] or by inferring a shared low-dimensional latent structure [e.g. MOFA, Argelaguet et al., 2018] that could be used within a data-reconstruction framework.

With the proliferation of alternative methods, comprehensive benchmarking is urgently required as for all areas of single-cell data analysis section 6.2. Early attempts by Zhang and Zhang [2018] and Andrews and Hemberg [2019] provide valuable insights into the performance of methods available at the time. But many more methods have since been proposed and even more comprehensive benchmarking platforms are needed. Many methods, especially those using deep learning, depend strongly on choice of hyperparameters [Hu and Greene, 2019]. There, more detailed comparisons that explore parameter spaces would be helpful, extending work like that from Sun et al. [2019] comparing dimensionality reduction methods. Learning from exemplary benchmarking studies [Soneson and Robinson, 2018, Saelens et al., 2019], it would be immensely beneficial to develop a community-supported benchmarking platform with a wide-range of synthetic and experiment ground-truth datasets (or as close as possible, in the case of experimental data) and a variety of thoughtful metrics for evaluating performance. Ideally, such a benchmarking platform would remain dynamic beyond an initial publication to allow ongoing

comparison of methods as new approaches are proposed. Detailed benchmarking would also help to establish when normalisation methods derived from explicit count models [e.g. Hafemeister and Satija, 2019, Townes et al., 2019] may be preferable to imputation.

Finally, scalability for large numbers of cells remains an ongoing concern for imputation, data smoothing and data reconstruction methods, as for all high-throughput single-cell methods and software (see section 2.3).

3.2 Challenge II: Defining flexible statistical frameworks for discovering complex differential patterns in gene expression

Beyond simple changes in average gene expression between cell types (or across bulk-collected libraries), scRNA-seq enables a high granularity of changes in expression to be unraveled. Interesting and informative changes in expression patterns can be revealed, as well as cell-type-specific changes in cell state across samples (Figure 6, Approach 1). Further understanding of gene expression changes will enable deeper knowledge across a myriad of applications, such as immune responses [Kang et al., 2018b, Stubbington et al., 2017], development [Karaiskos et al., 2017a] and drug response [Kim et al., 2015].

3.2.1 Status

Currently, the vast majority of differential expression detection methods assume that the groups of cells to be compared are known in advance (e.g., experimental conditions or cell types). However, most current analysis pipelines rely on clustering or cell type assignment to identify such groups, before

downstream differential analysis is performed, without propagating the uncertainty in these assignments or accounting for the double use of data (clustering, differential testing between clusters).

In this context, most methods have focused on comparing average expression between groups [Kharchenko et al., 2014, Finak et al., 2015], but it appears that single-cell-specific methods do not uniformly outperform the state-of-the-art bulk methods [Soneson and Robinson, 2018]. Instead, little attention has been given to more general patterns of differential expression (Figure 3), such as changes in variability that account for mean expression confounding [Eling et al., 2018], changes in trajectory along pseudotime [Campbell and Yau, 2018, van den Berge et al., 2019], or more generally, changes in distributions [Korthauer et al., 2016b]. Furthermore, methods for cross-sample comparisons of gene expression (e.g., cell-type-specific changes in cell state across samples, compare section 6.1, Figure 6 and Table 2) are now emerging, such as pseudo-bulk comparisons [Kang et al., 2018a], where expression is aggregated over multiple cells within each sample. With the expanding capacity of experimental techniques to generate multi-sample scRNA-seq datasets, further general and flexible statistical frameworks will be required to identify complex differential patterns across samples. This will be particularly critical in clinical applications, where cells are collected from multiple patients.

3.2.2 Open problems

Accounting for uncertainty in cell type assignment and for double use of data will require, first of all, a systematic study of their impact. Integrative approaches in which clustering and differential testing are simultaneously performed [Vavoulis et al., 2015]

population differences in

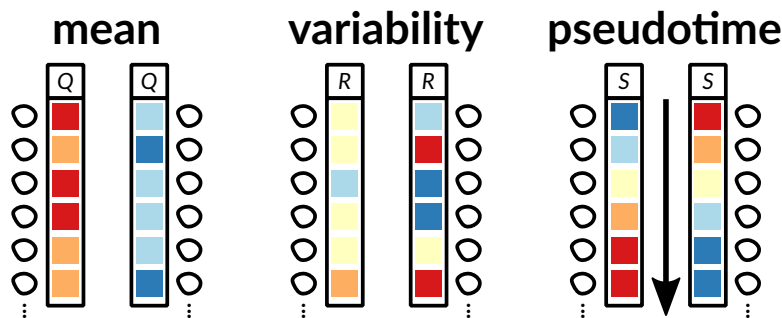


Figure 3: Differential expression of a gene or transcript between cell populations. The top row labels the specific gene or transcript, as is also done in Figure 6. A difference in **mean** gene expression manifests in a consistent difference of gene expression across all cells of a population (e.g. high vs. low). A difference in **variability** of gene expression means that in one population, all cells have a very similar expression level, whereas in another population some cells have a much higher expression and some a much lower expression. The resulting average expression level may be the same and in such cases, only single-cell measurements can find the difference between populations. A difference **across pseudotime** is a change of expression within a population, e.g. along a developmental trajectory (compare Figure 1). This also constitutes a difference between cell populations that is not apparent from population averages, but requires a pseudo-temporal ordering of measurements on single cells.

can address both issues. However, integrative methods typically require bespoke implementations, precluding a direct combination between arbitrary clustering and differential testing tools. In such cases, the adaptation of selective inference methods [Reid et al., 2018, Zhang et al., 2019b] could provide an alternative solution.

While some methods exist to identify more general patterns of gene expression changes (e.g. variability, distributions), these methods could be further improved by integrating with existing approaches that account for confounding effects such as cell cycle [Stegle et al., 2015] and complex batch effects [Butler et al., 2018a, Haghverdi et al., 2018]. Moreover, our capability to discover interesting gene expression patterns will be vastly expanded by connecting with other aspects of single-cell expression dynamics, such as

cell type composition, RNA velocity [Manno et al., 2018], splicing and allele-specificity. This will allow us to fully exploit the granularity contained in single-cell level expression measurements.

In the multi-donor setting, several promising methods have been applied to discover state transitions in high-dimensional cytometry datasets [Lun et al., 2017, Bruggner et al., 2014, Weber et al., 2018, Nowicka et al., 2017]. These approaches could be expanded to the higher dimensions and characteristic aspects of scRNA-seq data. Alternatively, there is a large space to explore other general and flexible approaches, such as hierarchical models where information is borrowed across samples, while allowing for sample-specific patterns.

3.3 Challenge III: Mapping single cells to a reference atlas

Classifying cells into cell types or states is essential for many secondary analyses. As an example, consider studying and classifying how expression varies across different cells and different biological conditions (for differential expression analyses, see section 3.2 and data integration Approach 1 in section 6.1, Figure 6 and Table 2). To put the results of such studies on a map, reliable reference systems are required.

The lack of appropriate, available references has so far implied that only reference-free approaches were conceivable, where unsupervised clustering approaches were the predominant option (see data integration Approach 0 in section 6.1, Figure 6 and Table 2). Method development for such unsupervised clustering of cells has already reached a certain level of maturity; see Duò et al. [2018], Freytag et al. [2018], Kiselev et al. [2019] for a systematic identification of available techniques.

However, unsupervised approaches involve manual cluster annotation. There are two major caveats: (i) manual annotation is a time-consuming process, which also (ii) puts certain limits to the reproducibility of the results. Cell atlases, as reference systems that systematically capture cell types and states, either tissue-specific or across different tissues, remedy this issue (see data integration Approach 2 in section 6.1, Figure 6 and Table 2; see also Figure 1 for an idea of what cell atlas type reference systems preferably could look like).

3.3.1 Status

See Table 1 for a list of cell atlas type references that have recently been published. For

human, similar endeavours as for the mouse are under way, with the intention to raise a Human Cell Atlas [Regev et al., 2017]. Towards this end, initial consortia focus on specific organs, for example the lung [Schiller et al., 2019].

The availability of these reference atlases has led to the active development of methods that make use of them in the context of supervised classification of cell types and states [Lieberman et al., 2018, Srivastava et al., 2018, Cao et al., 2019b, DePasquale et al., 2019, Kanter et al., 2019, Sato et al., 2019, Zhang et al., 2019a]. A field that serves as a source of inspiration is flow/mass cytometry, where several methods have addressed the classification of high-dimensional cell type data [Chester and Maecker, 2015, Weber and Robinson, 2016, Saeys et al., 2016, Williams et al., 2016]. Finally, as for benchmarking methods that map cells of unknown type or state onto reference atlases (see Section section 6.2 for benchmarking in general), atlases of model organisms where full lineages of cells have been integrated can form the basis for further developments [Spanjaard et al., 2018, Plass et al., 2018, Fincher et al., 2018, Farrell et al., 2018, Briggs et al., 2018]. Importantly, additional information available from lineage tracing can provide a cross-check with respect to the transcriptome-profile-based classification [Briggs et al., 2018, Kester and van Oudenarden, 2018].

3.3.2 Open problems

Cell atlases can still be considered under active development, with several computational challenges still open, in particular referring to the fundamental themes from above [Regev et al., 2017, Schiller et al., 2019, Hon et al., 2018]. Here, we focus on the mapping of cells or rather their molecular profiles onto stable existing reference atlases to fur-

organism	scale of cell atlas	citation
nematode <i>Caenorhabditis elegans</i>	whole organism at larval stage L2	[Cao et al., 2017]
planaria <i>Schmidtea mediterranea</i>	whole organism of the adult animal	[Fincher et al., 2018, Plass et al., 2018]
fruit fly <i>Drosophila melanogaster</i>	whole organism at embryonic stage	[Karaikos et al., 2017b]
Zebrafish	whole organism at embryonic stage	[Farrell et al., 2018, Wagner et al., 2018a]
frog <i>Xenopus tropicalis</i>	whole organism at embryonic stage	[Briggs et al., 2018]
Mouse	whole adult brain	[Rosenberg et al., 2018, Saunders et al., 2018, Zeisel et al., 2018]
Mouse	whole adult organism	[Tabula Muris Consortium, 2018, Han et al., 2018]

Table 1: Published cell atlases of whole tissues or whole organisms.

ther highlight the importance of these fundamental themes. A computationally and statistically sound method for mapping cells onto atlases for a range of conceivable research questions will need to: (i) enable operation at various levels of resolution of interest, and also cover continuous, transient cell states (see section 2.1); (ii) quantify the uncertainty of a particular mapping of cells of unknown type/state (see section 2.2); (iii) to scale to ever more cells and broader coverage of types and states (see section 2.3), and (iv) to eventually integrate information generated not only through scRNA-seq experiments, but also through other types of measurements, for example scDNA-seq or protein expression data (see below in section 6.1 for a discussion of data integration, especially data integration Approaches 4 and 5 in section 6.1, Figure 6 and Table 2).

3.4 Challenge IV: Generalizing trajectory inference

Several biological processes, such as differentiation, immune response or cancer expansion can be described and represented as continuous dynamic changes in cell type/state space using tree, graphical or probabilistic models. A potential path that a cell can undergo in this continuous space is often referred to as a trajectory (Trapnell et al. [2014] and Figure 1), and the ordering induced by this path is referred to as pseudotime. Several models have been proposed to describe cell state dynamics, starting from transcriptomic data [Saelens et al., 2019]. Trajectory inference is in principle not limited to transcriptomics. Nevertheless, modeling of other measurements, such as proteomic, metabolomic, and epigenomic, or even integrating multiple types of data (see section 6.1), is still at its infancy. We believe the study of complex tra-

jectories integrating different data-types especially epigenetics and proteomics information in addition to transcriptomics data will lead to a more systematic understanding of the processes determining cell fate.

3.4.1 Status

More than sixty trajectory methods have been proposed for trajectory inference from transcriptomic data using snapshot data at single or multiple time points [Saelens et al., 2019]. Briefly, those methods start from a count matrix where genes are rows and cells are columns. First, a feature selection or dimensionality reduction step is used to explore a subspace where distances between cells are more reliable. Next, clustering and minimum spanning trees [Trapnell et al., 2014, Ji and Ji, 2016], principal curve or graph fitting [Qiu et al., 2017, Chen et al., 2019, Rizvi et al., 2017], or random walks and diffusion operations on graphs (Haghverdi et al. [2016], Setty et al. [2016] among others) are used to infer pseudotime and/or branching trajectories. Alternative probabilistic descriptions can be obtained using optimal transport analysis [Schiebinger et al., 2017] or approximation of the Fokker-Planck equations [Weinreb et al., 2018] or by estimating pseudotime through dimensionality reduction with a Gaussian process latent variable model [Campbell and Yau, 2016, Reid and Wernisch, 2016, Ahmed et al., 2019].

3.4.2 Open problems

Potentially, many of the above-mentioned methods for trajectory inference can be extended to data obtained with non-transcriptomic assays. Thereby, the following aspects are crucial. First, it is necessary to define the features to use; while for transcriptomic data the features are well anno-

tated and correspond to expression levels of genes, clear-cut features are harder to determine for data such as methylation profiles and chromatin accessibility where signals can refer to individual genomic sites, but also be pooled over sequence features or sequence regions. Second, many of those recent technologies only allow measurement of a quite limited number of cells compared to transcriptomic assays where millions of cells can be profiled using droplet-based platforms [Macosko et al., 2015, Klein et al., 2015, Zheng et al., 2017]. Third, some of those measurements are technically challenging since the input material for each cell is limited (for example two copies of each chromosome for methylation or chromatin accessibility), giving rise to more sparsity than scRNA-seq. In the latter case it is necessary to define distance or similarity metrics that take this problem into account. An alternative approach consists of pooling/combining information from several cells or data imputation. For example, imputation has been used for single-cell DNA methylation [Angermueller et al., 2017], aggregation over chromatin accessibility peaks from bulk or pseudo-bulk sample [Cusanovich et al., 2018], and k-mer-based approaches have been proposed [Buenrostro et al., 2018, de Boer and Regev, 2018, Chen et al., 2019]. However, so far, no systematic evaluation (see section 6.2) of those choices has been performed and it is not clear how many cells are necessary to reliably define those features.

A pressing challenge is to assess how the different trajectory inference methods perform on different data types and importantly to define metrics that are suitable. Also, it is necessary to reason on the ground truth or propose reasonable surrogates (e.g. previous knowledge about developmental processes). Some recent papers explore this idea using scATAC-seq data, an assay to measure chro-

1 matin accessibility [Buenrostro et al., 2018,
2 Chen et al., 2019, Pliner et al., 2018].

3 Having defined robust methods to recon-
4 struct trajectories from each data type, an-
5 other future challenge is related to their com-
6 parison or alignment. Here, some ideas from
7 recent methods used to align transcriptomic
8 datasets may be extended [Butler et al.,
9 2018b, Haghverdi et al., 2018, Welch et al.,
10 2018]. A related unsolved problem is that
11 of comparing different trajectories obtained
12 from the same data type but across individu-
13 als or conditions to highlight unique and com-
14 mon aspects.

15 3.5 Challenge V: Finding 16 patterns in spatially resolved 17 measurements

18 Single-cell spatial transcriptomics or pro-
19 teomics [Crosetto et al., 2015, Strell et al.,
20 2018, Moffitt et al., 2018] technologies can
21 obtain transcript abundance measurements
22 while retaining spatial coordinates of cells or
23 even transcripts within a tissue (this can be
24 seen as an additional feature space to inte-
25 grate, see Approach 3 in section 6.1, Figure 6
26 and Table 2). With such data, the question
27 arises of how spatial information can best be
28 leveraged to find patterns, infer cell types or
29 functions and classify cells in a given tissue
30 [Tanay and Regev, 2017].

31 3.5.1 Status

32 Experimental approaches have been tailored
33 to either systematically extract foci of cells
34 and analyze them with scRNA-seq, or to mea-
35 sure RNA and proteins *in-situ*. Histological
36 sections can be projected in two dimensions
37 while preserving spatial information using se-
38 quencing arrays [Stahl et al., 2016]. Whole
39 tissues can be decomposed using the Niche-
40 seq approach [Medaglia et al., 2017]: here

a group of cells are specifically labeled with
a fluorescent signal, sorted and subjected to
scRNA-seq. The Slide-seq approach uses an
array of Drop-seq drops with known barcodes
to dissolve corresponding slide sites and se-
quence them with the respective barcodes
[Rodriques et al., 2019]. Ultimately, one
would like to sequence inside a tissue with-
out dissociating the cells and without compro-
miting on the unbiased nature of scRNA-seq.
A preliminary approach has been proposed
by Lee et al. [2015] coined FISSEQ (Fluo-
rescent *in-situ* sequencing). Lubeck et al.
[2014] have shown a first approach to itera-
tively apply fluorescence *in-situ* hybridization
to measure hundreds of RNA species simulta-
neously, called seqFISH. SeqFISH+ scales the
FISH barcoding strategy to 10,000 genes by
splitting each of four barcode locations to be
scanned into 20 separate readings to avoid sig-
nal crowding [Eng et al., 2019]. Based on a
related principle, MERFISH was proposed by
Chen et al. [2015], which enables to measure
hundreds to thousands of transcripts in sin-
gle cells simultaneously while retaining spa-
tial coordinates [Moffitt et al., 2016]. Here,
even the subcellular coordinates of each in-
dividual transcript are retained. In addition
to the methods that provide *in-situ* measure-
ments of RNA, Giesen et al. [2014] and An-
gelo et al. [2014] use mass cytometry tech-
nology to quantify the abundance of proteins
while preserving subcellular resolution. Fi-
nally, the recently described Digital Spatial
Profiling [DSP, Merritt et al., 2019, Van and
Blank, 2019] promises to provide both RNA
and protein measurements with spatial reso-
lution.

For determining cell types, or clustering
cells into groups that conduct a common func-
tion, several methods are available [Zhang
et al., 2019a, Kiselev et al., 2018, Butler et al.,
2018b]. None of these currently directly use
spatial information. In contrast, spatial cor-

relation methods have been used to detect aggregation of proteins [Shivanandan et al., 2016]. Shah et al. [2016] use seqFISH to measure transcript abundance of a set of marker genes while retaining the spatial coordinates of the cells. Cells are clustered by gene expression profiles and then assigned to regions in the brain based on their coordinates in the sample. Recently, Edsgård et al. [2018] presented a method to detect spatial differential expression patterns per gene based on marked point processes [Jacobsen, 2005]. Svensson et al. [2018a] provided a method to perform a spatially resolved differential expression analysis. Here, spatial dependence for each gene is learned by nonparametric regression, enabling the testing of the statistical significance for a gene to be differentially expressed in space.

3.5.2 Open problems

The central problem is to consider gene or transcript expression and spatial coordinates of cells, and derive an assignment of cells to classes, functional groups or cell types. While methods for both assigning cell types or functional groups and spatially resolved gene expression analysis are present, there is currently no method available that combines the two by leveraging information from spatial localization to determine the cell type or find groups of cells that conduct a common function. Depending on the studied biological question, it can be useful to constrain assignments with expectations on the homogeneity of the tissue. For example, a set of cells grouped together might be required to appear in one or multiple clusters where little to no other cells are present. Such constraints might depend on the investigated cell types or tissues. For example, in cancer, spatial patterns can occur on multiple scales, ranging from single infiltrating immune cells

[Fridman et al., 2011] and minor subclones [Swanton, 2012] to larger subclonal structures or the embedding in surrounding normal tissue and the tumor microenvironment [Cretu and Brooks, 2007]. Currently, to the best of our knowledge, there is no method available that would allow the encoding of such prior knowledge while inferring cell types by integrating spatial information with transcript or gene expression. Another important aspect when modeling the relation between space and expression is whether uncertainty in the measurements can be propagated to downstream analyses. For example, it is desirable to rely on transcript quantification methods that provide the posterior distribution of transcript expression [Kharchenko et al., 2014, Köster et al., 2017] and propagate this information to the spatial analysis. Finally, in light of issues with sparsity in single-cell measurements (section 3.1), it appears desirable to integrate spatial information into the quantification itself, and e.g. use neighboring cells within the same tissue for imputation or the inference of a posterior distribution of transcript expression.

4 Challenges in single-cell genomics

With every cell division in an organism, the genome can be altered through mutational events ranging from point mutations, over short insertions and deletions, to large scale copy number variation and complex structural variants. In cancer, the entire repertoire of these genetic events can occur during disease progression (Figure 4). The resulting tumor cell populations are highly heterogeneous. As tumor heterogeneity can predict patient survival and response to therapy, including immunotherapy, quantifying this het-

erogeneity and understanding its dynamics are crucial for improving diagnosis and therapeutic choices (Figure 4).

Classic bulk sequencing data of tumor samples taken during surgery are always a mixture of tumor and normal cells (including e.g. invading immune cells). This means that disentangling mutational profiles of tumour subclones will always be challenging, which especially holds for rare subclones that could nevertheless be the ones e.g. bearing resistance mutation combinations prior to a treatment (Figure 4). Here, the sequencing of (sufficient) single cells holds the exciting promise of directly identifying and characterizing those subclone profiles (Figure 4).

4.1 Challenge VI: Improving single-cell DNA sequencing data quality and scaling to more cells

Despite accumulating technological advances in the field, the task of characterizing tumor heterogeneity and inferring the evolutionary mechanisms that give rise to this heterogeneity is still hampered by multiple types of errors that occur during the process of scDNA-seq [Wang and Song, 2017, Hou et al., 2015, Gawad et al., 2016, Estévez-Gómez et al., 2018]. DNA sequencing technologies differ in their protocols of single-cell isolation and lysis, whole genome amplification (WGA), and library preparation [Zhang et al., 2016]. Failure of cell isolation leads to the presence—albeit usually in a small proportion—of doublets instead of single cells and the cell lysis step can introduce artificial sequence modification. The main source of error, however, is the WGA step. Single cells only carry two (in case of normal cells) up to tens (in amplified regions of disease cells) of copies of DNA molecules, which need to be substan-

tially amplified from pico to nanogram scale to read their sequence. Amplification-related artifacts include i) amplification errors, i.e. sequence alterations such as single nucleotide or indel errors introduced by the polymerase in the copy process, ii) allelic bias, i.e. the differential amplification of the alleles at a genomic locus (if one allele fails to amplify at all, this is an allele dropout, if both fail, a locus dropout), iii) chimeric sequences. The majority of WGA approaches can be broadly classified into PCR-based and multiple displacement amplification (MDA)-based methods. The PCR-based technologies include degenerate oligonucleotide-primed PCR (DOP-PCR) [Telenius et al., 1992], linker-adaptor PCR [Klein et al., 1999], primer extension pre-amplification PCR (PEP-PCR-/I-PEP-PCR) [Zhang et al., 1992, Arneson et al., 2008] and others. They require thermostable polymerases that withstand all temperatures during the cycling. More recent MDA-based technologies use the strand-displacing, high-fidelity Φ 29 DNA polymerase [Blanco et al., 1989, Dean et al., 2002, Spits et al., 2006b, Picher et al., 2016, Paez et al., 2004, Spits et al., 2006a] for an isothermal reaction, as it is not stable at common PCR temperature maxima. Another approach, called multiple annealing and looping-based amplification cycles (MALBAC) combines MDA and PCR, and relies on the *Bacillus stearothermophilus* polymerase for the MDA process [Zong et al., 2012].

4.1.1 Status

Ideally, scDNA-seq should provide information about the entire repertoire of distinct events that occurred in the genome of a single cell, such as copy number alterations, genomic rearrangements, together with SNVs and smaller insertion and deletion variants. However, amplification biases and errors present a

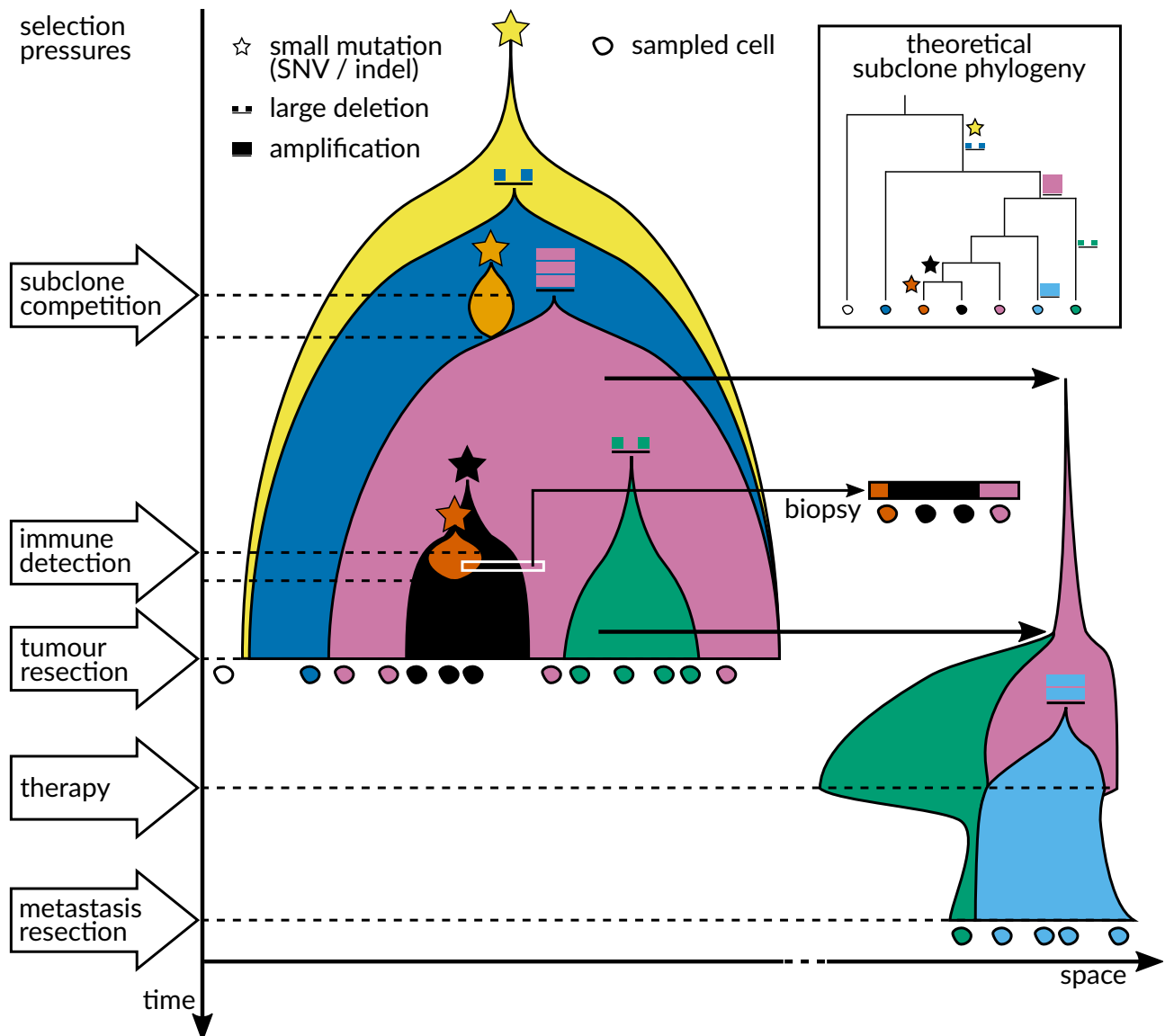


Figure 4: From initiation of a tumour to its detection, resection and possible metastasis, it will evolve somatically. New genomic mutations can confer a selective advantage to the resulting new subclone, that can allow it to outcompete other tumour subclones (subclone competition). At the same time, the acting selection pressures can change over time, e.g. due to new subclones arising, the immune system detecting certain subclones, or as a result of therapy. Understanding such selective regimes—and how specific mutations alter a subclone’s susceptibility to changes in selection pressures—will help construct an evolutionary model of tumorigenesis. And it is only within this evolutionary model, that more efficient and more patient-specific treatments can be developed. For such a model, unambiguously identifying mutation profiles of subclones via scDNA-seq of resected or biopsied single cells is crucial.

serious challenge to variant calling [de Bourcy et al., 2014, Hou et al., 2015, Huang et al., 2015, Estévez-Gómez et al., 2018]: It is broadly accepted that different WGA technologies should be used depending on whether SNVs or whether copy number variation (CNV)s are to be detected, as the distinct technologies differ in the magnitude of amplification bias, and the rates of amplification errors and chimera formation. Generally, PCR-based approaches with more uniform coverage should be used for CNV calling, while MDA-based methods that result in less single nucleotide errors should be applied for SNV calling. The goal must thus be to (i) improve the coverage uniformity of MDA-based methods, (ii) reduce the error rate of the PCR-based methods, or (iii) create new methods that exhibit both a low error rate and a more uniform amplification of alleles. Recent years witnessed intensive research in these directions, e.g.: (i) Improved coverage uniformity for MDA has been achieved using droplet microfluidics-based methods, resulting in emulsion WGA (eWGA, [Fu et al., 2015]), single droplet MDA (sd-MDA, [Hosokawa et al., 2017]) and digital droplet multiple displacement amplification (ddMDA, [Sidore et al., 2016]). A second approach has been to couple the Φ 29 DNA polymerase to a primase to reduce priming bias [Picher et al., 2016]. Both these approaches improve the calling of CNVs from the resulting data. (ii) One way to reduce the amplification error rate of the PCR-based methods (including MALBAC) would be to employ a thermostable polymerase (necessary for use in PCR) with proof-reading activity similar to Φ 29 DNA polymerase. While SD polymerase combines thermostability with strand displacement and has been tested for WGA [Blagodatskikh et al., 2017], we are not aware of any PCR DNA polymerases with a fidelity in the range of Φ 29 DNA polymerase [Potapov and Ong,

2017] having been used in PCR-based WGA. (iii) Three newer methods use an entirely different approach: They randomly insert transposons into the whole genome and then leverage these as priming sites for library preparation and amplification. Direct library preparation (DLP, [Zahn et al., 2017a]), as the name suggests, directly sequences the resulting shallow library without any amplification, allowing only for CNV calling. It has recently been further improved to account for doublets and dead cells and scaled to 80,000 single cells [Laks et al., 2018]. Transposon Barcoded (TnBC) follows the transposon integration with PCR amplification, making it useful for CNV calling, but suffering from amplification errors [Xi et al., 2017]. Finally, Linear Amplification via Transposon Insertion (LIANTI, [Chen et al., 2017]) introduces a new approach to dealing with amplification errors. Instead of exponential amplification, their amplification process is linear: From promoters included in the transposon insertion, they transcribe the original tagged sequence multiple times and then use reverse transcription and second-strand synthesis to obtain double-stranded DNA for sequencing. As errors introduced by the individual processes are not propagated, they should be unique to individual copies and accordingly the authors report a false positive rate that is even lower than for MDA [Chen et al., 2017].

4.1.2 Open problems

These recent developments promise scalable methodology for scDNA-seq comparable to that already available for scRNA-seq, while at the same time reducing previously limiting errors and biases. In addition to further improvements over the described existing methods, the major challenge will be to continuously and systematically evaluate the whole range of promising WGA methods for

1 the identification of all types of genetic varia-
2 tion from SNVs over smaller insertions and
3 deletions up to copy number variation and
4 structural variants.

5 4.2 Challenge VII: Errors and 6 missing data in the 7 identification of features / 8 variation from single-cell 9 DNA sequencing data.

10 The aim of scDNA sequencing usually is to
11 track somatic evolution at the cellular level,
12 that is, at the finest resolution possible rela-
13 tive to the laws of reproduction (cell division,
14 Figure 5). Examples refer to identifying het-
15 erogeneity and tracking evolution in cancer,
16 as the likely most predominant use case (also
17 see below in section 5), but also to monitor-
18 ing the interaction of somatic mutation with
19 developmental and differentiation processes.
20 To track genetic drifts, selective pressures, or
21 other phenomena inherent to the development
22 of cell clones or types (Figure 4)—but also to
23 stratify cancer patients for the presence of re-
24 sistant subclones—it is instrumental to geno-
25 type and also phase genetic variants in single
26 cells with sufficiently high confidence.

27 The major disturbing factor in scDNA-seq
28 data is the WGA process (see section 4.1).
29 All methodologies introduce amplification er-
30 rors (false positive alternative alleles), but
31 more drastic is the effect of amplification bias:
32 the insufficient or complete failure of am-
33 plification, which leads to imbalanced pro-
34 portions or complete lack of variant alleles.
35 Overall, one can distinguish between three
36 cases: (i) an imbalanced proportion of al-
37 leles, i.e. loci harboring heterozygous muta-
38 tions where preferential amplification of one
39 of the two alleles leads to read counts that
40 are distorted, sometimes heavily; (ii) allele
41 drop-out, i.e. loci harboring heterozygous mu-

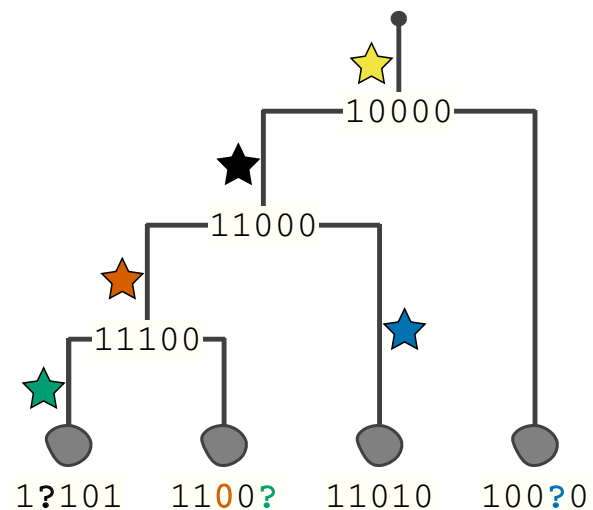


Figure 5: Mutations (colored stars) accumu-
late in cells during somatic cell divisions
and can be used to reconstruct the develop-
mental lineages of individual cells within an
organism (leaf nodes of the tree with muta-
tional presence / absence profiles attached).
However, insufficient or unbalanced WGA
can lead to the dropout of one or both al-
leles at a genomic site. This can be mitigated
by better amplification methods, but also
by computational and statistical methods
that can account for or impute the missing
values.

tations where only one of the alleles was am-
plified and sequenced, and (iii) site drop-out,
which is the complete failure of amplification
of both alleles at a site and the resulting lack
of any observation of a certain position of the
genome. Note that (ii) can be considered an
extreme case of (i).

A sound imputation of missing alleles and
a sufficiently accurate quantification of un-
certainties will yield massive improvements in
geno- and haplotyping (phasing) somatic vari-
ants. This, in turn, is necessary to substan-
tially improve the identification of subclonal
genotypes and the tracking of evolutionary
developments. Potential improvements in this

area include (i) more explicit accounting for possible scDNA-seq error types, (ii) integrating with different data types with error profiles different from scDNA-seq (e.g. bulk sequencing or RNA sequencing), or (iii) integrating further knowledge of the process of somatic evolution, such as the constraints of phylogenetic relationships among cells, into variant calling models. In this latter context, it is important to realize that somatic evolution is asexual. Thus, no recombination occurs during mitosis, eliminating a major disturbing factor usually encountered when aiming to reconstruct species or population trees from germline mutation profiles.

4.2.1 Status

Current single-cell specific SNV callers include Monovar [Zafar et al., 2016] and SCcaller [Dong et al., 2017]. SCcaller detects somatic variants independently for each cell, but accounts for local allelic amplification biases by integrating across neighbouring germline single-nucleotide polymorphisms. It exploits the fact that allele dropout affects contiguous regions of the genome large enough to harbor several, and not only one, heterozygous mutation loci. Monovar uses an orthogonal approach to variant calling. It does not assume any dependency across sites, but instead handles low and uneven coverage and false positive alternative alleles by integrating the sequencing information across multiple cells. While Monovar merely creates a consensus across cells, integrating across cells is particularly powerful if further knowledge about the dependency structure among cells is incorporated. As pointed out above, due to the lack of recombination, any sample of cells derived from an organism shares an evolutionary history that can be described by a cell lineage tree (see section 5). This tree, however, is in general un-

known and can in turn only be reconstructed from single-cell mutation profiles. A possible solution is to infer both mutation calls and a cell lineage tree at the same time, an approach taken by a number of existing tools: single-cell Genotyper [Roth et al., 2016], SciCloneFit [Zafar et al., 2018] and SciPhi [Singer et al., 2018].

Finally, SSrGE, identifies SNVs correlated with gene expression from scRNA-seq data [Poirion et al., 2018].

Some basic approaches to CNV calling from scDNA-seq data are available. These are usually based on hidden markov models (HMMs) where the hidden variables correspond to copy number states, as e.g. in Aneupfinder [Bakker et al., 2016]. Another tool, Ginkgo, provides interactive CNV detection using circular binary segmentation, but is only available as a web-based tool [Garvin et al., 2015]. ScRNA-seq data, which does not suffer from the errors and biases of WGA, can also be used to call CNVs or loss of heterozygosity events: an approach called HoneyBADGER [Fan et al., 2018] utilizes a probabilistic hidden Markov model, whereas the R package inferCNV simply averages the expression over adjacent genes [Patel et al., 2014].

4.2.2 Open problems

SNV callers for scDNA-seq data have already incorporated amplification error rates and allele dropout in their models. But beyond these rates, the challenge remains to further extend this into a full statistical modelling of the amplification process, that would inherently account for both errors and biases, and more accurately quantify the resulting uncertainties (see section 2.2). This could be achieved by expanding models that accurately quantify uncertainties in related

settings² and would ultimately even allow reliable control of false discovery rates in the variant discovery and genotyping process. Such expanded models can build on a number of recent studies in this context, e.g. on a formalisation in a recent preprint [Koptagel et al., 2018]. Furthermore, such models could integrate the structure of cell lineage trees with the structure implicit in haplotypes that link alleles. For haplotype phasing, Satas and Raphael [2018] recently proposed an approach based on contiguous stretches of amplification bias (similar to SCcaller, see above), whereas others propose read-backed phasing in two recent studies [Bohrson et al., 2019, Hård et al., 2019]. In addition, the integration with deep bulk sequencing data, as well as with (sc)RNA-seq data remains unexplored, although it promises to improve the precision of callers without compromising sensitivity.

Identification of short insertions and deletions (indels) is another major challenge to be addressed: we are not aware of any scDNA-seq variant callers with those respective capabilities.

For copy number variation calling, software has previously been published mostly in conjunction with data-driven studies. Here, a systematic analysis of biases in the most common WGA methods for copy number variation calling (including newer methods to come) could further inform method development. The already mentioned approach of leveraging amplification bias for phasing could also be informative [Satas and Raphael, 2018].

The final challenge is a systematic comparison of tools beyond the respective software

publications, which is still lacking for both SNV and CNV callers. This requires systematic benchmarks, which in turn require simulation tools to generate synthetic datasets, as well as sample-based benchmarking datasets with a reasonably reliable ground truth (see section 6.2).

5 Challenges in single-cell phylogenomics

Single-cell variant profiles from scDNA-seq, as described above (section 4.2), can be used in computational models of somatic evolution, including cancer evolution as an important special case (Figure 4). For cancer, there is an on-going, lively discussion about the very nature of evolutionary processes at play, with competing theories such as linear, branching, neutral, and punctuated evolution [Davis et al., 2017].

Models of cancer evolution may range from a simple binary representation of the presence versus the absence of a particular mutational event (Figure 5), to elaborate models of the mechanisms and rates of distinct mutational events. There are two main modeling approaches that lend themselves to the analysis of tumour evolution [Altrock et al., 2015]: phylogenetics and population genetics.

Phylogenetics comes with a rich repertoire of computational methods for likelihood-based inference of phylogenetic trees [Felsenstein, 1981]. Traditionally, these methods are used to reconstruct the evolutionary history of a set of distinct species. However, they can also be applied to cancer cells or subclones (Figure 4). In this setting, tips of the phylogeny (also called leaves or taxa) represent sampled and sequenced cells or subclones, whereas inner nodes (also called ancestral) represent their hypothetical common ances-

²<https://varlociraptor.github.io>

tors. The input for a phylogenetic inference commonly consists of a multiple sequence alignment (MSA) of molecular sequences for the species of interest. For cancer phylogenies, one would concatenate the SNVs (and possibly other variant types) to assemble the input MSA. The key challenge for phylogenetic method development comprises designing sequence evolution models that are (i) biologically realistic and yet (ii) computationally tractable for the increasingly large number of sequenced cells per patient and study.

In population genetics, the tumor is understood as a population of evolving cells (Figure 4). To date, population genetic theory has been used to model the initiation, progression and spread of tumors from bulk sequencing data [Foo et al., 2011, Beerenwinkel et al., 2007, Haeno et al., 2012]. The general mathematical framework behind these models are branching processes [Kimmel and Axelrod, 2015], e.g. in models of the accumulation of driver and passenger mutations [Bozic et al., 2016, 2010]. Here, the driver mutations carry a fitness advantage, as might epistatic interactions among them [Bauer et al., 2014]. On the other hand, passenger mutations are assumed to be neutral regarding fitness; they merely hitchhike along the fitness advantage of driver mutations they are linked to via their haplotype. The parameters of population genetic models describe inherent features of individual cells that are relevant for the evolution of their populations, e.g. fitness and the rates of birth, death, and mutations. Such cell-specific parameters should more naturally apply to and be derived from information gathered by sequencing of individual cells, as opposed to sequencing of bulk tissue samples. Models using these parameters and the information about the evolutionary dynamics of cancer they contain, will e.g. be essential in the design of adaptive cancer treatment strategies that aim at managing subclonal tu-

mour composition [Acar et al., 2019, Zhang et al., 2017].

5.1 Challenge VIII: Scaling phylogenetic models to many cells and many sites

Even if given perfect data, phylogenetic models of tumor evolution would still face the challenge of computational tractability, which is mainly induced by: (i) the increasing numbers of cells that are sequenced in cancer studies (see section 2.3) and (ii) the increasing numbers of sites that can be queried per genome (also see section 2.3).

5.1.1 Open problems

(i) While adding data from more single cells will help improve the resolution of tumour phylogenies [Graybeal, 1998, Pollock et al., 2002], this exacerbates one of the main challenges of phylogenetic inference in general: the immense space of possible tree topologies that grows super-exponentially with the number of taxa—in our case the number of single cells. Therefore, phylogenetic inference is NP-hard [Roch, 2006] under most scoring criteria (a scoring criterion takes a given tree and MSA to calculate how well the tree explains the observed data). Calculating the given score on all possible trees to find the tree that best explains the data is computationally not feasible for MSAs containing more than approximately 20 single cells, and thus requires heuristic approaches to explore only promising parts of the tree search space.

(ii) In addition to the growing number of cells (taxa), the breadth of genomic sites and genomic alterations that can be queried per genome also increases. Classical approaches thus need not only scale with the number of

single cells queried (see above), but also with the length of the input MSA. Here, previous efforts for parallelization [Aberer et al., 2014, Ayres, 2017] and other optimisation efforts [Ogilvie et al., 2017] exist and can be built upon. The breadth of sequencing data also allows determination of large numbers of invariant sites, which further raises the question of whether including them will change results of phylogenetic inferences in the context of cancer. Excluding invariant sites from the inference has been coined ascertainment bias, and for phylogenetic analyses of closely related individuals from a few populations it has been shown that accounting for ascertainment bias alters branch lengths, but not the resulting tree topologies per se [Leaché et al., 2015].

5.2 Challenge IX: Integrating multiple types of features / variation into phylogenetic models

Naturally, downstream analyses—like characterising intratumor heterogeneity and inferring its evolutionary history—suffer from the unreliable variant detection in single cells. The better the quality of the variant calls gets, however, the more important it becomes to model all types of available signal in mathematical models of tumour evolution, with the goal of increasing the resolution and reliability of the resulting trees; from SNVs, over smaller insertions and deletions, to large structural variation and CNVs (Figure 4). Finally, to model somatic phylogenies comprehensively, all available types of variants will have to be integrated into a comprehensive model. In the context of cancer, with genomic destabilisation occurring, this will be especially challenging.

5.2.1 Status

For phylogenetic tree inference from SNVs of single cells, a considerable number of tools exist. The early tools OncoNEM [Ross and Markowitz, 2016] and SCITE [Jahn et al., 2016] use a binary representation of presence or absence of a particular SNV. They account for false negatives, false positives and missing information in SNV calls, where false negatives are orders of magnitude more likely to occur than false positives. The more recent tool SiFit [Zafar et al., 2017] also uses a binary SNV representation, but infers tumor phylogenies allowing for both noise in the calls and for violations of the infinite sites assumption. Another approach allowing for violations of the infinite sites assumption is the extension of the Dollo parsimony model to allow for k losses of a mutation (Dollo- k) [El-Kebir, 2018, Ciccolella et al., 2018]. Single cell genotyper [Roth et al., 2016], SciCloneFit [Zafar et al., 2018], or SciPhi [Singer et al., 2018] jointly call mutations in individual cells and estimate the tumor phylogeny of these cells, directly from single-cell raw sequencing data. In a recent work [Kozlov, 2018], a standard phylogenetic inference tool RAXML-NG [Kozlov et al., 2019] has been extended to handle single-cell SNV data. In particular, this implements (i) a 10-state substitution model to represent all possible unphased diploid genotypes and (ii) an explicit error model for allelic dropout and genotyping/amplification errors. Initial experiments showed that—although a 10-state model incorporates more information—it outperformed the ternary model (as used by SiFit) only slightly and only in simulations with very high error rates (10%-50%). However, further analysis suggests that benefits of the genotype model become much more pronounced with an increasing number of cells and, in particular, an increasing number of SNVs (Kozlov, personal communication).

While there are no tools yet available to identify insertions and deletions from scDNA-seq (see challenge above), it is only a matter of time until such callers will become available. As they can already be identified from bulk sequencing data, some precious efforts to incorporate indels in addition to substitutions into classical phylogenetic models exist: A decade ago, a simple probabilistic model of indel evolution was proposed [Rivas and Eddy, 2008]. But although some progress has been made since then, such models are less tractable than the respective substitution models [Holmes, 2017].

Incorporating CNVs in the reconstruction of tumor phylogeny can be helpful for understanding tumor progressions, as they represent one of the most common mutation types associated to tumor hypermutability [Kim et al., 2013]. CNVs in single cells were extensively studied in the context of tumor evolution and clonal dynamics [Navin et al., 2011, Eirew et al., 2015]. Reconstructing a phylogeny with CNVs is not straightforward. The challenges are not only related to experimental limits, such as the complexity of bulk sequencing data [Zaccaria et al., 2017] and amplification biases [Gawad et al., 2016], but also involve computational constraints. First of all, the causal mechanisms, such as breakage-fusion-bridge cycles [Bignell et al., 2007] and chromosome missegregation [Santaguida et al., 2017], can lead to overlapping copy number events [Schwarz et al., 2014]. Secondly, inferring a phylogeny with CNV data requires quantifying transition probabilities for changes in copy numbers based on the causal mechanisms. Towards that goal, approaches to calculate the distance between whole copy number profiles [Zeira and Shamir, 2018] are a first step. But for them, a number of challenges remain, with several of the underlying problems known to be NP-hard [Zeira and Shamir, 2018].

Co-occurrence of all of the above variation types further complicates mathematical modeling, as these events are not independent. For example, multiple SNVs that occurred in the process of tumor evolution may disappear at once via a deletion of a large genomic region. In addition, recent analyses revealed recurrence and loss of particular mutational hits at specific sites in the life histories of tumors [Kuipers et al., 2017], undermining the validity of the so called infinite sites assumption, commonly made by phylogenetic models: it assumes an infinite number of genomic sites, thus rendering a repeated mutational hit of the same genomic site along a phylogeny impossible.

5.2.2 Open problems

For phylogenetic reconstruction from SNVs, we anticipate a shift towards leveraging improvements in input data quality as they are achieved through better amplification methods and SNV callers (see challenges above). For indels, variant callers for scDNA-seq data remain to be developed (see challenge above), but are anticipated. Thus, indel modelling efforts for phylogenetic reconstruction from bulk sequencing data should be adapted. For phylogenetic inference from CNVs, the major challenges are (i) determining correct mutational profiles and (ii) computing realistic transition probabilities between those profiles.

The final challenge will be to incorporate all of the above phenomena into a holistic model of cancer evolution. However, this will substantially increase the computational cost of reconstructing the evolutionary history of tumor cells. Thus, one needs to carefully determine which phenomena actually do matter (e.g. which parameters even affect the final tree topology) and which features can be

measured (section 4.1) and called (section 4.2) with sufficient accuracy to actually improve modelling results. As a consequence one might be able to devise more lightweight models for answering specific questions and invest considerable effort into optimizing novel tools at the algorithmic and technical level (see challenge below).

5.3 Challenge X: Inferring population genetic parameters of tumor heterogeneity by model integration

Tumor heterogeneity is the result of an evolutionary journey of tumor cell populations through both time and space [Swanton, 2012, McGranahan and Swanton, 2017]. Microenvironmental factors like access to the vascular system and infiltration with immune cells differ greatly—for regions within the original tumor as well as between the main tumour and metastases, and across different time points [Yang and Lin, 2017]. This imposes different selective pressures on different tumour cells, driving the formation of tumour subclones and thus determining disease progression (including metastatic potential), patient outcome and susceptibility to treatment (Junttila and de Sauvage [2013], Corredor et al. [2018] and Figure 4). However, even the answers to very basic questions about the resulting dynamics remain unanswered [Tura-jlic and Swanton, 2016]: for example, whether metastatic seeding from the primary tumor occurs early and multiple times in parallel, with metastases diverging genetically from the primary tumor, or whether seeding of metastases occurs late, from a far-developed subclone in the primary tumour, with that subclone seeding multiple locations with a genotype closer to the late-stage primary tu-

mour; and whether a single cell can seed a metastasis, or whether the joint migration of a set of cells is required. Here, sc-seq can provide invaluable resolution [Navin et al., 2011].

Although many mathematical models of tumor evolution have been proposed [Bozic et al., 2010, 2016, Altrock et al., 2015, Foo et al., 2011, Michor et al., 2004], fundamental parameters characterizing the evolutionary processes remain elusive. To quantitatively describe the tumor evolution process and evaluate different possible modes against each other (e.g. modes of metastatic seeding), we would like to estimate fitness values of individual mutations and mutation combinations, as well as rates of mutation, cell birth and cell death—if possible, on the level of subclones. These parameters determine the underlying fitness landscape of individual cells within their microenvironment, which in turn determines the evolutionary dynamics of cancer progression.

5.3.1 Status

Recent technological advances already allow for measuring the arrangement and relationships of tumor cells in space, with cell location basically amounting to a second measurement type requiring data integration within a cell (Approach 3 in section 6.1, Figure 6 and Table 2). While in vivo imaging techniques might also become interesting for obtaining time series data in the future [Larue et al., 2017], the automated analysis of whole slide immunohistochemistry images [Ghaznavi et al., 2013, Saco et al., 2016] seems the most promising in the context of cancer and mutational profiles from scDNA-seq. It is already amenable to single-cell extraction of characterised cells with known spatial context and subsequent scDNA-seq. Using laser capture microdissection [Datta et al., 2015] hundreds of single cells have recently been

isolated from tissue sections and analysed for copy number variation [Casasent et al., 2018]. For cell and tissue characterisation in immunohistochemical images, machine learning models are trained to segment the images and recognize structures within tissues and cells [Gurcan et al., 2009, Irshad et al., 2014, Komura and Ishikawa, 2018]: They can e.g. determine the densities and quantities of mitotic nuclei, vascular invasion, immune cell infiltration on the tissue level, as well as stained biomarkers on the level of the individual cell. These are key parameters of the tumor microenvironment, characterising the interaction tumor cells with their environment in space [Yuan, 2016, Heindl et al., 2015].

Mathematical models of tumor population genetics have classically assumed well mixed populations, ignoring any spatial structure, let alone evolutionary microenvironments. Recently, methods have been extended to account for some spatial structure and have already led to refined predictions of the waiting time to cancer [Martens et al., 2011] and intratumor heterogeneity [Waclaw et al., 2015]. In particular, spatial statistics has been proposed for the quantitative statistical analysis of cancer digital pathology imaging [Heindl et al., 2015], but the idea is applicable to other spatially resolved readouts. A number of methods were proposed to model cell-cell interactions [Schapiro et al., 2017, Arnol et al., 2018] or to predict single-cell expression from microenvironmental features [Goltsev et al., 2018, Battich et al., 2015]. With the advent of spatially resolved DNA sequencing, models can be adapted to the new data.

Regarding temporal resolution, it is already common to sequence tumor material from different timepoints: biopsies used for diagnosis, resected tumours, lymph nodes and metastases upon surgery and tumours after relapse. These time-points already lend them-

selves to temporal analyses of clonal dynamics using bulk DNA sequencing data [Johnson et al., 2014]. But scDNA-seq will help to increase the resolution of subclonal genotypes. And integrating this clonal stratification across timepoints and with other readouts, such as cell state markers, will allow to determine central model parameters for the detection of positive and negative selection, e.g. rates of proliferation, mutation and death.

To also leverage the kinship relationships between cells, population genetic methods and models could be integrated with approaches from phylogenetics. One prominent example of this recent trend is the use of the multi-species coalescent model for analyzing MSAs that contain several individuals for several populations [Rannala and Yang, 2017, Liu et al., 2015]. This naturally translates into analyzing tumour subclones as populations of single cells, capturing some of the population structure seen in cancers. This phylogenetic context also lends itself to modelling differences in mutational rates and signatures between different cell populations, e.g. between normal somatic evolution before tumour initiation and cancer evolution after tumour initiation, or between different tumor subclones.

In this setting, we will have to account for heterotachy (see e.g. Kolaczkowski and Thornton [2008]), that is, we cannot assume a single model of substitution for the entire tree, but have to allow different models to act on distinct branches or subtrees/subclones. Here, anything from a simple model of rate heterogeneity (e.g. Yang [1994]) to an empirical mixture model as used for protein evolution [Le et al., 2012] could be considered.

A recent example integrating population genetics approaches with phylogenetics, is a computational model for inference of fitness landscapes of cancer clone populations using

sDNA-Seq data, SCIFIL [Skums et al., 2019]. It estimates the maximum likelihood fitness of clone variants by fitting a replicator equation model onto a character-based tumor phylogeny.

For the detection of positive selection, a number of phylogenetic and population genetic approaches have been proposed. Phylogenetic trees may be used for detecting branches on which positive [Zhang et al., 2005] or diversifying episodic selection [Smith et al., 2015] is acting. The tests from the area of “classic” phylogenetics might serve as a starting point for exploring and adapting appropriate methods that will allow to associate positive selection events to branches of the tumor tree or specific evolutionary events. Evolutionary pressures are often quantified by the dN/dS ratio of non-synonymous and synonymous substitutions. In application to tumor cell populations, however, this ratio may not be applicable, as it has been shown to be relatively insensitive when applied to populations within the same species [Kryazhimskiy and Plotkin, 2008]. Other measures have been proposed as better suited for detecting selection within populations based on time-series data and could potentially be transferred to tumor cell populations [Neher et al., 2014, Gray et al., 2011, Steinbrück and McHardy, 2011]. An open question is to which extent the above tests will be sensitive to errors in cancer data as they are known to produce high false positive rates in the classic phylogenetic setting if the error rate in the input data is too high [Fletcher and Yang, 2010]. Computationally intense solutions for decreasing the high false positive rate have been proposed [Redelings, 2014], but they might not scale to cancer datasets. Importantly, development of tests for positive selection could contribute to the discussion of whether the evolution of tumors is driven by selection or

neutral.

For the detection of negative selection, time resolved measurements and resulting proliferation and death rates could prove equally promising. Further, approaches were developed to discover epistatic interactions—particularly synthetic lethality—from genomic and transcriptomic data in tumor genomes and cancer cell lines [Szczurek et al., 2013, Jerby-Arnon et al., 2014], and patient survival [Matlak and Szczurek, 2017]. Some of these epistatic interactions, however, can be hard to spot in bulk sequencing data, as they may simply disappear because of a low frequency. ScDNA-seq, ideally in a time resolved fashion and across individuals, provides much more insight into epistatic interactions than bulk sequencing. The key feature is that it is possible to identify pairs of mutations that often occur simultaneously in the same genome, and pairs that rarely or never do. That is, cells affected by negatively selected or synthetic lethal mutations will go extinct in the tumor population and thus their genotype with the synthetic lethal mutations occurring together will not be observed. Cell death, however, can be the result of mere chance, so to detect significant negative pressures, large cohorts of repeated time resolved experiments would have to be performed.

5.3.2 Open problems

With an increased resolution of scDNA-seq (section 4.1) and more work on the scDNA-seq challenges described in other sections, it will be possible to determine subclone genotypes in more detail.

The first challenge will be to integrate this with the spatial location of single cells obtained from other measurements. This will enable determining whether cells from the same subclones are co-located, whether

1 metastases are founded recurrently by the
2 same subclone(s) and whether individual
3 metastases are founded by individual or mul-
4 tiple subclones. A number of studies utilizing
5 multiple region samples from the same tumor
6 and from distant metastases already paved
7 the way in investigating these questions [Tu-
8 rajlic and Swanton, 2016]. Still, only single-
9 cell spatial resolution will allow identification
10 of specific individual genotypes in specific lo-
11 cations and the drawing precise conclusions.

12 The second challenge will be to determine
13 rates of proliferation and death per subclone.
14 This could be achieved by measuring num-
15 bers of mitotic and apoptotic cells per sub-
16 clone or by integrating subclone abundance
17 profiles across time points. Good estimates
18 of these basic parameters will greatly benefit
19 models, e.g. for the detection of positive and
20 negative selection in cancer.

21 A third challenge will be to determine
22 subclone-specific rates of mutation. Here, in-
23 tegration of models from population genetics
24 and phylogenetics holds promise.

25 A fourth challenge will be to devise ways
26 to determine further relevant model param-
27 eters. For example, comparing expanded sub-
28 clones in drug screens to determine subclone
29 fitness under different treatment regimes can
30 both help to predict subclone resistance (and
31 thus expected treatment success) and further
32 inform cancer evolution models.

33 A final step will then be to put all these
34 parameters into context with further infor-
35 mation about local microenvironments (such
36 as vascular invasion and immune cell infiltra-
37 tion), to estimate the selection potential of
38 such local factors for or against different sub-
39 clones.

6 Overarching challenges 40

6.1 Challenge XI: Integration of 41 single-cell data: across 42 samples, experiments and 43 types of measurement 44

45 Biological processes are complex and dy-
46 namic, varying across cells and organisms. To
47 comprehensively analyze such processes, dif-
48 ferent types of measurements from multiple
49 experiments need to be obtained and inte-
50 grated. Depending on the actual research
51 question, such experiments will refer to dif-
52 ferent time points, tissues or organisms. For
53 different measurement types, we put particu-
54 lar emphasis on the combination of scRNA-
55 seq and scDNA-seq data, although augment-
56 ing sequencing data with records on protein
57 or metabolite levels is also possible.

58 Since the exploration of complex, dynamic
59 and variable processes requires the integration
60 of data from multiple experiments, we need
61 flexible but rigorous statistical and compu-
62 tational frameworks to support that integra-
63 tion. See Table 2 and Figure 6 for an overview
64 of how the issues in creating such frameworks
65 can vary relative to the particular problem³.

66 When aiming at the identification of pat-
67 terns of differential expression, so as to char-
68 acterize variability across organisms, individ-
69 uals, or location, data refers to the same
70 (unique) measurement type (for example,
71 only scRNA-seq), but stems from different
72 time points, different locations (such as dif-
73 ferent tissues or sites in a tumor), or different
74 organisms. See Approach 1 in Figure 6 and
75 Table 2 for methodological challenges arising

³Graph representation in Figure 6 Approaches 2 and 5 taken from Wolf et al. [2019], Fig. 3, provided under Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>)

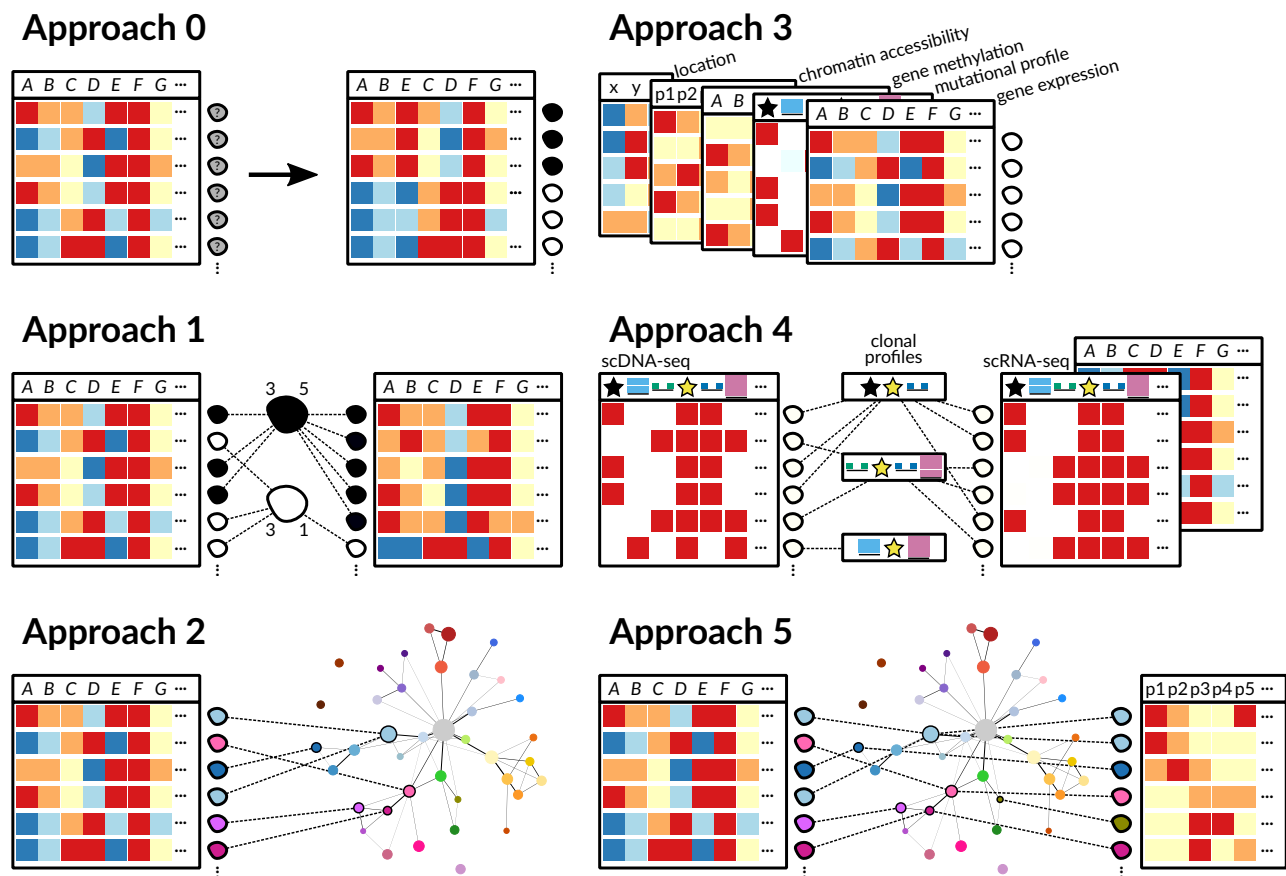


Figure 6: Approaches for integrating single-cell measurement datasets across measurement types, samples and experiments, as also described in Table 2.

Approach 0 Clustering of cells from one sample from one experiment, no data integration is needed. **Approach 1** Cell populations / clusters from multiple samples but the same measurement type need to be linked. **Approach 2** For cell populations / clusters across multiple experiments, stable reference systems like cell atlases are needed (compare Figure 1). **Approach 3** Whenever multiple measurement types can be obtained from the same cell, they are automatically linked. However, this setup highlights the problem of data sparsity of all available measurement types and the dependency of measurement types that needs to be accounted for. **Approach 4** When multiple measurement types cannot be obtained from the same cell, a solution is to obtain them from cells of the same cell population. However, this combines the problems of Approach 1 with those of Approach 3. **Approach 5** One possibility for easing data integration across measurement types from separate cells would be to have a stable reference (cell atlas) across multiple measurement types. Effectively, this combines the problems of Approaches 2, 3 and 4.

	Integration	example MTs	example AMs	Promises	Challenges
0	none	scDNA-seq, scRNA-seq, merFISH	clustering / unsupervised	identify new cell types and states	technical noise
1	within 1 MT, within 1 exp, across > 1 smps	scDNA-seq, scRNA-seq, merFISH	differential analyses, time series, spatial sampling	identify effects across sample groups, time and space	technical noise; batch effects; validate cell type assignments
2	within 1 MT, across > 1 exp, across > 1 smps,	scRNA-seq, merFISH	map cells to stable reference (cell atlas)	accelerate analyses; increase sample size & generalize obser- vations	technical noise; batch effects; validate cell type assignments; standards across experimental centres
3	across > 1 MTs, within 1 exp, within 1 cell	scG&T-seq, scM&T-seq, seqFISH	MOFA, DIABLO, MINT	holistic view of biol. processes within cell; quantification of dependency of MTs	scaling cell throughput; MT combinations limited; dependency of MTs; data sparsity
4	across > 1 MTs, within 1 exp, across > 1 cells, within 1 cell pop	scDNA-seq + scRNA-seq, DNA-seq + scRNA-seq	Cardelino, Clonealign, MATCHER	use existing datasets (faster than 3); flexible experimen- tal design	technical noise; validate cell / data grouping; test assumptions for integrating data
5	across > 1 MTs, across > 1 exps, across > 1 smps, within cells	hypothetical: any combina- tion	hypothetical: multi-omic HCA, single-cell TCGA	comprehensive char- acterizations of bio- logical systems	all from approaches 2, 3 & 4; standards across experimental centres

Table 2: Approaches for data integration and their potential.

Abbreviations: AM – analysis method; exp(s) – experiment(s); HCA – human cell atlas;

MT – measurement type; smps – samples; TCGA – The Cancer Genome Atlas

from this scenario.

Another scenario arises when aiming at a general increase in sample sizes, so as to generalize (and statistically corroborate) observations. The increase in generality may further support the construction of a reference system, such as a cell atlas, the existence of which can support decisive speed-ups when classifying cells or cell states, investigated in subsequent experiments (see section 3.3). Increasing sample sizes often means that data is raised across multiple experiments of identical setup, for example experimental replicates possibly raised in different laboratories, such that statistically accounting for batch effects is a decisive factor. See Approach 2 in Figure 6 and Table 2 for respective methodological challenges.

Yet another scenario manifests when trying to unravel complexity and coordination of intracellular biological processes, as well as their mutual dependencies, so as to draw a comprehensive picture of a single cell. In this, an optimal setup is to raise data from just one single cell across multiple experiments referring to different types of measurements, such as scDNA-seq, scRNA-seq, possibly further augmented by measurements of chromatin accessibility, gene methylation, proteins or metabolites. See Approach 3 in Figure 6 and Table 2 for this scenario.

Co-measuring different and possibly concurring types of quantities, for example scRNA-seq and scDNA-seq [Kong et al., 2019], in just one single cell can be experimentally challenging or even just impossible at this point in time. An exit strategy to this problem is to raise a population of cells that is coherent in terms of cell type and state. One then spreads the different measurements across several single cells, all of which are drawn from this population. Upon having applied the different measurements on different single cells, one needs to combine the data

raised in a way that is biologically meaningful, respecting that each measurement stems from a different cell. Note that this approach encompasses the possibility to raise data both from single cells, and from bulks of cells. An example for the latter are bulk sequencing derived genotypes which one uses for imputation of missing values or the quantification of data that have remained uncertain in single cells that stem from the same population as the bulk. The integration of different types of data raised across multiple single cells, possibly including bulk data, casts issues that deserve attention in their own right (see Approach 4 in Figure 6 and Table 2), because these issues can substantially differ from the methods referring to Approach 3.

The most comprehensive goal, finally, may be to gain deeper insight into the complexity of (intra-) cellular circuits, and to chart their variability across time, tissues, and populations. Mapping cellular circuits in this comprehensive manner requires to take complementary and concurring measurements in single cells and across multiple single cells, possibly also across time, tissues and populations. Approach 5 in Figure 6 and Table 2 deals with this holistic approach to examining single cells. The ultimate goal is to comprehensively characterize biological systems, which requires to operate at the single-cell level, because one would not gain sufficient insight otherwise.

The challenges just outlined in terms of Approaches 1-5 in Figure 6 and Table 2 all are affected by the issues that influence single-cell data analysis in general, namely: (i) the varying resolution levels that are of interest depending on the research question at hand (section 2.1); (ii) the uncertainty of any measurements and how to quantify it for and during the analyses (section 2.2) and (iii) the scaling of single-cell methodology to more

cells and more features measured at once (section 2.3). All of these further compound the most important challenge in the integration of single-cell data: to link data from the different sources in a way that is biologically meaningful and supports the intended analysis. It is an immediate insight that the maps that describe how data from the different sources is linked, increase in complexity on increasing amounts of samples, time points and types of measurements (Figure 6, Table 2): Linking multiple samples referring to the same quantity measured within one experiment (Approach 1 in Figure 6 and Table 2) or across several experiments (Approach 2) needs to account for batch effects. Of course, whenever possible, batch effects should be minimized by establishing (global) standards affecting experimental centres worldwide to streamline common initiatives. Nevertheless, even if standards have been successfully established, additional validation of, for example, assignments of cells to types and states may be required.

The integration of measurements on multiple quantities (such as scRNA-seq and scDNA-seq) raised in one single cell (Approach 3) needs to account for dependencies if phenomena are concurrent. An illustrative example is to measure copy number variation (through scDNA-seq) or methylation so as to investigate their effects on RNA levels (measured through scRNA-seq).

Linking multiple types of measurement across different cells from the same cell population (Approach 4) may require the grouping of cells after experiments have been performed, because only then does disturbing variability among the (prior to the experiment assumed coherent) different cells become evident. An example is to group cells based on commonalities or differences in their genotype profile, having become evident only after the application of a scDNA-seq experiment.

Any assumptions that underlie these possible groupings need to resist thorough statistical testing and functional validation.

6.1.1 Status

For *unsupervised clustering* (Approach 0 in Figure 6 and Table 2), method development is a well-established field. Remaining challenges have already been identified systematically, see Duò et al. [2018], Freytag et al. [2018], Kiselev et al. [2019].

For *integrating multiple datasets of the same measurement type across different samples in one experiment* (Approach 1), a few approaches are available. See for example MNN [Haghverdi et al., 2018], and the methodologies included in the Seurat package [Satija et al., 2015, Butler et al., 2018b, Stuart et al., 2018]. For the challenges and promises referring to the integration of sc-seq data that vary in terms of spatial and temporal origin, see the discussions in the section 3.5 and section 5.3 below.

For *integrating multiple datasets of the same measurement type across experiments* (Approach 2), mapping cells to reference datasets such as the Human Cell Atlas [Regev et al., 2017] are currently emerging as the most promising strategy. We refer the reader to more particular and detailed discussions in section 3.3. If applicable reference systems are not available (note that the human cell atlas is not yet fully operable), assembling cell type clusters from different experiments is a reasonable strategy, as implemented by several recently published tools [Zhang et al., 2018, Barkas et al., 2018, Gao et al., 2018, Kiselev et al., 2018, Park et al., 2018, Wagner and Yanai, 2018, Boufeia et al., 2019, Johansen and Quon, 2019, Johnson et al., 2019].

The integration of data raised from one cell, referring to multiple types of measurements (Approach 3) is described in some particular

experimental protocols that address the issue [Macaulay et al., 2017]. These focus on combining scDNA-seq and scRNA-seq (Dey et al. [2015], Macaulay et al. [2016, 2017]), methylation data and scRNA-seq [Angermueller et al., 2016], or even all of scRNA-seq, scDNA-seq, methylation and chromatin accessibility data [Clark et al., 2018], or targeted queries on a cell's methylation, transcription (scRNA-seq) and genotype status (sc-GEM, Cheow et al. [2016]). Beyond these single-cell specific approaches, bulk approaches that address the integration of data from different types of experiments have the potential to be leveraged to account for single-cell specific noise characteristics or adapted to also qualify for corresponding single-cell analyses (MOFA, Arge-laguet et al. [2018]), DIABLO [Rohart et al., 2017b, Singh et al., 2018] and MINT [Rohart et al., 2017a]).

For the integration of different measurements performed on several cells all of which stem from a population of cells that is coherent with respect to the intended analysis (Approach 4), technologies such as 10X ge-nomics [Zheng et al., 2017] for scRNA-seq and direct library preparation (DLP, Zahn et al. [2017b]) for scDNA-seq establish an ex-perimental basis. As above-mentioned, the greater analytical challenge is to, upon hav-ing performed experiments, identify subpop-ulations that had hitherto remained invis-ible, and whose identification is crucial so as to not combine different types of data in mistaken ways. An example for this are the identification of cancer clones although single cells had been sampled from identi-cal tumor tissue—only performing scDNA-seq experiments can definitively reveal the clonal structure of a tumor. If one wishes to correctly link mutation with transcription profiles—the latter of which are examined via scRNA-seq experiments—ignoring the clonal structure of a tumor would be misleading.

Several analytical methods that address this problem have recently emerged: (i) clonealign [Campbell et al., 2019] assumes a copy-num-ber dosage effect on transcription to assign gene expression states to clones. (ii) cardelino [McCarthy et al., 2018] aligns clone-specific SNVs in scRNA-seq to those inferred from bulk exome data to infer clone-specific ex-pression patterns. (iii) MATCHER [Welch et al., 2017] uses manifold alignment to com-bine scM&T-seq [Angermueller et al., 2016] with sc-GEM [Cheow et al., 2016], leverag-ing the common set of loci. All of these methods are based on biologically coherent assumptions on how to summarize measure-ments across different types and samples in a reasonable way, despite their different physi-cal origin.

6.1.2 Open problems

Experimental technologies that deal with tak-ing measurements of different kinds on one single cell (Approach 3 in Figure 6 and Ta-ble 2) are on the rise and will allow to as-say more cells at higher fidelity and reduced cost. Yet, however, many methods for evau-lating combinations of different types of mea-surements performed on one single cell have not been in the focus. It is to be expected that the corresponding open problems will be-come more urgent. As an example, consider combined measurements of scDNA-seq and scRNA-seq, where one uses the transcripts de-rived from the latter to impute missing values in the genotype profile derived from the first.

While this may make Approach 4 look as if becoming gradually obsolete, the advances with respect to Approach 3 and the corre-sponding advances in terms of the resolution of how intracellular measurements of different types are linked with one another will benefit from ground work on Approach 4. Further, work using Approach 4 will mean a boost for

reference systems, such as cell atlases (see also Approach 2), because our understanding of the link between the different substrates measured will improve. As an example consider how gene expression increases on increasing genomic copy number, known as measurement linkage [Loper et al., 2019], are important to account for in such a reference system. This, in turn, will yield techniques that map different cellular quantities with greater accuracy, eventually allowing analyses at higher resolution and finer granularity. As a consequence, approaches that address taking different measurement across different cells from the same population (Approach 4) will deliver more finegrained results, hence also thanks to these approaches being easier to perform and being more cost efficient, likely will not experience a loss in popularity.

As just mentioned, advances with respect to Approach 3 and 4 will be partially based on advances in terms of mappings that connect cells across their types and states, see Approach 2. With combinations of measurement types gradually being shifted in the focus of attention, extensions of Approach 2 (which predominantly addresses how to connect different cells based on a single measurement) are necessary. These extensions will have to address how to connect different cells also in terms of multiple types of measurements, or even combinations thereof, such as integrative genotype-expression-profiles (raised by evaluating combined experiments on both scRNA-seq and scDNA-seq, for example), which points out the need for improvements addressing Approach 5.

Amounts of material that underlie most measurements will remain tiny, oftentimes limited by the amounts within a single cell and by a limited number of cells available from a particular cell population. This means that one overarching theme will persist: that the analyses we have just discussed will suf-

fer from missing entire views—samples, time points, or measurement types missing entirely at the time of training models or mapping quantities on one another. This will add to the difficulties in terms of missing data one experiences in non-integrative approaches.

6.2 Challenge XII: Validating and benchmarking analysis tools for single-cell measurements

With the advances in sc-seq and other single-cell technologies, more and more analysis tools become available for researchers, and even more are being developed and will be published in the near future. Thus, the need for datasets and methods that support systematic benchmarking and evaluation of these tools is becoming more pressing. To be useful and reliable, algorithms and pipelines should be able to pass the following quality control tests: (i) They should produce the expected results (e.g. reconstruct phylogenies, estimate differential expressions or cluster the data) of high quality and outperform existing methods, if such methods exist. (ii) They should be robust to high levels of sequencing noise and technological biases, including PCR bias, allele dropout and chimeric signals. In any case, benchmarking should be conducted in a systematic way, following established recommendations [Mangul et al., 2019, Weber et al., 2019].

Evaluation of tool performance requires benchmarking datasets with known ground truth. Such data should include cell populations with known genomic compositions and population structures, i.e. where frequencies of clones and alleles are known. Currently, such datasets are scarce—with some notable exceptions [Grün et al., 2014, Tian et al., 2019]—because generating them in genuine

laboratory settings is time-, labor- and cost-intensive. Experimental benchmark datasets for evolutionary analysis of single-cell populations are even harder to obtain, as they require follow-up samples with known information about evolutionary trajectories and developmental times. With lack of time-resolved measurements, only anecdotal evidence exists on, for instance, how the accuracy of phylogenetic inferences is affected by data quality. Availability of such gold-standard datasets would benefit single-cell genomics research enormously.

Due to aforementioned difficulties, the most affordable sources of benchmarking and validation data are in silico simulations. Simulations provide ground truth test examples that can be rapidly and cost-effectively generated under different assumptions. However, development of reliable simulation tools require design and implementation of models which capture the essence of underlying biological processes and technological details of single-cell technologies and high-throughput sequencing platforms, establishing single-cell data simulation as a methodologically involved challenge.

6.2.1 Status

Recent studies [Soneson and Robinson, 2018, Saelens et al., 2019] show that systematic benchmarking of different single-cell analysis methodologies has begun. However, to the best of our knowledge, there is still a shortage of single-cell data simulation tools. Many single-cell data analysis packages include their own ad hoc data simulators [Vallejos et al., 2015, Korthauer et al., 2016a, Lun et al., 2016, Lun and Marioni, 2017, Jahn et al., 2016, Satas and Raphael, 2018, Rizzetto et al., 2017, Köster et al., 2017]. However, these simulators are usually not available as separate tools or even as a source code, tailored to specific

problems studied in corresponding papers and sometimes not comprehensively documented, thus limiting their utility for the broad research community. Furthermore, since such simulators are used only as auxiliary subroutines inside particular projects and are not published as stand-alone tools, they themselves are usually not evaluated, and therefore the accuracy of their reflection of real biological and technological processes remain unclear. There are few exceptions known to us, including the tools Splatter [Zappia et al., 2017], powsimR [Vieth et al., 2017], and SymSim [Zhang et al., 2019d], which provide frameworks for simulation of scRNA-seq data and whose accuracy has been validated by comparison of its results with real data. For single-cell phylogenomics, cancer genome evolution simulators are being designed [Semeraro et al., 2018, Xia et al., 2018, Meng and Chen, 2018].

6.2.2 Open problems

Simulation tools mostly concentrate on differential expression analysis, while comprehensive simulation methods for other important aspects of sc-seq analysis are still to be developed. In particular, to the best of our knowledge, no such tool is available for scDNA-seq data.

With single-cell phylogenomics, one would like to assess the accuracy of methods for phylogenetic inference and subclone identification, or the power of population genetics methods for estimating parameters of interest (e.g. tests for selection and epistatic interactions in cancer, see section 5.3). To this end, realistic and comprehensive (w.r.t. the evolutionary phenomena) simulation tools are required.

Another interesting computational problem is development of tools for validation of simulated sc-seq datasets themselves by their com-

parison with real data using a comprehensive set of biological parameters. The first such tool for scRNA-seq data is countsimQC [Soneson and Robinson, 2017], but similar tools for scDNA-seq data are needed. Finally, most of the simulators concentrate on modeling of biologically meaningful data, while ignoring or simplifying models for sc-seq errors and artifacts.

Another important challenge in single-cell analysis tool validation is the selection of comprehensive evaluation metrics, which should be used for comparison of different analysis results with each other and with the ground truth. For single-cell data it is particularly complicated, since many analysis tools deal with heterogeneous clone populations, which possesses multiple biological characteristics to be inferred and analyzed. Development of a single measure which captures several of these characteristics is complicated, and in many cases impossible. For example, validation of tools for imputation of cellular and transcriptional heterogeneity should simultaneously evaluate two measures: (i) how close are the reconstructed and true cellular genomic profiles and (ii) how close are reconstructed and true SNV/haplotype frequency distributions. Development of synthetic measures which capture several such characteristics (e.g. based on utilization of earth mover's distance [Knyazev et al., 2018]) is highly important.

When simulating datasets in general, the circularity of simulating and inferring parameters under the same—possibly simplistic model—should be critically assessed, as should potential biases. Thus, further evaluation on empirical datasets for which some ground truth is known will be invaluable. Ideally, all single-cell analysis fields should define a standard set of benchmark datasets that will allow for assessing and comparing methods or come up with a regular data analysis chal-

lenge. This approach has been very successful, e.g. in protein structure prediction⁴ and metagenomic analyses⁵. A first step in this direction was the recent single-cell transcriptomics DREAM challenge⁶.

Acknowledgements

We are deeply grateful to the Lorentz Center for hosting the workshop “Single Cell Data Science: Making Sense of Data from Billions of Single Cells” (4–8 June 2018). In particular, we would like to thank the Lorentz Center staff, who turned organizing and attending the workshop into a great pleasure. For a week, the authors of this review came together—researchers from the fields of statistics and medicine, computer science and biology, and any combinations thereof. In interactive workshop sessions, we brought together our knowledge of single-cell analyses, ranging from the wet-lab to the server cluster, from statistical models to algorithms, from cancer biology to evolutionary genetics. During these sessions, we formulated an initial set of challenges that was further systematized and refined in the following months, and substantiated with extensive literature research of the respective state-of-the-art for this review.

Acronyms

CNV copy number variation. 22, 24, 25, 27, 28

MSA multiple sequence alignment. 26, 27, 30

⁴<http://predictioncenter.org/>

⁵<https://data.cami-challenge.org>

⁶<https://www.synapse.org/#!Synapse:syn15665609/wiki/582909>

1 **sc-seq** single-cell sequencing. 3–6, 29, 36,
2 38–40

3 **scDNA-seq** single-cell DNA sequencing. 3,
4 6–8, 16, 20–25, 28–32, 34–40

5 **SCDS** Single Cell Data Science. 4, 7

6 **scRNA-seq** single-cell RNA sequencing. 3,
7 5–9, 11–14, 16–18, 22, 24, 32, 34–40

8 **SNV** single nucleotide variation. 6, 20, 22–
9 25, 27, 28, 37, 40

10 **WGA** whole genome amplification. 20, 22–25

11 References

12 Andre J Aberer, Kassian Kobert, and Alexan-
13 dros Stamatakis. ExaBayes: massively par-
14 allel bayesian tree inference for the whole-
15 genome era. *Mol. Biol. Evol.*, 31(10):2553–
16 2556, October 2014.

17 Ahmet Acar, Daniel Nichol, Javier
18 Fernandez-Mateos, George D. Cress-
19 well, Iros Barozzi, Sung Pil Hong,
20 Inmaculada Spiteri, Mark Stubbs, Rose-
21 mary Burke, Adam Stewart, Georgios
22 Vlachogiannis, Carlo C. Maley, Luca
23 Magnani, Nicola Valeri, Udai Banerji, and
24 Andrea Sottoriva. Exploiting evolution-
25 ary herding to control drug resistance
26 in cancer. *bioRxiv*, page 566950, March
27 2019. doi: 10.1101/566950. URL
28 [https://www.biorxiv.org/content/10.](https://www.biorxiv.org/content/10.1101/566950v1)
29 [1101/566950v1](https://www.biorxiv.org/content/10.1101/566950v1).

30 Sumon Ahmed, Magnus Rattray, and
31 Alexis Boukouvalas. GrandPrix: scaling
32 up the Bayesian GPLVM for single-
33 cell data. *Bioinformatics*, 35(1):47–
34 54, January 2019. ISSN 1367-4803.
35 doi: 10.1093/bioinformatics/bty533.
36 URL <https://academic.oup.com/>

[bioinformatics/article/35/1/47/](https://academic.oup.com/bioinformatics/article/35/1/47/5047752)
5047752. 37 38

Philipp M Altrock, Lin L Liu, and Franziska 39
Michor. The mathematics of cancer: in- 40
tegrating quantitative models. *Nat. Rev.* 41
Cancer, 15(12):730–745, December 2015. 42

Matthew Amodio, David van Dijk, Krish- 43
nan Srinivasan, William S Chen, Hussein 44
Mohsen, Kevin R Moon, Allison Campbell, 45
Yujiao Zhao, Xiaomei Wang, Manjunatha 46
Venkataswamy, Anita Desai, V Ravi, Priti 47
Kumar, Ruth Montgomery, Guy Wolf, and 48
Smita Krishnaswamy. Exploring Single- 49
Cell data with deep multitasking neural 50
networks. January 2019. 51

Benedict Anchang, Tom D. P. Hart, Sean C. 52
Bendall, Peng Qiu, Zach Bjornson, Michael 53
Linderman, Garry P. Nolan, and Sylvia K. 54
Plevritis. Visualization and cellular 55
hierarchy inference of single-cell data 56
using SPADE. *Nature Protocols*, 11(7): 57
1264–1279, July 2016. ISSN 1754-2189. 58
doi: 10.1038/nprot.2016.066. URL [http:](http://www.nature.com/nprot/journal/v11/n7/full/nprot.2016.066.html)
59 [//www.nature.com/nprot/journal/v11/](http://www.nature.com/nprot/journal/v11/n7/full/nprot.2016.066.html)
60 [n7/full/nprot.2016.066.html](http://www.nature.com/nprot/journal/v11/n7/full/nprot.2016.066.html). 61

Tallulah S. Andrews and Martin Hem- 62
berg. False signals induced by single-cell 63
imputation. *F1000Research*, 7:1740, 64
March 2019. ISSN 2046-1402. doi: 65
10.12688/f1000research.16613.2. URL 66
[https://f1000research.com/articles/](https://f1000research.com/articles/7-1740/v2)
67 [7-1740/v2](https://f1000research.com/articles/7-1740/v2). 68

Michael Angelo, Sean C. Bendall, Rachel 69
Finck, Matthew B. Hale, Chuck Hitzman, 70
Alexander D. Borowsky, Richard M. Lev- 71
enson, John B. Lowe, Scot D. Liu, Shuchun 72
Zhao, Yasodha Natkunam, and Garry P. 73
Nolan. Multiplexed ion beam imaging of 74
human breast tumors. *Nature Medicine*, 20 75
(4):436–442, April 2014. ISSN 1546-170X. 76
doi: 10.1038/nm.3488. 77

- 1 Christof Angermueller, Stephen J. Clark,
2 Heather J. Lee, Iain C. Macaulay, Mabel J.
3 Teng, Tim Xiaoming Hu, Felix Krueger,
4 Sebastien Smallwood, Chris P. Ponting,
5 Thierry Voet, Gavin Kelsey, Oliver Ste-
6 gle, and Wolf Reik. Parallel single-cell se-
7 quencing links transcriptional and epige-
8 netic heterogeneity. *Nature Methods*, 13(3):
9 229–232, March 2016. ISSN 1548-7105. doi:
10 10.1038/nmeth.3728.
- 11 Christof Angermueller, Heather J. Lee,
12 Wolf Reik, and Oliver Stegle. Deep-
13 CpG: accurate prediction of single-cell
14 DNA methylation states using deep learn-
15 ing. *Genome Biology*, 18(1):67, April
16 2017. ISSN 1474-760X. doi: 10.1186/
17 s13059-017-1189-z. URL [https://doi.](https://doi.org/10.1186/s13059-017-1189-z)
18 [org/10.1186/s13059-017-1189-z](https://doi.org/10.1186/s13059-017-1189-z).
- 19 Ricard Argelaguet, Britta Velten, Damien
20 Arnol, Sascha Dietrich, Thorsten Zenz,
21 John C. Marioni, Florian Buettner, Wolf-
22 gang Huber, and Oliver Stegle. Multi-
23 Omics Factor Analysis—a framework for
24 unsupervised integration of multi-omics
25 data sets. *Molecular Systems Biology*,
26 14(6):e8124, June 2018. ISSN 1744-
27 4292, 1744-4292. doi: 10.15252/msb.
28 20178124. URL [http://msb.embopress.](http://msb.embopress.org/content/14/6/e8124)
29 [org/content/14/6/e8124](http://msb.embopress.org/content/14/6/e8124).
- 30 C Arisdakessian, O Poirion, B Yunits,
31 X Zhu, and L Garmire. DeepIm-
32 pute: an accurate, fast and scalable
33 deep neural network method to im-
34 pute single-cell RNA-Seq data. *bioRxiv*,
35 2018. URL [https://www.biorxiv.org/](https://www.biorxiv.org/content/10.1101/353607v1.abstract)
36 [content/10.1101/353607v1.abstract](https://www.biorxiv.org/content/10.1101/353607v1.abstract).
- 37 Nona Arneson, Simon Hughes, Richard Houl-
38 ston, and Susan Done. Whole-Genome am-
39 plification by improved primer extension
40 preamplification PCR (I-PEP-PCR). *CSH*
41 *Protoc.*, 2008:db.prot4921, January 2008.
- Damien Arnol, Denis Schapiro, Bernd Bo- 42
denmiller, Julio Saez-Rodriguez, and Oliver 43
Stegle. Modelling cell-cell interactions from 44
spatial molecular data with spatial variance 45
component analysis, 2018. 46
- Daniel L. Ayres. *Research And Applica- 47
tion Of Parallel Computing Algorithms For 48
Statistical Phylogenetic Inference*. PhD 49
thesis, University of Maryland, 2017. 50
URL [http://drum.lib.umd.edu/handle/](http://drum.lib.umd.edu/handle/1903/19951) 51
[1903/19951](http://drum.lib.umd.edu/handle/1903/19951). 52
- Elham Azizi, Sandhya Prabhakaran, Ambrose 53
Carr, and Dana Pe’er. Bayesian Infer- 54
ence for Single-cell Clustering and Imput- 55
ing. *Genomics and Computational Biol- 56
ogy*, 3(1):46, January 2017. ISSN 2365- 57
7154. doi: 10.18547/gcb.2017.vol3.iss1.e46. 58
URL [https://genomicscomputbiol.org/](https://genomicscomputbiol.org/ojs/index.php/GCB/article/view/46) 59
[ojs/index.php/GCB/article/view/46](https://genomicscomputbiol.org/ojs/index.php/GCB/article/view/46). 60
- Rhonda Bacher and Christina Kendzierski. 61
Design and computational analysis 62
of single-cell RNA-sequencing ex- 63
periments. *Genome Biology*, 17(1): 64
63, April 2016. ISSN 1474-760X. 65
doi: 10.1186/s13059-016-0927-y. 66
URL [https://doi.org/10.1186/](https://doi.org/10.1186/s13059-016-0927-y) 67
[s13059-016-0927-y](https://doi.org/10.1186/s13059-016-0927-y). 68
- Md Bahadur Badsha, Rui Li, Boxiang Liu, 69
Yang I Li, Min Xian, Nicholas E Banovich, 70
and Audrey Qiuyan Fu. Imputation of 71
single-cell gene expression with an autoen- 72
coder neural network. December 2018. 73
- Bjorn Bakker, Aaron Taudt, Mirjam E 74
Belderbos, David Porubsky, Diana C J 75
Spierings, Tristan V de Jong, Nancy 76
Halsema, Hinke G Kazemier, Karina 77
Hoekstra-Wakker, Allan Bradley, Eveline S 78
J M de Bont, Anke van den Berg, Victor 79
Guryev, Peter M Lansdorp, Maria Colomé- 80
Tatché, and Floris Foiyer. Single-cell se- 81
quencing reveals karyotype heterogeneity in 82

- 1 murine and human malignancies. *Genome*
2 *Biol.*, 17(1):115, May 2016.
- 3 Nikolas Barkas, Viktor Petukhov, Daria Niko-
4 laeva, Yaroslav Lozinsky, Samuel Demhar-
5 ter, Konstantin Khodosevich, and Peter V.
6 Kharchenko. Wiring together large single-
7 cell RNA-seq sample collections. *bioRxiv*,
8 page 460246, November 2018. doi: 10.1101/
9 460246. URL [https://www.biorxiv.org/](https://www.biorxiv.org/content/10.1101/460246v1)
10 [content/10.1101/460246v1](https://www.biorxiv.org/content/10.1101/460246v1).
- 11 Nico Battich, Thomas Stoeger, and Lucas
12 Pelkmans. Control of transcript variabil-
13 ity in single mammalian cells. *Cell*, 163(7):
14 1596–1610, December 2015.
- 15 Benedikt Bauer, Reiner Siebert, and Arne
16 Traulsen. Cancer initiation with epistatic
17 interactions between driver and passenger
18 mutations. *J. Theor. Biol.*, 358:52–60, Oc-
19 tober 2014.
- 20 Niko Beerenwinkel, Tibor Antal, David
21 Dingli, Arne Traulsen, Kenneth W Kinzler,
22 Victor E Velculescu, Bert Vogelstein, and
23 Martin A Nowak. Genetic progression and
24 the waiting time to cancer. *PLoS Comput.*
25 *Biol.*, 3(11):e225, November 2007.
- 26 Graham R. Bignell, Thomas Santarius, Jes-
27 sica C.M. Pole, Adam P. Butler, Janet
28 Perry, Erin Pleasance, Chris Greenman,
29 Andrew Menzies, Sheila Taylor, Sarah
30 Edkins, Peter Campbell, Michael Quail,
31 Bob Plumb, Lucy Matthews, Kirsten
32 McLay, Paul A.W. Edwards, Jane Rogers,
33 Richard Wooster, P. Andrew Futreal,
34 and Michael R. Stratton. Architec-
35 tures of somatic genomic rearrangement
36 in human cancer amplicons at sequence-
37 level resolution. *Genome Research*, 17
38 (9):1296–1303, 2007. doi: 10.1101/gr.
39 6522707. URL [http://genome.cshlp.](http://genome.cshlp.org/content/17/9/1296.abstract)
40 [org/content/17/9/1296.abstract](http://genome.cshlp.org/content/17/9/1296.abstract).
- Konstantin A Blagodatskikh, Vladimir M
Kramarov, Ekaterina V Barsova, Alexey V
Garkovenko, Dmitriy S Shcherbo, An-
drew A Shelenkov, Vera V Ustinova,
Maria R Tokarenko, Simon C Baker, Ta-
tiana V Kramarova, and Konstantin B Ig-
natov. Improved DOP-PCR (iDOP-PCR):
A robust and simple WGA method for effi-
cient amplification of low copy number ge-
nomic DNA. *PLoS One*, 12(9):e0184507,
September 2017.
- L Blanco, A Bernad, J M Lázaro, G Martín,
C Garmendia, and M Salas. Highly ef-
ficient DNA synthesis by the phage phi
29 DNA polymerase. symmetrical mode of
DNA replication. *J. Biol. Chem.*, 264(15):
8935–8940, May 1989.
- Craig L. Bohrsen, Alison R. Barton,
Michael A. Lodato, Rachel E. Rodin,
Lovelace J. Luquette, Vinay V. Viswanad-
ham, Doga C. Gulhan, Isidro Cortés-
Ciriano, Maxwell A. Sherman, Min-
seok Kwon, Michael E. Coulter, Alon
Galor, Christopher A. Walsh, and
Peter J. Park. Linked-read analysis
identifies mutations in single-cell DNA-
sequencing data. *Nature Genetics*,
page 1, March 2019. ISSN 1546-1718.
doi: 10.1038/s41588-019-0366-2. URL
[https://www.nature.com/articles/](https://www.nature.com/articles/s41588-019-0366-2)
s41588-019-0366-2.
- Katerina Boufea, Sohan Seth, and Nizar N.
Batada. scID: Identification of equivalent
transcriptional cell populations across
single cell RNA-seq data using discrim-
inant analysis. *bioRxiv*, page 470203,
January 2019. doi: 10.1101/470203. URL
[https://www.biorxiv.org/content/10.](https://www.biorxiv.org/content/10.1101/470203v2)
1101/470203v2.
- Ivana Bozic, Tibor Antal, Hisashi Ohtsuki,
Hannah Carter, Dewey Kim, Sining Chen,

- 1 Rachel Karchin, Kenneth W Kinzler, Bert
2 Vogelstein, and Martin A Nowak. Accu-
3 mulation of driver and passenger muta-
4 tions during tumor progression. *Proc. Natl.*
5 *Acad. Sci. U. S. A.*, 107(43):18545–18550,
6 October 2010.
- 7 Ivana Bozic, Jeffrey M Gerold, and Mar-
8 tin A Nowak. Quantifying clonal and sub-
9 clonal passenger mutations in cancer evolu-
10 tion. *PLoS Comput. Biol.*, 12(2):e1004731,
11 February 2016.
- 12 James A. Briggs, Caleb Weinreb, Daniel E.
13 Wagner, Sean Megason, Leonid Peshkin,
14 Marc W. Kirschner, and Allon M. Klein.
15 The dynamics of gene expression in ver-
16 tebrate embryogenesis at single-cell res-
17 olution. *Science*, 360(6392):eaar5780,
18 June 2018. ISSN 0036-8075, 1095-
19 9203. doi: 10.1126/science.aar5780.
20 URL [http://science.sciencemag.org/](http://science.sciencemag.org/content/360/6392/eaar5780)
21 [content/360/6392/eaar5780](http://science.sciencemag.org/content/360/6392/eaar5780).
- 22 Jane Bromley, James W. Bentz, Léon Bot-
23 tou, Isabelle Guyon, Yann Lecun, Cliff
24 Moore, Eduard Säckinger, and Roopak
25 Shah. Signature verification using a
26 “siamese” time delay neural network.
27 *International Journal of Pattern Recog-*
28 *nition and Artificial Intelligence*, 07(04):
29 669–688, August 1993. ISSN 0218-0014.
30 doi: 10.1142/S0218001493000339. URL
31 [https://www.worldscientific.com/](https://www.worldscientific.com/doi/10.1142/S0218001493000339)
32 [doi/10.1142/S0218001493000339](https://www.worldscientific.com/doi/10.1142/S0218001493000339).
- 33 Robert V Bruggner, Bernd Bodenmiller,
34 David L Dill, Robert J Tibshirani, and
35 Garry P Nolan. Automated identification
36 of stratifying signatures in cellular subpop-
37 ulations. *Proc. Natl. Acad. Sci. U. S. A.*,
38 111(26):E2770–7, July 2014.
- 39 Jason D. Buenrostro, Beijing Wu, Ulrike M.
40 Litzénburger, Dave Ruff, Michael L.
Gonzales, Michael P. Snyder, Howard Y.
Chang, and William J. Greenleaf. Single-
cell chromatin accessibility reveals princi-
ples of regulatory variation. *Nature*, 523
(7561):486–490, July 2015. ISSN 1476-
4687. doi: 10.1038/nature14590. URL
[https://www.nature.com/articles/](https://www.nature.com/articles/nature14590)
[nature14590](https://www.nature.com/articles/nature14590).
- Jason D. Buenrostro, M. Ryan Corces,
Caleb A. Lareau, Beijing Wu, Alicia N.
Schip, Martin J. Aryee, Ravindra Ma-
jeti, Howard Y. Chang, and William J.
Greenleaf. Integrated Single-Cell Analy-
sis Maps the Continuous Regulatory Land-
scape of Human Hematopoietic Differen-
tiation. *Cell*, 173(6):1535–1548.e16, 2018.
ISSN 1097-4172. doi: 10.1016/j.cell.2018.
03.074.
- Florian Buettner, Naruemon Pratanwanich,
Davis J McCarthy, John C Marioni, and
Oliver Stegle. f-scLVM: scalable and ver-
satile factor analysis for single-cell RNA-
seq. *Genome biology*, 18(1):212, Novem-
ber 2017. ISSN 1465-6906. doi: 10.1186/
s13059-017-1334-8. URL [http://dx.doi.](http://dx.doi.org/10.1186/s13059-017-1334-8)
[org/10.1186/s13059-017-1334-8](http://dx.doi.org/10.1186/s13059-017-1334-8).
- Andrew Butler, Paul Hoffman, Peter Smib-
ert, Efthymia Papalexi, and Rahul Satija.
Integrating single-cell transcriptomic data
across different conditions, technologies,
and species. *Nat. Biotechnol.*, 36(5):411–
420, June 2018a.
- Andrew Butler, Paul Hoffman, Peter Smib-
ert, Efthymia Papalexi, and Rahul Satija.
Integrating single-cell transcriptomic data
across different conditions, technologies,
and species. *Nature Biotechnology*, 36(5):
411–420, May 2018b. ISSN 1546-1696.
doi: 10.1038/nbt.4096. URL [https://](https://www.nature.com/articles/nbt.4096)
www.nature.com/articles/nbt.4096.

- 1 Kieran R. Campbell and Christopher
2 Yau. Order Under Uncertainty: Robust
3 Differential Expression Analysis Using
4 Probabilistic Models for Pseudotime In-
5 ference. *PLOS Computational Biology*, 12
6 (11):e1005212, November 2016. ISSN 1553-
7 7358. doi: 10.1371/journal.pcbi.1005212.
8 URL [https://journals.plos.org/
9 ploscompbiol/article?id=10.1371/
10 journal.pcbi.1005212](https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005212).

11 Kieran R. Campbell and Christopher Yau.
12 Uncovering pseudotemporal trajectories
13 with covariates from single cell and bulk
14 expression data. *Nature Communications*,
15 9(1):2442, June 2018. ISSN 2041-1723.
16 doi: 10.1038/s41467-018-04696-6. URL
17 [https://www.nature.com/articles/
18 s41467-018-04696-6](https://www.nature.com/articles/s41467-018-04696-6).

19 Kieran R. Campbell, Adi Steif, Emma Laks,
20 Hans Zahn, Daniel Lai, Andrew McPher-
21 son, Hossein Farahani, Farhia Kabeer,
22 Ciara O’Flanagan, Justina Biele, Jazmine
23 Brimhall, Beixi Wang, Pascale Walters,
24 IMAXT Consortium, Alexandre Bouchard-
25 Côté, Samuel Aparicio, and Sohrab P.
26 Shah. clonealign: statistical integra-
27 tion of independent single-cell RNA and
28 DNA sequencing data from human can-
29 cers. *Genome Biology*, 20(1):54, March
30 2019. ISSN 1474-760X. doi: 10.1186/
31 s13059-019-1645-z. URL [https://doi.
32 org/10.1186/s13059-019-1645-z](https://doi.org/10.1186/s13059-019-1645-z).

33 Junyue Cao, Jonathan S. Packer, Vijay Ra-
34 mani, Darren A. Cusanovich, Chau Huynh,
35 Riza Daza, Xiaojie Qiu, Choli Lee, Scott N.
36 Furlan, Frank J. Steemers, Andrew Adey,
37 Robert H. Waterston, Cole Trapnell, and
38 Jay Shendure. Comprehensive single-
39 cell transcriptional profiling of a multi-
40 cellular organism. *Science*, 357(6352):
41 661–667, August 2017. ISSN 0036-8075,
42 1095-9203. doi: 10.1126/science.aam8940.
URL [http://science.sciencemag.org/
43 content/357/6352/661](http://science.sciencemag.org/content/357/6352/661).
44
45 Junyue Cao, Darren A. Cusanovich, Vijay
46 Ramani, Delasa Aghamirzaie, Hannah A.
47 Pliner, Andrew J. Hill, Riza M. Daza,
48 Jose L. McFaline-Figueroa, Jonathan S.
49 Packer, Lena Christiansen, Frank J.
50 Steemers, Andrew C. Adey, Cole Trap-
51 nell, and Jay Shendure. Joint profil-
52 ing of chromatin accessibility and gene
53 expression in thousands of single cells.
54 *Science*, 361(6409):1380–1385, Septem-
55 ber 2018. ISSN 0036-8075, 1095-
56 9203. doi: 10.1126/science.aau0730.
57 URL [https://science.sciencemag.org/
58 content/361/6409/1380](https://science.sciencemag.org/content/361/6409/1380).

59 Junyue Cao, Malte Spielmann, Xiaojie Qiu,
60 Xingfan Huang, Daniel M. Ibrahim, An-
61 drew J. Hill, Fan Zhang, Stefan Mundlos,
62 Lena Christiansen, Frank J. Steemers, Cole
63 Trapnell, and Jay Shendure. The single-
64 cell transcriptional landscape of mam-
65 malian organogenesis. *Nature*, 566(7745):
66 496, February 2019a. ISSN 1476-4687.
67 doi: 10.1038/s41586-019-0969-x. URL
68 [https://www.nature.com/articles/
69 s41586-019-0969-x](https://www.nature.com/articles/s41586-019-0969-x).

70 Zhi-Jie Cao, Lin Wei, Shen Lu, De-Chang
71 Yang, and Ge Gao. Cell BLAST: Search-
72 ing large-scale scRNA-seq database via
73 unbiased cell embedding. *bioRxiv*, page
74 587360, March 2019b. doi: 10.1101/
75 587360. URL [https://www.biorxiv.org/
76 content/10.1101/587360v1](https://www.biorxiv.org/content/10.1101/587360v1).

77 Anna K Casasent, Aislyn Schalck, Ruli
78 Gao, Emi Sei, Annalyssa Long, William
79 Pangburn, Tod Casasent, Funda Meric-
80 Bernstam, Mary E Edgerton, and
81 Nicholas E Navin. Multiclonal invasion in
82 breast tumors identified by topographic

- 1 single cell sequencing. *Cell*, 172(1-2):
2 205–217.e12, January 2018.
- 3 Chong Chen, Changjing Wu, Linjie Wu,
4 Yishu Wang, Minghua Deng, and Ruibin
5 Xi. scRMD: Imputation for single cell
6 RNA-seq data via robust matrix de-
7 composition. November 2018. URL
8 [https://www.biorxiv.org/content/10.](https://www.biorxiv.org/content/10.1101/459404v2)
9 [1101/459404v2](https://www.biorxiv.org/content/10.1101/459404v2).
- 10 Chongyi Chen, Dong Xing, Longzhi Tan,
11 Heng Li, Guangyu Zhou, Lei Huang, and
12 X Sunney Xie. Single-cell whole-genome
13 analyses by linear amplification via trans-
14 poson insertion (LIANTI). *Science*, 356
15 (6334):189–194, April 2017.
- 16 Huidong Chen, Luca Albergante,
17 Jonathan Y. Hsu, Caleb A. Lareau,
18 Giosuè Lo Bosco, Jihong Guan, Shuigeng
19 Zhou, Alexander N. Gorban, Daniel E.
20 Bauer, Martin J. Aryee, David M. Lan-
21 genau, Andrei Zinovyev, Jason D. Buen-
22 rostro, Guo-Cheng Yuan, and Luca Pinello.
23 Single-cell trajectories reconstruction, ex-
24 ploration and mapping of omics data with
25 STREAM. *Nature Communications*, 10
26 (1):1903, April 2019. ISSN 2041-1723.
27 doi: 10.1038/s41467-019-09670-4. URL
28 [https://www.nature.com/articles/](https://www.nature.com/articles/s41467-019-09670-4)
29 [s41467-019-09670-4](https://www.nature.com/articles/s41467-019-09670-4).
- 30 Kok Hao Chen, Alistair N Boettiger, Jef-
31 frey R Moffitt, Siyuan Wang, and Xiaowei
32 Zhuang. RNA imaging. spatially resolved,
33 highly multiplexed RNA profiling in sin-
34 gle cells. *Science*, 348(6233):aaa6090, April
35 2015.
- 36 Mengjie Chen and Xiang Zhou. VIPER:
37 variability-preserving imputation for accu-
38 rate gene expression recovery in single-cell
39 RNA sequencing studies. *Genome Biol.*, 19
40 (1):196, November 2018.
- Lih Feng Cheow, Elise T. Courtois, Yuliana
Tan, Ramya Viswanathan, Qiaorui Xing,
Rui Zhen Tan, Daniel S. W. Tan, Paul Rob-
son, Yui-Han Loh, Stephen R. Quake, and
William F. Burkholder. Single-cell multi-
modal profiling reveals cellular epigenetic
heterogeneity. *Nature Methods*, 13(10):833–
836, October 2016. ISSN 1548-7105. doi:
10.1038/nmeth.3961. URL [https://www.](https://www.nature.com/articles/nmeth.3961)
[nature.com/articles/nmeth.3961](https://www.nature.com/articles/nmeth.3961).
- Cariad Chester and Holden T Maecker. Algo-
rithmic tools for mining High-Dimensional
cytometry data. *J. Immunol.*, 195(3):773–
779, August 2015.
- Hyunghoon Cho, Bonnie Berger, and Jian
Peng. Generalizable and scalable visualiza-
tion of Single-Cell data using neural net-
works. *Cell Syst*, 7(2):185–191.e4, August
2018.
- Simone Ciccolella, Mauricio Soto Gomez,
Murray Patterson, Gianluca Della Ve-
dova, Iman Hajirasouliha, and Paola
Bonizzoni. Inferring Cancer Progres-
sion from Single-cell Sequencing while Al-
lowing Mutation Losses. *bioRxiv*, page
268243, April 2018. doi: 10.1101/
268243. URL [https://www.biorxiv.org/](https://www.biorxiv.org/content/10.1101/268243v2)
[content/10.1101/268243v2](https://www.biorxiv.org/content/10.1101/268243v2).
- Stephen J. Clark, Ricard Argelaguet,
Chantierint-Andreas Kapourani,
Thomas M. Stubbs, Heather J. Lee, Celia
Alda-Catalinas, Felix Krueger, Guido San-
guinetti, Gavin Kelsey, John C. Marioni,
Oliver Stegle, and Wolf Reik. scNMT-seq
enables joint profiling of chromatin accessi-
bility DNA methylation and transcription
in single cells. *Nature Communications*, 9
(1):781, February 2018. ISSN 2041-1723.
doi: 10.1038/s41467-018-03149-4. URL
[https://www.nature.com/articles/](https://www.nature.com/articles/s41467-018-03149-4)
[s41467-018-03149-4](https://www.nature.com/articles/s41467-018-03149-4).

- 1 Germán Corredor, Xiangxue Wang, Yu Zhou,
2 Cheng Lu, Pingfu Fu, Konstantinos Syri-
3 gos, David L Rimm, Michael Yang, Ed-
4 uardo Romero, Kurt A Schalper, Vam-
5 sidhar Velcheti, and Anant Madabhushi.
6 Spatial architecture and arrangement of
7 Tumor-Infiltrating lymphocytes for pre-
8 dicting likelihood of recurrence in Early-
9 Stage Non-Small cell lung cancer. *Clin.*
10 *Cancer Res.*, September 2018.
- 11 Alexandra Cretu and Peter C Brooks. Im-
12 pact of the non-cellular tumor microenvi-
13 ronment on metastasis: potential thera-
14 peutic and imaging opportunities. *J. Cell.*
15 *Physiol.*, 213(2):391–402, November 2007.
- 16 Nicola Crosetto, Magda Bienko, and Alexan-
17 der van Oudenaarden. Spatially resolved
18 transcriptomics and beyond. *Nat. Rev.*
19 *Genet.*, 16(1):57–66, January 2015.
- 20 Darren A. Cusanovich, Riza Daza, Andrew
21 Adey, Hannah A. Pliner, Lena Chris-
22 tiansen, Kevin L. Gunderson, Frank J.
23 Steemers, Cole Trapnell, and Jay Shen-
24 dure. Multiplex single cell profiling of chro-
25 matin accessibility by combinatorial cellu-
26 lar indexing. *Science (New York, N.Y.)*,
27 348(6237):910–914, May 2015. ISSN 1095-
28 9203. doi: 10.1126/science.aab1601.
- 29 Darren A. Cusanovich, James P. Redding-
30 ton, David A. Garfield, Riza M. Daza, De-
31 lasa Aghamirzaie, Raquel Marco-Ferreres,
32 Hannah A. Pliner, Lena Christiansen, Xi-
33 aojie Qiu, Frank J. Steemers, Cole Trap-
34 nell, Jay Shendure, and Eileen E. M. Fur-
35 long. The cis-regulatory dynamics of em-
36 bryonic development at single-cell resolu-
37 tion. *Nature*, 555(7697):538–542, March
38 2018. ISSN 1476-4687. doi: 10.1038/
39 nature25981. URL [https://www.nature.](https://www.nature.com/articles/nature25981)
40 [com/articles/nature25981](https://www.nature.com/articles/nature25981).
- Sayantana Das, Gonçalo R Abecasis, and
Brian L Browning. Genotype Impu-
tation from Large Reference Panels. *Annual review of genomics and hu-*
man genetics, 19:73–96, August 2018.
ISSN 1527-8204, 1545-293X. doi:
10.1146/annurev-genom-083117-021602.
URL [http://dx.doi.org/10.1146/](http://dx.doi.org/10.1146/annurev-genom-083117-021602)
[annurev-genom-083117-021602](http://dx.doi.org/10.1146/annurev-genom-083117-021602).
- Soma Datta, Lavina Malhotra, Ryan Dicker-
son, Scott Chaffee, Chandan K Sen, and
Sashwati Roy. Laser capture microdissec-
tion: Big data from small samples. *Histol.*
Histopathol., 30(11):1255–1269, November
2015.
- Alexander Davis, Ruli Gao, and Nicholas
Navin. Tumor evolution: Linear, branch-
ing, neutral or punctuated? *Biochim. Bio-*
phys. Acta, 1867(2):151–161, April 2017.
- Carl G. de Boer and Aviv Regev. BROCK-
MAN: deciphering variance in epige-
nomic regulators by k-mer factoriza-
tion. *BMC Bioinformatics*, 19(1):253, July
2018. ISSN 1471-2105. doi: 10.1186/
s12859-018-2255-6. URL [https://doi.](https://doi.org/10.1186/s12859-018-2255-6)
[org/10.1186/s12859-018-2255-6](https://doi.org/10.1186/s12859-018-2255-6).
- Charles F A de Bourcy, Iwijn De Vlam-
inck, Jad N Kanbar, Jianbin Wang, Charles
Gawad, and Stephen R Quake. A quantita-
tive comparison of single-cell whole genome
amplification methods. *PLoS One*, 9(8):
e105585, August 2014.
- Frank B Dean, Seiyu Hosono, Linhua Fang,
Xiaohong Wu, A Fawad Faruqi, Patricia
Bray-Ward, Zhenyu Sun, Qiuling Zong,
Yuefen Du, Jing Du, Mark Driscoll, Wan-
min Song, Stephen F Kingsmore, Michael
Egholm, and Roger S Lasken. Compre-
hensive human genome amplification using
multiple displacement amplification. *Proc.*

- 1 *Natl. Acad. Sci. U. S. A.*, 99(8):5261–5266,
2 April 2002.
- 3 Yue Deng, Feng Bao, Qionghai Dai, Lani F
4 Wu, and Steven J Altschuler. Scalable anal-
5 ysis of cell-type composition from single-
6 cell transcriptomics using deep recurrent
7 learning. *Nature methods*, March 2019.
8 ISSN 1548-7091, 1548-7105. doi: 10.1038/
9 s41592-019-0353-7. URL [https://doi.](https://doi.org/10.1038/s41592-019-0353-7)
10 [org/10.1038/s41592-019-0353-7](https://doi.org/10.1038/s41592-019-0353-7).
- 11 Erica AK DePasquale, Kyle Ferchen, Stuart
12 Hay, H. Leighton Grimes, and Nathan
13 Salomonis. cellHarmony: Cell-level
14 matching and comparison of single-cell
15 transcriptomes. *bioRxiv*, page 412080,
16 January 2019. doi: 10.1101/412080. URL
17 [https://www.biorxiv.org/content/10.](https://www.biorxiv.org/content/10.1101/412080v4)
18 [1101/412080v4](https://www.biorxiv.org/content/10.1101/412080v4).
- 19 Siddharth S. Dey, Lennart Kester, Bastiaan
20 Spanjaard, Magda Bienko, and Alexan-
21 der van Oudenaarden. Integrated genome
22 and transcriptome sequencing of the same
23 cell. *Nature Biotechnology*, 33(3):285–289,
24 March 2015. ISSN 1546-1696. doi: 10.1038/
25 nbt.3129. URL [https://www.nature.](https://www.nature.com/articles/nbt.3129)
26 [com/articles/nbt.3129](https://www.nature.com/articles/nbt.3129).
- 27 David van Dijk, Roshan Sharma, Juozas
28 Nainys, Kristina Yim, Pooja Kathail,
29 Ambrose J. Carr, Cassandra Burdziak,
30 Kevin R. Moon, Christine L. Chaf-
31 fer, Diwakar Pattabiraman, Brian Bie-
32 rie, Linas Mazutis, Guy Wolf, Smita Krish-
33 naswamy, and Dana Pe’er. Recovering
34 Gene Interactions from Single-Cell Data
35 Using Data Diffusion. *Cell*, 174(3):716–
36 729.e27, July 2018. ISSN 0092-8674,
37 1097-4172. doi: 10.1016/j.cell.2018.05.
38 061. URL [https://www.cell.com/cell/](https://www.cell.com/cell/abstract/S0092-8674(18)30724-4)
39 [abstract/S0092-8674\(18\)30724-4](https://www.cell.com/cell/abstract/S0092-8674(18)30724-4).
- 40 Jiarui Ding, Anne Condon, and Sohrab P
41 Shah. Interpretable dimensionality reduc-
tion of single cell transcriptome data with
deep generative models. *Nat. Commun.*, 9
(1):2002, May 2018.
- Xiao Dong, Lei Zhang, Brandon Milholland,
Moonsook Lee, Alexander Y. Maslov, Tao
Wang, and Jan Vijg. Accurate iden-
tification of single-nucleotide variants in
whole-genome-amplified single cells. *Na-
ture Methods*, 14(5):491–493, May 2017.
ISSN 1548-7105. doi: 10.1038/nmeth.
4227. URL [https://www.nature.com/](https://www.nature.com/articles/nmeth.4227)
articles/nmeth.4227.
- Angelo Duò, Mark D Robinson, and Char-
lotte Soneson. A systematic performance
evaluation of clustering methods for single-
cell RNA-seq data. *F1000Res.*, 7, July
2018.
- G Durif, L Modolo, J E Mold, S Lambert-
Lacroix, and F Picard. Probabilistic
Count Matrix Factorization for Single
Cell Expression Data Analysis. *Bioin-
formatics*, March 2019. ISSN 1367-4803,
1367-4811. doi: 10.1093/bioinformatics/
btz177. URL [http://dx.doi.org/10.](http://dx.doi.org/10.1093/bioinformatics/btz177)
1093/bioinformatics/btz177.
- Daniel Edsgård, Per Johnsson, and Rickard
Sandberg. Identification of spatial expres-
sion trends in single-cell gene expression
data. *Nat. Methods*, 15(5):339–342, May
2018.
- Peter Eirew, Adi Steif, Jaswinder Khattra,
Gavin Ha, Damian Yap, Hossein Fara-
hani, Karen Gelmon, Stephen Chia, Colin
Mar, Adrian Wan, Emma Laks, Justina
Biele, Karey Shumansky, Jamie Rosner,
Andrew McPherson, Cydney Nielsen,
Andrew J. L. Roth, Calvin Lefebvre, Ali
Bashashati, Camila de Souza, Celia Siu,
Radhouane Aniba, Jazmine Brimhall,
Arusha Oloumi, Tomo Osako, Alejandra

- 1 Bruna, Jose L. Sandoval, Teresa Algara,
2 Wendy Greenwood, Kaston Leung, Hong-
3 wei Cheng, Hui Xue, Yuzhuo Wang,
4 Dong Lin, Andrew J. Mungall, Richard
5 Moore, Yongjun Zhao, Julie Lorette, Long
6 Nguyen, David Huntsman, Connie J.
7 Eaves, Carl Hansen, Marco A. Marra, Car-
8 los Caldas, Sohrab P. Shah, and Samuel
9 Aparicio. Dynamics of genomic clones
10 in breast cancer patient xenografts at
11 single-cell resolution. *Nature*, 518(7539):
12 422–426, February 2015. ISSN 0028-0836.
13 doi: 10.1038/nature13952. URL [http:](http://www.nature.com/nature/journal/v518/n7539/full/nature13952.html)
14 [//www.nature.com/nature/journal/](http://www.nature.com/nature/journal/v518/n7539/full/nature13952.html)
15 [v518/n7539/full/nature13952.html](http://www.nature.com/nature/journal/v518/n7539/full/nature13952.html).
16 Mohammed El-Kebir. SPhyR: tumor
17 phylogeny estimation from single-cell
18 sequencing data under loss and er-
19 ror. *Bioinformatics*, 34(17):i671–i679,
20 September 2018. ISSN 1367-4803.
21 doi: 10.1093/bioinformatics/bty589.
22 URL [https://academic.oup.com/](https://academic.oup.com/bioinformatics/article/34/17/i671/5093218)
23 [bioinformatics/article/34/17/i671/](https://academic.oup.com/bioinformatics/article/34/17/i671/5093218)
24 [5093218](https://academic.oup.com/bioinformatics/article/34/17/i671/5093218).
25 Nils Eling, Arianne C. Richard, Sylvia
26 Richardson, John C. Marioni, and
27 Catalina A. Vallejos. Correcting the
28 Mean-Variance Dependency for Differ-
29 ential Variability Testing Using Single-Cell
30 RNA Sequencing Data. *Cell Systems*, 7(3):
31 284–294.e12, September 2018. ISSN 2405-
32 4712. doi: 10.1016/j.cels.2018.06.011. URL
33 [https://www.cell.com/cell-systems/](https://www.cell.com/cell-systems/abstract/S2405-4712(18)30278-3)
34 [abstract/S2405-4712\(18\)30278-3](https://www.cell.com/cell-systems/abstract/S2405-4712(18)30278-3).
35 Chee-Huat Linus Eng, Michael Lawson,
36 Qian Zhu, Ruben Dries, Noushin Koulana,
37 Yodai Takei, Jina Yun, Christopher
38 Cronin, Christoph Karp, Guo-Cheng
39 Yuan, and Long Cai. Transcriptome-
40 scale super-resolved imaging in tissues by
41 RNA seqFISH+. *Nature*, 568(7751):
42 235, April 2019. ISSN 1476-4687.
doi: 10.1038/s41586-019-1049-y. URL
[https://www.nature.com/articles/](https://www.nature.com/articles/s41586-019-1049-y)
[s41586-019-1049-y](https://www.nature.com/articles/s41586-019-1049-y).
Gökçen Eraslan, Lukas M. Simon, Maria
Mircea, Nikola S. Mueller, and Fabian J.
Theis. Single-cell RNA-seq denois-
ing using a deep count autoencoder.
Nature Communications, 10(1):390,
January 2019. ISSN 2041-1723. doi:
10.1038/s41467-018-07931-2. URL
[https://www.nature.com/articles/](https://www.nature.com/articles/s41467-018-07931-2)
[s41467-018-07931-2](https://www.nature.com/articles/s41467-018-07931-2).
Nuria Estévez-Gómez, Tamara Prieto, Amy
Guillaumet-Adkins, Holger Heyn, Sonia
Prado-López, and David Posada. Com-
parison of single-cell whole-genome am-
plification strategies. *bioRxiv*, page
443754, October 2018. doi: 10.1101/
443754. URL [https://www.biorxiv.org/](https://www.biorxiv.org/content/10.1101/443754v1)
[content/10.1101/443754v1](https://www.biorxiv.org/content/10.1101/443754v1).
Jean Fan, Hae-Ock Lee, Soohyun Lee, Da-
Eun Ryu, Semin Lee, Catherine Xue,
Seok Jin Kim, Kihyun Kim, Nikolaos
Barkas, Peter J Park, Woong-Yang Park,
and Peter V Kharchenko. Linking tran-
scriptional and genetic tumor heterogeneity
through allele analysis of single-cell RNA-
seq data. *Genome Res.*, 28(8):1217–1227,
August 2018.
Jeffrey A. Farrell, Yiqun Wang, Saman-
tha J. Riesenfeld, Karthik Shekhar,
Aviv Regev, and Alexander F. Schier.
Single-cell reconstruction of develop-
mental trajectories during zebrafish
embryogenesis. *Science*, 360(6392):
eaar3131, June 2018. ISSN 0036-8075,
1095-9203. doi: 10.1126/science.aar3131.
URL [http://science.sciencemag.org/](http://science.sciencemag.org/content/360/6392/eaar3131)
[content/360/6392/eaar3131](http://science.sciencemag.org/content/360/6392/eaar3131).
Joseph Felsenstein. Evolutionary trees from
DNA sequences: A maximum likelihood

- 1 approach. *J. Mol. Evol.*, 17(6):368–376,
2 1981.
- 3 Greg Finak, Andrew McDavid, Masanao
4 Yajima, Jingyuan Deng, Vivian Gersuk,
5 Alex K. Shalek, Chloe K. Slichter, Han-
6 nah W. Miller, M. Juliana McElrath,
7 Martin Prlic, Peter S. Linsley, and Raphael
8 Gottardo. MAST: a flexible statistical
9 framework for assessing transcriptional
10 changes and characterizing heterogene-
11 ity in single-cell RNA sequencing data.
12 *Genome Biology*, 16, 2015. ISSN 1474-
13 7596. doi: 10.1186/s13059-015-0844-5.
14 URL [https://www.ncbi.nlm.nih.gov/
15 pmc/articles/PMC4676162/](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4676162/).
- 16 Christopher T. Fincher, Omri Wurtzel,
17 Thom de Hoog, Kellie M. Kravarik, and
18 Peter W. Reddien. Cell type transcriptome
19 atlas for the planarian *Schmidtea mediter-*
20 *ranea*. *Science*, 360(6391):eaaq1736,
21 May 2018. ISSN 0036-8075, 1095-
22 9203. doi: 10.1126/science.aaq1736.
23 URL [http://science.sciencemag.org/
24 content/360/6391/eaaq1736](http://science.sciencemag.org/content/360/6391/eaaq1736).
- 25 William Fletcher and Ziheng Yang. The ef-
26 fect of insertions, deletions, and alignment
27 errors on the branch-site test of positive se-
28 lection. *Mol. Biol. Evol.*, 27(10):2257–2267,
29 October 2010.
- 30 Jasmine Foo, Kevin Leder, and Franziska Mi-
31 chor. Stochastic dynamics of cancer initi-
32 ation. *Phys. Biol.*, 8(1):015002, February
33 2011.
- 34 Joshua M. Francis, Cheng-Zhong Zhang,
35 Cecile L. Maire, Joonil Jung, Veronica E.
36 Manzo, Viktor A. Adalsteinsson, Heather
37 Homer, Sam Haidar, Brendan Blumenstiel,
38 Chandra Sekhar Pedomallu, Azra H.
39 Ligon, J. Christopher Love, Matthew
40 Meyerson, and Keith L. Ligon. EGFR
Variant Heterogeneity in Glioblastoma
Resolved through Single-Nucleus Sequenc-
ing. *Cancer Discovery*, 4(8):956–971,
August 2014. ISSN 2159-8274, 2159-8290.
doi: 10.1158/2159-8290.CD-13-0879.
URL [http://cancerdiscovery.
aacrjournals.org/content/4/8/956](http://cancerdiscovery.aacrjournals.org/content/4/8/956).
- Saskia Freytag, Luyi Tian, Ingrid Lönnst-
edt, Milica Ng, and Melanie Bahlo. Comparison of clustering tools in R for
medium-sized 10x Genomics single-cell
RNA-sequencing data. *F1000Research*,
7, December 2018. ISSN 2046-1402. doi:
10.12688/f1000research.15809.2. URL
[https://www.ncbi.nlm.nih.gov/pmc/
articles/PMC6124389/](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6124389/).
- Wolf H Fridman, Jérôme Galon, Marie-
Caroline Dieu-Nosjean, Isabelle Cremer,
Sylvain Fisson, Diane Damotte, Franck
Pagès, Eric Tartour, and Catherine Sautès-
Fridman. Immune infiltration in human
cancer: Prognostic significance and disease
control. In Glenn Dranoff, editor, *Cancer
Immunology and Immunotherapy*, pages 1–
24. Springer Berlin Heidelberg, Berlin, Hei-
delberg, 2011.
- Yusi Fu, Chunmei Li, Sijia Lu, Wenxiong
Zhou, Fuchou Tang, X Sunney Xie, and
Yanyi Huang. Uniform and accurate
single-cell sequencing based on emulsion
whole-genome amplification. *Proc. Natl.
Acad. Sci. U. S. A.*, 112(38):11923–11928,
September 2015.
- Dan Gao, Feng Jin, Min Zhou, and Yuyang
Jiang. Recent advances in single cell ma-
nipulation and biochemical analysis on mi-
crofluidics. *Analyst*, 144(3):766–781, Jan-
uary 2019.
- Xin Gao, Deqing Hu, Madelaine Gogol,
and Hua Li. ClusterMap: Com-
paring analyses across multiple Single

- 1 Cell RNA-Seq profiles. *bioRxiv*, page 41
2 331330, June 2018. doi: 10.1101/ 42
3 331330. URL [https://www.biorxiv.org/](https://www.biorxiv.org/content/10.1101/331330v2)
4 [content/10.1101/331330v2](https://www.biorxiv.org/content/10.1101/331330v2).
- 5 Tyler Garvin, Robert Aboukhalil, Jude 43
6 Kendall, Timour Baslan, Gurinder S At- 44
7 wal, James Hicks, Michael Wigler, and 45
8 Michael C Schatz. Interactive analysis and 46
9 assessment of single-cell copy-number vari- 47
10 ations. *Nat. Methods*, 12(11):1058–1060, 48
11 November 2015. URL <https://doi.org/10.1186/s12859-018-2226-y>.
- 12 Charles Gawad, Winston Koh, and Stephen R 49
13 Quake. Single-cell genome sequencing: cur- 50
14 rent state of the science. *Nat. Rev. Genet.*, 51
15 17(3):175–188, March 2016.
- 16 Farzad Ghaznavi, Andrew Evans, Anant 52
17 Madabhushi, and Michael Feldman. Digital 53
18 imaging in pathology: whole-slide imaging 54
19 and beyond. *Annu. Rev. Pathol.*, 8:331–
20 359, January 2013.
- 21 Charlotte Giesen, Hao A. O. Wang, Denis 55
22 Schapiro, Nevena Zivanovic, Andrea Ja- 56
23 cobs, Bodo Hattendorf, Peter J. Schüffler, 57
24 Daniel Grolimund, Joachim M. Buhmann, 58
25 Simone Brandt, Zsuzsanna Varga, Peter J. 59
26 Wild, Detlef Günther, and Bernd Boden- 60
27 miller. Highly multiplexed imaging of tu- 61
28 mor tissues with subcellular resolution by 62
29 mass cytometry. *Nature Methods*, 11(4): 63
30 417–422, April 2014. ISSN 1548-7105. doi: 64
31 10.1038/nmeth.2869. URL [https://www.](https://www.nature.com/articles/nmeth.2869) 65
32 [nature.com/articles/nmeth.2869](https://www.nature.com/articles/nmeth.2869).
- 33 Yury Goltsev, Nikolay Samusik, Julia 66
34 Kennedy-Darling, Salil Bhate, Matthew 67
35 Hale, Gustavo Vazquez, Sarah Black, and 68
36 Garry P. Nolan. Deep Profiling of Mouse 69
37 Splenic Architecture with CODEX Multi- 70
38 plexed Imaging. *Cell*, 174(4):968–981.e15, 71
39 August 2018. ISSN 0092-8674. doi: 72
40 10.1016/j.cell.2018.07.010. URL [http://](http://www.sciencedirect.com/science/article/pii/S0092867418309048) 73
[www.sciencedirect.com/science/](http://www.sciencedirect.com/science/article/pii/S0092867418309048) 74
[article/pii/S0092867418309048](http://www.sciencedirect.com/science/article/pii/S0092867418309048). 75
76
77
78
79
80
Wuming Gong, Il-Youp Kwak, Pruthvi Pota, 43
Naoko Koyano-Nakagawa, and Daniel J. 44
Garry. DrImpute: imputing dropout 45
events in single cell RNA sequencing 46
data. *BMC Bioinformatics*, 19(1):220, June 47
2018. ISSN 1471-2105. doi: 10.1186/ 48
s12859-018-2226-y. URL [https://doi.](https://doi.org/10.1186/s12859-018-2226-y) 49
[org/10.1186/s12859-018-2226-y](https://doi.org/10.1186/s12859-018-2226-y). 50
R R Gray, O G Pybus, and M Salemi. Mea- 51
suring the temporal structure in Serially- 52
Sampled phylogenies. *Methods Ecol. Evol.*, 53
2(5):437–445, October 2011. 54
Anna Graybeal. Is it better to add taxa or 55
characters to a difficult phylogenetic prob- 56
lem? *Syst. Biol.*, 47(1):9–17, 1998. 57
Christopher Heje Grønbech, Maximil- 58
lian Fornitz Vording, Pascal N Timshel, 59
Casper Kaae Sønderby, Tune Hannes 60
Pers, and Ole Winther. scVAE: Vari- 61
ational auto-encoders for single-cell 62
gene expression data. May 2018. URL 63
[https://www.biorxiv.org/content/](https://www.biorxiv.org/content/early/2018/05/16/318295) 64
[early/2018/05/16/318295](https://www.biorxiv.org/content/early/2018/05/16/318295). 65
Dominic Grün, Lennart Kester, and Alexan- 66
der van Oudenaarden. Validation of noise 67
models for single-cell transcriptomics. *Na-* 68
ture Methods, 11(6):637–640, June 2014. 69
ISSN 1548-7105. doi: 10.1038/nmeth. 70
2930. URL [https://www.nature.com/](https://www.nature.com/articles/nmeth.2930) 71
[articles/nmeth.2930](https://www.nature.com/articles/nmeth.2930). 72
Martin Guillems, Charles-Antoine Dutertre, 73
Charlotte L. Scott, Naomi McGovern, 74
Dorine Sichien, Svetoslav Chakarov, Sofie 75
Van Gassen, Jinmiao Chen, Michael 76
Poidinger, Sofie De Prijck, Simon J. 77
Tavernier, Ivy Low, Sergio Erdal Irac, 78
Citra Nurfarah Mattar, Hermi Rizal Suma- 79
toh, Gillian Hui Ling Low, Tam John Kit 80

- 1 Chung, Dedrick Kok Hong Chan, Ker Kan
2 Tan, Tony Lim Kiat Hon, Even Fos-
3 sum, Bjarne Bogen, Mahesh Choolani,
4 Jerry Kok Yen Chan, Anis Larbi, Hervé
5 Luche, Sandrine Henri, Yvan Saeys,
6 Evan William Newell, Bart N. Lambrecht,
7 Bernard Malissen, and Florent Ginhoux.
8 Unsupervised High-Dimensional Analysis
9 Aligns Dendritic Cells across Tissues
10 and Species. *Immunity*, 45(3):669–684,
11 September 2016. ISSN 1074-7613. doi:
12 10.1016/j.immuni.2016.08.015. URL [http://www.sciencedirect.com/science/
13 article/pii/S1074761316303399](http://www.sciencedirect.com/science/article/pii/S1074761316303399).
14
- 15 Metin N Gurcan, Laura Boucheron, Ali Can,
16 Anant Madabhushi, Nasir Rajpoot, and
17 Bulent Yener. Histopathological image
18 analysis: A review. *IEEE Rev. Biomed.*
19 *Eng.*, 2:147, 2009.
- 20 Hiroshi Haeno, Mithat Gonen, Meghan B
21 Davis, Joseph M Herman, Christine A
22 Iacobuzio-Donahue, and Franziska Michor.
23 Computational modeling of pancreatic can-
24 cer reveals kinetics of metastasis suggesting
25 optimum treatment strategies. *Cell*, 148(1-
26 2):362–375, January 2012.
- 27 Christoph Hafemeister and Rahul Satija. Nor-
28 malization and variance stabilization of
29 single-cell RNA-seq data using regular-
30 ized negative binomial regression. March
31 2019. URL [https://www.biorxiv.org/
32 content/10.1101/576827v2](https://www.biorxiv.org/content/10.1101/576827v2).
- 33 Laleh Haghverdi, Maren Büttner, F. Alexan-
34 der Wolf, Florian Buettner, and Fabian J.
35 Theis. Diffusion pseudotime robustly
36 reconstructs lineage branching. *Nature*
37 *Methods*, 13(10):845–848, October 2016.
38 ISSN 1548-7105. doi: 10.1038/nmeth.
39 3971. URL [https://www.nature.com/
40 articles/nmeth.3971](https://www.nature.com/articles/nmeth.3971).
- Laleh Haghverdi, Aaron T. L. Lun, 41
Michael D. Morgan, and John C. Marionni. 42
Batch effects in single-cell RNA-sequencing 43
data are corrected by matching mutual 44
nearest neighbors. *Nature Biotechnology*, 45
36(5):421–427, 2018. ISSN 1546-1696. doi: 46
10.1038/nbt.4091. 47
- Xiaoping Han, Renying Wang, Yincong Zhou, 48
Lijiang Fei, Huiyu Sun, Shujing Lai, Assieh 49
Saadatpour, Ziming Zhou, Haide Chen, 50
Fang Ye, Daosheng Huang, Yang Xu, Wen- 51
tao Huang, Mengmeng Jiang, Xinyi Jiang, 52
Jie Mao, Yao Chen, Chenyu Lu, Jin Xie, 53
Qun Fang, Yibin Wang, Rui Yue, Tiefeng 54
Li, He Huang, Stuart H Orkin, Guo-Cheng 55
Yuan, Ming Chen, and Guoji Guo. Map- 56
ping the mouse cell atlas by Microwell-Seq. 57
Cell, 172(5):1091–1107.e17, February 2018. 58
- Andreas Heindl, Sidra Nawaz, and Yinyin 59
Yuan. Mapping spatial heterogeneity in 60
the tumor microenvironment: a new era for 61
digital pathology. *Lab. Invest.*, 95(4):377– 62
384, April 2015. 63
- Stephanie C. Hicks and Roger D. Peng. 64
Elements and Principles of Data Anal- 65
ysis. *arXiv:1903.07639 [stat]*, March 66
2019. URL [http://arxiv.org/abs/1903.](http://arxiv.org/abs/1903.07639)
67 07639. arXiv: 1903.07639. 68
- Stephanie C. Hicks, F. William Townes, 69
Mingxiang Teng, and Rafael A. Irizarry. 70
Missing data and technical variability 71
in single-cell RNA-sequencing exper- 72
iments. *Biostatistics*, 19(4):562–578, 73
October 2018. ISSN 1465-4644. doi: 74
10.1093/biostatistics/kxx053. URL [https: 75
://academic.oup.com/biostatistics/
76 article/19/4/562/4599254](https://academic.oup.com/biostatistics/article/19/4/562/4599254). 77
- Elad Hoffer and Nir Ailon. Deep Metric 78
Learning Using Triplet Network. In Aasa 79
Feragen, Marcello Pelillo, and Marco Loog, 80

- 1 editors, *Similarity-Based Pattern Recognition*, Lecture Notes in Computer Science,
2 pages 84–92. Springer International Pub-
3 lishing, 2015. ISBN 978-3-319-24261-3. 40
- 5 Ian H Holmes. Solving the master equation
6 for indels. *BMC Bioinformatics*, 18(1):255,
7 May 2017. 41
- 8 Chung-Chau Hon, Jay W. Shin, Piero Carn-
9 inci, and Michael J. T. Stubbington. 42
10 The Human Cell Atlas: Technical ap-
11 proaches and challenges. *Briefings in*
12 *Functional Genomics*, 17(4):283–294, July
13 2018. ISSN 2041-2649. doi: 10.1093/bfgp/
14 elx029. URL [https://academic.oup.](https://academic.oup.com/bfg/article/17/4/283/4571849)
15 [com/bfg/article/17/4/283/4571849](https://academic.oup.com/bfg/article/17/4/283/4571849). 43
- 16 Masahito Hosokawa, Yohei Nishikawa,
17 Masato Kogawa, and Haruko Takeyama. 44
18 Massively parallel whole genome amplifica-
19 tion for single-cell sequencing using droplet
20 microfluidics. *Sci. Rep.*, 7(1):5199, July
21 2017. 45
- 22 Yong Hou, Kui Wu, Xulian Shi, Fuqiang Li,
23 Luting Song, Hanjie Wu, Michael Dean,
24 Guibo Li, Shirley Tsang, Runze Jiang,
25 Xiaolong Zhang, Bo Li, Geng Liu, Ni-
26 harika Bedekar, Na Lu, Guoyun Xie, Han
27 Liang, Liao Chang, Ting Wang, Jianghao
28 Chen, Yingrui Li, Xiuqing Zhang, Huan-
29 ming Yang, Xun Xu, Ling Wang, and Jun
30 Wang. Comparison of variations detection
31 between whole-genome amplification meth-
32 ods used in single-cell resequencing. *Giga-*
33 *science*, 4:37, August 2015. 46
- 34 Qiwen Hu and Casey S Greene. Param-
35 eter tuning is a key part of dimension-
36 ality reduction via deep variational au-
37 toencoders for single cell RNA transcrip-
38 tomics. *Pacific Symposium on Biocomput-*
39 *ing*. *Pacific Symposium on Biocomputing*,
24:362–373, 2019. ISSN 2335-6936, 2335-
6928. URL [https://www.ncbi.nlm.nih.](https://www.ncbi.nlm.nih.gov/pubmed/30963075)
gov/pubmed/30963075. 47
- Lei Huang, Fei Ma, Alec Chapman, Sijia
Lu, and Xiaoliang Sunney Xie. Single-Cell
Whole-Genome amplification and sequenc-
ing: Methodology and applications. *Annu.*
Rev. Genomics Hum. Genet., 16:79–102,
June 2015. 48
- Mo Huang, Jingshu Wang, Eduardo Torre,
Hannah Dueck, Sydney Shaffer, Roberto
Bonasio, John I. Murray, Arjun Raj,
Mingyao Li, and Nancy R. Zhang. 49
SAVER: gene expression recovery for
single-cell RNA sequencing. *Nature Meth-*
ods, 15(7):539, July 2018. ISSN 1548-7105.
doi: 10.1038/s41592-018-0033-z. URL
[https://www.nature.com/articles/](https://www.nature.com/articles/s41592-018-0033-z)
[s41592-018-0033-z](https://www.nature.com/articles/s41592-018-0033-z). 50
- Joanna Hård, Ezeddin Al Hakim, Marie
Kindblom, Åsa K. Björklund, Bengt
Sennblad, Ilke Demirci, Marta Paterlini,
Pedro Reu, Erik Borgström, Patrik L.
Ståhl, Jakob Michaelsson, Jeff E. Mold,
and Jonas Frisén. Conbase: a software
for unsupervised discovery of clonal so-
matic mutations in single cells through read
phasing. *Genome Biology*, 20(1):68, April
2019. ISSN 1474-760X. doi: 10.1186/
s13059-019-1673-8. URL [https://doi.](https://doi.org/10.1186/s13059-019-1673-8)
org/10.1186/s13059-019-1673-8. 51
- T. Höllt, N. Pezzotti, V. van Unen, F. Kon-
ing, B. P. F. Lelieveldt, and A. Vilanova. 52
CyteGuide: Visual Guidance for Hierar-
chical Single-Cell Analysis. *IEEE Trans-*
actions on Visualization and Computer
Graphics, 24(1):739–748, January 2018.
ISSN 1077-2626. doi: 10.1109/TVCG.2017.
2744318. 53
- Giovanni Iacono, Elisabetta Mereu, Amy
Guillaumet-Adkins, Roser Corominas, Ivon 54

- 1 Cuscó, Gustavo Rodríguez-Esteban, Marta
2 Gut, Luis Alberto Pérez-Jurado, Ivo Gut,
3 and Holger Heyn. bigSCale: an analyti-
4 cal framework for big-scale single-cell data.
5 *Genome Res.*, 28(6):878–890, June 2018.
- 6 Humayun Irshad, Antoine Veillard, Ludovic
7 Roux, and Daniel Racocceanu. Methods
8 for nuclei detection, segmentation, and
9 classification in digital histopathology: A
10 Review—Current status and future poten-
11 tial, 2014.
- 12 Martin Jacobsen. *Point Process Theory and*
13 *Applications: Marked Point and Piecewise*
14 *Deterministic Processes*. Springer Science
15 & Business Media, December 2005.
- 16 Katharina Jahn, Jack Kuipers, and Niko
17 Beerenwinkel. Tree inference for single-cell
18 data. *Genome Biol.*, 17:86, May 2016.
- 19 Livnat Jerby-Arnon, Nadja Pfetzer, Yedael Y
20 Waldman, Lynn McGarry, Daniel James,
21 Emma Shanks, Brinton Seashore-Ludlow,
22 Adam Weinstock, Tamar Geiger, Paul A
23 Clemons, Eyal Gottlieb, and Eytan Rup-
24 pin. Predicting cancer-specific vulnerabil-
25 ity via data-driven detection of synthetic
26 lethality. *Cell*, 158(5):1199–1209, August
27 2014.
- 28 Zhicheng Ji and Hongkai Ji. TSCAN:
29 Pseudo-time reconstruction and evaluation
30 in single-cell RNA-seq analysis. *Nucleic*
31 *Acids Research*, 44(13):e117, 2016. ISSN
32 1362-4962. doi: 10.1093/nar/gkw430.
- 33 Nelson Johansen and Gerald Quon. scAlign:
34 a tool for alignment, integration and
35 rare cell identification from scRNA-seq
36 data. *bioRxiv*, page 504944, March
37 2019. doi: 10.1101/504944. URL
38 [https://www.biorxiv.org/content/10.](https://www.biorxiv.org/content/10.1101/504944v4)
39 [1101/504944v4](https://www.biorxiv.org/content/10.1101/504944v4).
- Brett E Johnson, Tali Mazor, Chibo Hong,
Michael Barnes, Koki Aihara, Cory Y
McLean, Shaun D Fouse, Shogo Ya-
mamoto, Hiroki Ueda, Kenji Tatsuno,
Saurabh Asthana, Llewellyn E Jalbert,
Sarah J Nelson, Andrew W Bollen, W Clay
Gustafson, Elise Charron, William A
Weiss, Ivan V Smirnov, Jun S Song,
Adam B Olshen, Soonmee Cha, Yongjun
Zhao, Richard A Moore, Andrew J
Mungall, Steven J M Jones, Martin Hirst,
Marco A Marra, Nobuhito Saito, Hiroyuki
Aburatani, Akitake Mukasa, Mitchel S
Berger, Susan M Chang, Barry S Taylor,
and Joseph F Costello. Mutational anal-
ysis reveals the origin and therapy-driven
evolution of recurrent glioma. *Science*, 343
(6167):189–193, January 2014.
- Travis S. Johnson, Tongxin Wang, Zhi
Huang, Christina Y. Yu, Yi Wu, Yatong
Han, Yan Zhang, Kun Huang, and Jie
Zhang. LAMBDA: Label Ambiguous
Domain Adaptation Dataset Integration
Reduces Batch Effects and Improves
Subtype Detection. *Bioinformatics*, April
2019. doi: 10.1093/bioinformatics/btz295.
URL [https://academic.oup.com/](https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btz295/5481958)
[bioinformatics/advance-article/](https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btz295/5481958)
[doi/10.1093/bioinformatics/btz295/](https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btz295/5481958)
[5481958](https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btz295/5481958).
- Altuna Akalin Jonathan Ronen. netsmooth:
Network-smoothing based imputation for
single cell RNA-seq. *F1000Res.*, 7, 2018.
- Min Jung, Daniel Wells, Jannette Rusch,
Suhaira Ahmad, Jonathan Marchini, Si-
mon R Myers, and Donald F Conrad. Uni-
fied single-cell analysis of testis gene regu-
lation and pathology in five mouse strains.
eLife, 8, June 2019. ISSN 2050-084X.
doi: 10.7554/eLife.43966. URL [http://](http://dx.doi.org/10.7554/eLife.43966)
dx.doi.org/10.7554/eLife.43966.

- 1 Melissa R Junttila and Frederic J de Sauvage.
2 Influence of tumour micro-environment
3 heterogeneity on therapeutic response. *Nature*, 501(7467):346–354, September 2013.
4
- 5 Hyun Min Kang, Meena Subramaniam, Sasha
6 Targ, Michelle Nguyen, Lenka Maliskova,
7 Elizabeth McCarthy, Eunice Wan, Simon
8 Wong, Lauren Byrnes, Cristina Lanata,
9 Rachel Gate, Sara Mostafavi, Alexander
10 Marson, Noah Zaitlen, Lindsey A
11 Criswell, and Chun Jimmie Ye. Multi-
12 plexed droplet single-cell RNA-sequencing
13 using natural genetic variation. *Nature*
14 *biotechnology*, 36(1):89–94, January
15 2018a. ISSN 1087-0156. doi: 10.1038/nbt.
16 4042. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5784859/>.
17
- 18 Hyun Min Kang, Meena Subramaniam, Sasha
19 Targ, Michelle Nguyen, Lenka Maliskova,
20 Elizabeth McCarthy, Eunice Wan, Simon
21 Wong, Lauren Byrnes, Cristina M Lanata,
22 Rachel E Gate, Sara Mostafavi, Alexander
23 Marson, Noah Zaitlen, Lindsey A Criswell,
24 and Chun Jimmie Ye. Multiplexed droplet
25 single-cell RNA-sequencing using natural
26 genetic variation. *Nat. Biotechnol.*, 36(1):
27 89–94, January 2018b.
- 28 Jurrian Kornelis de Kanter, Philip Lijnzaad,
29 Tito Candelli, Thanasis Margaritis, and
30 Frank Holstege. CHETAH: a selective, hi-
31 erarchical cell type identification method
32 for single-cell RNA sequencing. *bioRxiv*,
33 page 558908, February 2019. doi: 10.1101/
34 558908. URL <https://www.biorxiv.org/content/10.1101/558908v1>.
35
- 36 Nikos Karaikos, Philipp Wahle, Jonathan
37 Alles, Anastasiya Boltengagen, Salah Ay-
38 oub, Claudia Kipar, Christine Kocks, Niko-
39 laus Rajewsky, and Robert P Zinzen. The
40 drosophila embryo at single-cell transcrip-
tome resolution. *Science*, 358(6360):194–
199, October 2017a. 41 42
- 43 Nikos Karaikos, Philipp Wahle, Jonathan
44 Alles, Anastasiya Boltengagen, Salah Ay-
45 oub, Claudia Kipar, Christine Kocks, Niko-
46 laus Rajewsky, and Robert P. Zinzen. The
47 Drosophila embryo at single-cell transcrip-
48 tome resolution. *Science*, 358(6360):194–
49 199, October 2017b. ISSN 0036-8075,
50 1095-9203. doi: 10.1126/science.aan3235.
51 URL <http://science.sciencemag.org/content/358/6360/194>.
52
- 53 Ino D. Karemaker and Michiel Vermeulen.
54 Single-Cell DNA Methylation Profiling:
55 Technologies and Biological Applications.
56 *Trends in Biotechnology*, 36(9):952–965,
57 September 2018. ISSN 0167-7799, 1879-
58 3096. doi: 10.1016/j.tibtech.2018.04.002.
59 URL [https://www.cell.com/trends/biotechnology/abstract/S0167-7799\(18\)30115-X](https://www.cell.com/trends/biotechnology/abstract/S0167-7799(18)30115-X).
60 61
- 62 Lennart Kester and Alexander van Oudenaar-
63 den. Single-Cell transcriptomics meets lin-
64 eage tracing. *Cell Stem Cell*, May 2018.
- 65 Peter V Kharchenko, Lev Silberstein, and
66 David T Scadden. Bayesian approach
67 to single-cell differential expression analy-
68 sis. *Nature Methods*, 11(7):740–742, July
69 2014. ISSN 1548-7091. doi: 10.1038/
70 nmeth.2967. URL <http://www.nature.com/doifinder/10.1038/nmeth.2967>.
71
- 72 Kyu-Tae Kim, Hye Won Lee, Hae-Ock Lee,
73 Sang Cheol Kim, Yun Jee Seo, Woosung
74 Chung, Hye Hyeon Eum, Do-Hyun Nam,
75 Junhyong Kim, Kyeong Min Joo, and
76 Woong-Yang Park. Single-cell mRNA se-
77 quencing identifies subclonal heterogeneity
78 in anti-cancer drug responses of lung ade-
79 noma cells. *Genome Biol.*, 16:127,
80 June 2015.

- 1 Tae-Min Kim, Ruibin Xi, Lovelace J. Lu-
2 quette, Richard W. Park, Mark D. John-
3 son, and Peter J. Park. Functional
4 genomic analysis of chromosomal aber-
5 rations in a compendium of 8000 can-
6 cer genomes. *Genome Research*, 23(2):
7 217–227, 2013. doi: 10.1101/gr.140301.
8 112. URL [http://genome.cshlp.org/
9 content/23/2/217.abstract](http://genome.cshlp.org/content/23/2/217.abstract).
- 10 Marek Kimmel and David Axelrod. *Branch-*
11 *ing Processes in Biology*. Interdisci-
12 plinary Applied Mathematics. Springer-
13 Verlag, New York, 2 edition, 2015. ISBN
14 978-1-4939-1558-3. URL [https://www.
15 springer.com/gp/book/9781493915583](https://www.springer.com/gp/book/9781493915583).
- 16 Savvas Kinalis, Finn Cilius Nielsen, Ole
17 Winther, and Frederik Otzen Bagger.
18 Deconvolution of autoencoders to learn bi-
19 ological regulatory modules from single cell
20 mRNA sequencing data. *BMC bioinforma-*
21 *tics*, 20(1):379, July 2019. ISSN 1471-
22 2105. doi: 10.1186/s12859-019-2952-9.
23 URL [http://dx.doi.org/10.1186/
24 s12859-019-2952-9](http://dx.doi.org/10.1186/s12859-019-2952-9).
- 25 Vladimir Yu Kiselev, Andrew Yiu, and Mar-
26 tin Hemberg. scmap: projection of single-
27 cell RNA-seq data across data sets. *Na-*
28 *ture Methods*, 15(5):359–362, May 2018.
29 ISSN 1548-7105. doi: 10.1038/nmeth.
30 4644. URL [https://www.nature.com/
31 articles/nmeth.4644](https://www.nature.com/articles/nmeth.4644).
- 32 Vladimir Yu Kiselev, Tallulah S. Andrews,
33 and Martin Hemberg. Challenges in
34 unsupervised clustering of single-cell
35 RNA-seq data. *Nature Reviews Genetics*,
36 page 1, January 2019. ISSN 1471-0064.
37 doi: 10.1038/s41576-018-0088-9. URL
38 [https://www.nature.com/articles/
39 s41576-018-0088-9](https://www.nature.com/articles/s41576-018-0088-9).
- 40 Allon M. Klein, Linas Mazutis, Ilke Akar-
41 tuna, Naren Tallapragada, Adrian Veres,
Victor Li, Leonid Peshkin, David A. Weitz,
and Marc W. Kirschner. Droplet barcod-
ing for single-cell transcriptomics applied
to embryonic stem cells. *Cell*, 161(5):1187–
1201, May 2015. ISSN 1097-4172. doi:
10.1016/j.cell.2015.04.044.
- C A Klein, O Schmidt-Kittler, J A Schardt,
K Pantel, M R Speicher, and G Rieth-
müller. Comparative genomic hybridiza-
tion, loss of heterozygosity, and DNA se-
quence analysis of single cells. *Proc. Natl.*
Acad. Sci. U. S. A., 96(8):4494–4499, April
1999.
- Sergey Knyazev, Viachaslau Tsyvina, An-
drew Melnyk, Alexander Artyomenko, Ta-
tiana Malygina, Yuri B Porozov, Ellsworth
Campbell, William M Switzer, Pavel
Skums, and Alex Zelikovsky. CliqueSNV:
Scalable reconstruction of Intra-Host viral
populations from NGS reads, 2018.
- Bryan Kolaczowski and Joseph W Thornton.
A mixed branch length model of hetero-
taxy improves phylogenetic accuracy. *Mol.*
Biol. Evol., 25(6):1054–1066, June 2008.
- Daisuke Komura and Shumpei Ishikawa. Ma-
chine learning methods for histopatho-
logical image analysis. *Comput. Struct.*
Biotechnol. J., 16:34–42, February 2018.
- Say Li Kong, Huipeng Li, Joyce A Tai, Elise T
Courtois, Huay Mei Poh, Dawn Pingxi Lau,
Yu Xuan Haw, Narayanan Gopalakrishna
Iyer, Daniel Shao Weng Tan, Shyam Prab-
hakar, Dave Ruff, and Axel M Hillmer.
Concurrent Single-Cell RNA and targeted
DNA sequencing on an automated platform
for comeasurement of genomic and tran-
scriptomic signatures. *Clin. Chem.*, 65(2):
272–281, February 2019.
- Hazal Koptagel, Seong-Hwan Jun, and
Jens Lagergren. SCuPhr: A Prob-

- 1 abilistic Framework for Cell Lineage
2 Tree Reconstruction. *bioRxiv*, page
3 357442, June 2018. doi: 10.1101/
4 357442. URL [https://www.biorxiv.org/
5 content/early/2018/06/29/357442](https://www.biorxiv.org/content/early/2018/06/29/357442).
- 6 Keegan D Korthauer, Li-Fang Chu, Michael A
7 Newton, Yuan Li, James Thomson, Ron
8 Stewart, and Christina Kendzierski. A sta-
9 tistical approach for identifying differential
10 distributions in single-cell RNA-seq exper-
11 iments. *Genome Biol.*, 17(1):222, October
12 2016a.
- 13 Keegan D. Korthauer, Li-Fang Chu,
14 Michael A. Newton, Yuan Li, James
15 Thomson, Ron Stewart, and Christina
16 Kendzierski. A statistical approach for
17 identifying differential distributions in
18 single-cell RNA-seq experiments. *Genome
19 Biology*, 17(1):222, 2016b. ISSN 1474-
20 760X. doi: 10.1186/s13059-016-1077-y.
- 21 Johannes Köster, Myles Brown, and Xi-
22 aole Shirley Liu. A bayesian model for
23 single cell transcript expression analysis on
24 MERFISH data, September 2017.
- 25 Dylan Kotliar, Adrian Veres, M Aurel Nagy,
26 Shervin Tabrizi, Eran Hodis, Douglas A
27 Melton, and Pardis C Sabeti. Identify-
28 ing gene expression programs of cell-type
29 identity and cellular activity with single-
30 cell RNA-Seq. *Elife*, 8:e43803, July 2019.
- 31 Alexey M. Kozlov, Diego Darriba, Tomáš
32 Flouri, Benoit Morel, and Alexan-
33 dros Stamatakis. RAxML-NG: a fast,
34 scalable and user-friendly tool for
35 maximum likelihood phylogenetic in-
36 ference. *Bioinformatics*, May 2019.
37 doi: 10.1093/bioinformatics/btz305.
38 URL [https://academic.oup.com/
39 bioinformatics/advance-article/
40 doi/10.1093/bioinformatics/btz305/
41 5487384](https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btz305/5487384).
- 42 O Kozlov. *Models, Optimizations, and
43 Tools for Large-Scale Phylogenetic Infer-
44 ence, Handling Sequence Uncertainty, and
45 Taxonomic Validation*. PhD thesis, Karl-
46 sruhe Institute of Technology (KIT), Octo-
47 ber 2018.
- 48 Sergey Kryazhimskiy and Joshua B Plotkin.
49 The population genetics of dN/dS. *PLoS
50 Genet.*, 4(12):e1000304, December 2008.
- 51 Jack Kuipers, Katharina Jahn, Benjamin J
52 Raphael, and Niko Beerenwinkel. Single-
53 cell sequencing data reveal widespread re-
54 currence and loss of mutational hits in the
55 life histories of tumors. *Genome Res.*, 27
56 (11):1885–1894, November 2017.
- 57 Emma Laks, Hans Zahn, Daniel Lai, Andrew
58 McPherson, Adi Steif, Jazmine Brimhall,
59 Justina Biele, Beixi Wang, Tehmina
60 Masud, Diljot Grewal, Cydney Nielsen,
61 Samantha Leung, Viktoria Bojilova, Maia
62 Smith, Oleg Golovko, Steven Poon, Peter
63 Eirew, Farhia Kabeer, Teresa Ruiz de Al-
64 gara, So Ra Lee, M. Jafar Taghiyar, Curtis
65 Huebner, Jessica Ngo, Tim Chan, Spencer
66 Vatrtr-Watts, Pascale Walters, Nafis Abrar,
67 Sophia Chan, Matt Wiens, Lauren Mar-
68 tin, R. Wilder Scott, Michael T. Under-
69 hill, Elizabeth Chavez, Christian Steidl,
70 Daniel Da Costa, Yusanne Ma, Robin J. N.
71 Coope, Richard Corbett, Stephen Plea-
72 sance, Richard Moore, Andy J. Mungall,
73 Cruk Imaxt Consortium, Marco A. Marra,
74 Carl Hansen, Sohrab Shah, and Samuel
75 Aparicio. Resource: Scalable whole genome
76 sequencing of 40,000 single cells identifies
77 stochastic aneuploidies, genome replication
78 states and clonal repertoires. *bioRxiv*, page
79 411058, September 2018. doi: 10.1101/
80 411058. URL [https://www.biorxiv.org/
81 content/early/2018/09/13/411058](https://www.biorxiv.org/content/early/2018/09/13/411058).
- 82 Ruben T H M Larue, Gilles Defraene,

- 1 Dirk De Ruyscher, Philippe Lambin, and
2 Wouter van Elmpt. Quantitative radiomics
3 studies for tissue characterization: a re-
4 view of technology and methodological pro-
5 cedures. *Br. J. Radiol.*, 90(1070):20160665,
6 February 2017.
- 7 Devon A. Lawson, Kai Kessenbrock,
8 Ryan T. Davis, Nicholas Pervolarakis,
9 and Zena Werb. Tumour heterogeneity
10 and metastasis at single-cell resolu-
11 tion. *Nature Cell Biology*, 20(12):
12 1349, December 2018. ISSN 1476-4679.
13 doi: 10.1038/s41556-018-0236-7. URL
14 [https://www.nature.com/articles/
15 s41556-018-0236-7](https://www.nature.com/articles/s41556-018-0236-7).
- 16 Si Quang Le, Cuong Cao Dang, and Olivier
17 Gascuel. Modeling protein evolution with
18 several amino acid replacement matrices
19 depending on site rates. *Mol. Biol. Evol.*,
20 29(10):2921–2936, October 2012.
- 21 Adam D Leaché, Barbara L Banbury, Joseph
22 Felsenstein, Adrián Nieto-Montes de Oca,
23 and Alexandros Stamatakis. Short tree,
24 long tree, right tree, wrong tree: New ac-
25 quisition bias corrections for inferring SNP
26 phylogenies. *Syst. Biol.*, 64(6):1032–1047,
27 2015.
- 28 Je Hyuk Lee, Evan R Daugharthy, Jonathan
29 Scheiman, Reza Kalhor, Thomas C Fer-
30 rante, Richard Terry, Brian M Turczyk,
31 Joyce L Yang, Ho Suk Lee, John Aach, Kun
32 Zhang, and George M Church. Fluorescent
33 in situ sequencing (FISSEQ) of RNA for
34 gene expression profiling in intact cells and
35 tissues. *Nat. Protoc.*, 10(3):442–458, March
36 2015.
- 37 Jeffrey T. Leek, Robert B. Scharpf, Héc-
38 tor Corrada Bravo, David Simcha, Ben-
39 jamin Langmead, W. Evan Johnson, Don-
40 ald Geman, Keith Baggerly, and Rafael A.
Irizarry. Tackling the widespread and
critical impact of batch effects in high-
throughput data. *Nature Reviews Genetics*,
11(10):733–739, October 2010. ISSN 1471-
0064. doi: 10.1038/nrg2825. URL <https://www.nature.com/articles/nrg2825>.
- A C Leote, X Wu, and A Beyer. Network-
based imputation of dropouts in single-cell
RNA sequencing data. *bioRxiv*, 2019.
- Wei Vivian Li and Jingyi Jessica Li. An accu-
rate and robust imputation method scim-
pute for single-cell RNA-seq data. *Nat.*
Commun., 9(1):997, March 2018.
- Yuval Lieberman, Lior Rokach, and Tal
Shay. CaSTLe – Classification of single
cells by transfer learning: Harnessing
the power of publicly available single cell
RNA sequencing experiments to annotate
new experiments. *PLOS ONE*, 13(10):
e0205499, October 2018. ISSN 1932-6203.
doi: 10.1371/journal.pone.0205499. URL
[https://journals.plos.org/plosone/
article?id=10.1371/journal.pone.
0205499](https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0205499).
- Chieh Lin, Siddhartha Jain, Hannah Kim,
and Ziv Bar-Joseph. Using neural networks
for reducing the dimensions of single-cell
RNA-Seq data. *Nucleic Acids Res.*, 45(17):
e156, September 2017a.
- Peijie Lin, Michael Troup, and Joshua W. K.
Ho. CIDR: Ultrafast and accurate clus-
tering through imputation for single-cell
RNA-seq data. *Genome Biology*, 18(1):59,
March 2017b. ISSN 1474-760X. doi: 10.
1186/s13059-017-1188-0. URL [https://
doi.org/10.1186/s13059-017-1188-0](https://doi.org/10.1186/s13059-017-1188-0).
- G C Linderman, J Zhao, and Y Kluger. Zero-
preserving imputation of scRNA-seq data
using low-rank approximation. *bioRxiv*,

- 1 2018. URL [https://www.biorxiv.org/](https://www.biorxiv.org/content/10.1101/397588v1.abstract)
2 [content/10.1101/397588v1.abstract](https://www.biorxiv.org/content/10.1101/397588v1.abstract).
- 3 Liang Liu, Zhenxiang Xi, Shaoyuan Wu,
4 Charles C Davis, and Scott V Edwards. Es-
5 timating phylogenetic trees from genome-
6 scale data. *Ann. N. Y. Acad. Sci.*, 1360:
7 36–53, December 2015.
- 8 Jackson Loper, Trygve Bakken, Uygur
9 Sumbul, Gabe Murphy, Hongkui Zeng,
10 David Blei, and Liam Paninski. The
11 Markov link method: a nonparametric
12 approach to combine observations from
13 multiple experiments. *bioRxiv*, page
14 457283, January 2019. doi: 10.1101/
15 457283. URL [https://www.biorxiv.org/](https://www.biorxiv.org/content/10.1101/457283v3)
16 [content/10.1101/457283v3](https://www.biorxiv.org/content/10.1101/457283v3).
- 17 Romain Lopez, Jeffrey Regier, Michael B
18 Cole, Michael I Jordan, and Nir Yosef.
19 Deep generative modeling for single-cell
20 transcriptomics. *Nature methods*, 15
21 (12):1053–1058, December 2018. ISSN
22 1548-7091, 1548-7105. doi: 10.1038/
23 s41592-018-0229-2. URL [http://dx.doi.](http://dx.doi.org/10.1038/s41592-018-0229-2)
24 [org/10.1038/s41592-018-0229-2](http://dx.doi.org/10.1038/s41592-018-0229-2).
- 25 Eric Lubeck, Ahmet F Coskun, Timur
26 Zhiyentayev, Mubhij Ahmad, and Long
27 Cai. Single-cell in situ RNA profiling by
28 sequential hybridization. *Nat. Methods*, 11
29 (4):360–361, April 2014.
- 30 Aaron T L Lun and John C Marioni. Over-
31 coming confounding plate effects in dif-
32 ferential expression analyses of single-cell
33 RNA-seq data. *Biostatistics*, 18(3):451–
34 464, July 2017.
- 35 Aaron T L Lun, Karsten Bach, and John C
36 Marioni. Pooling across cells to normalize
37 single-cell RNA sequencing data with many
38 zero counts. *Genome Biol.*, 17:75, April
39 2016.
- 40 Aaron T L Lun, Arianne C Richard, and
41 John C Marioni. Testing for differential
42 abundance in mass cytometry data. *Nat.*
43 *Methods*, 14(7):707–709, July 2017.
- 44 Tao Luo, Lei Fan, Rong Zhu, and Dong Sun.
45 Microfluidic Single-Cell manipulation and
46 analysis: Methods and applications. *Micro-*
47 *machines (Basel)*, 10(2), February 2019.
- 48 Iain C. Macaulay, Mabel J. Teng, Wilfried
49 Haerty, Parveen Kumar, Chris P. Ponting,
50 and Thierry Voet. Separation and paral-
51 lel sequencing of the genomes and tran-
52 scriptomes of single cells using G&T-
53 seq. *Nature Protocols*, 11(11):2081–2103,
54 November 2016. ISSN 1750-2799. doi: 10.
55 1038/nprot.2016.138. URL [https://www.](https://www.nature.com/articles/nprot.2016.138)
56 [nature.com/articles/nprot.2016.138](https://www.nature.com/articles/nprot.2016.138).
- 57 Iain C. Macaulay, Chris P. Ponting, and
58 Thierry Voet. Single-Cell Multiomics:
59 Multiple Measurements from Single
60 Cells. *Trends in Genetics*, 33(2):155–
61 168, February 2017. ISSN 0168-9525.
62 doi: 10.1016/j.tig.2016.12.003. URL
63 [https://www.ncbi.nlm.nih.gov/pmc/](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5303816/)
64 [articles/PMC5303816/](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5303816/).
- 65 Evan Z. Macosko, Anindita Basu, Rahul
66 Satija, James Nemesh, Karthik Shekhar,
67 Melissa Goldman, Itay Tirosh, Allison R.
68 Bialas, Nolan Kamitaki, Emily M. Marter-
69 steck, John J. Trombetta, David A. Weitz,
70 Joshua R. Sanes, Alex K. Shalek, Aviv
71 Regev, and Steven A. McCarroll. Highly
72 Parallel Genome-wide Expression Profil-
73 ing of Individual Cells Using Nanoliter
74 Droplets. *Cell*, 161(5):1202–1214, May
75 2015. ISSN 1097-4172. doi: 10.1016/j.cell.
76 2015.05.002.
- 77 Serghei Mangul, Lana S. Martin, Brian L.
78 Hill, Angela Ka-Mei Lam, Margaret G.
79 Distler, Alex Zelikovsky, Eleazar Es-
80 kin, and Jonathan Flint. Systematic

- 1 benchmarking of omics computational
2 tools. *Nature Communications*, 10(1):
3 1393, March 2019. ISSN 2041-1723.
4 doi: 10.1038/s41467-019-09406-4. URL
5 [https://www.nature.com/articles/
6 s41467-019-09406-4](https://www.nature.com/articles/s41467-019-09406-4).
- 7 Gioele La Manno, Ruslan Soldatov, Amit
8 Zeisel, Emelie Braun, Hannah Hochgerner,
9 Viktor Petukhov, Katja Lidschreiber,
10 Maria E. Kastriiti, Peter Lönnerberg,
11 Alessandro Furlan, Jean Fan, Lars E.
12 Borm, Zehua Liu, David van Bruggen,
13 Jimin Guo, Xiaoling He, Roger Barker,
14 Erik Sundström, Gonçalo Castelo-Branco,
15 Patrick Cramer, Igor Adameyko, Sten
16 Linnarsson, and Peter V. Kharchenko.
17 RNA velocity of single cells. *Nature*, 560
18 (7719):494, August 2018. ISSN 1476-4687.
19 doi: 10.1038/s41586-018-0414-6. URL
20 [https://www.nature.com/articles/
21 s41586-018-0414-6](https://www.nature.com/articles/s41586-018-0414-6).
- 22 Erik A Martens, Rumen Kostadinov, Carlo C
23 Maley, and Oskar Hallatschek. Spatial
24 structure increases the waiting time for can-
25 cer. *New J. Phys.*, 13, November 2011.
- 26 Dariusz Matlak and Ewa Szczurek. Epista-
27 sis in genomic and survival data of can-
28 cer patients. *PLoS Comput. Biol.*, 13(7):
29 e1005626, July 2017.
- 30 Davis James McCarthy, Raghd Rostom,
31 Yuanhua Huang, Daniel J. Kunz, Petr
32 Danecek, Marc Jan Bonder, Tzachi Hagai,
33 HipSci Consortium, Wenyi Wang, Daniel J.
34 Gaffney, Benjamin D. Simons, Oliver Ste-
35 gle, and Sarah A. Teichmann. Cardelino:
36 Integrating whole exomes and single-cell
37 transcriptomes to reveal phenotypic im-
38 pact of somatic variants. *bioRxiv*, page
39 413047, November 2018. doi: 10.1101/
40 413047. URL [https://www.biorxiv.org/
41 content/10.1101/413047v2](https://www.biorxiv.org/content/10.1101/413047v2).
- Nicholas McGranahan and Charles Swanton. 42
Clonal heterogeneity and tumor evolution: 43
Past, present, and the future. *Cell*, 168(4): 44
613–628, February 2017. 45
- Chiara Medaglia, Amir Giladi, Liat Stoler- 46
Barak, Marco De Giovanni, Tomer Meir 47
Salame, Adi Biram, Eyal David, Han- 48
jie Li, Matteo Iannacone, Ziv Shul- 49
man, and Ido Amit. Spatial recon- 50
struction of immune niches by combin- 51
ing photoactivatable reporters and scRNA- 52
seq. *Science*, 358(6370):1622–1626, De- 53
cember 2017. ISSN 0036-8075, 1095- 54
9203. doi: 10.1126/science.aao4277. 55
URL [http://science.sciencemag.org/
56 content/358/6370/1622](http://science.sciencemag.org/content/358/6370/1622). 57
- Jing Meng and Yi-Ping Phoebe Chen. A 58
database of simulated tumor genomes to- 59
wards accurate detection of somatic small 60
variants in cancer. *PLoS One*, 13(8): 61
e0202982, August 2018. 62
- Christopher R. Merritt, Giang T. Ong, 63
Sarah Church, Kristi Barker, Gary Geiss, 64
Margaret Hoang, Jaemyeong Jung, Yan 65
Liang, Jill McKay-Fleisch, Karen Nguyen, 66
Kristina Sorg, Isaac Sprague, Charles War- 67
ren, Sarah Warren, Zoey Zhou, Daniel R. 68
Zollinger, Dwayne L. Dunaway, Gordon B. 69
Mills, and Joseph M. Beechem. High 70
multiplex, digital spatial profiling of pro- 71
teins and RNA in fixed tissue using ge- 72
nomic detection methods. *bioRxiv*, page 73
559021, February 2019. doi: 10.1101/
74 559021. URL [https://www.biorxiv.org/
75 content/10.1101/559021v2](https://www.biorxiv.org/content/10.1101/559021v2). 76
- Zhun Miao, Jiaqi Li, and Xuegong Zhang. 77
screcover: Discriminating true and false ze- 78
ros in single-cell RNA-seq data for imputa- 79
tion. June 2019. 80
- Franziska Michor, Yoh Iwasa, and Martin A 81
Nowak. Dynamics of cancer progression. 82

- 1 *Nat. Rev. Cancer*, 4(3):197–205, March
2 2004.
- 3 Jeffrey R Moffitt, Junjie Hao, Dhanan-
4 jay Bambah-Mukku, Tian Lu, Cather-
5 ine Dulac, and Xiaowei Zhuang. High-
6 performance multiplexed fluorescence in
7 situ hybridization in culture and tissue with
8 matrix imprinting and clearing. *Proc. Natl.*
9 *Acad. Sci. U. S. A.*, 113(50):14456–14461,
10 December 2016.
- 11 Jeffrey R. Moffitt, Dhananjay Bambah-
12 Mukku, Stephen W. Eichhorn, Eric
13 Vaughn, Karthik Shekhar, Julio D.
14 Perez, Nimrod D. Rubinstein, Junjie
15 Hao, Aviv Regev, Catherine Dulac,
16 and Xiaowei Zhuang. Molecular, spa-
17 tial, and functional single-cell profiling
18 of the hypothalamic preoptic region.
19 *Science*, 362(6416):eaau5324, Novem-
20 ber 2018. ISSN 0036-8075, 1095-9203.
21 doi: 10.1126/science.aau5324. URL
22 [http://science.sciencemag.org/](http://science.sciencemag.org/content/362/6416/eaau5324)
23 [content/362/6416/eaau5324](http://science.sciencemag.org/content/362/6416/eaau5324).
- 24 Kevin R Moon, Jay S Stanley, Daniel
25 Burkhardt, David van Dijk, Guy Wolf, and
26 Smita Krishnaswamy. Manifold learning-
27 based methods for analyzing single-cell
28 RNA-sequencing data. *Current Opinion in*
29 *Systems Biology*, 7:36–46, 2018.
- 30 Marmar Moussa and Ion I. Măndoiu.
31 Locality Sensitive Imputation for Sin-
32 gle Cell RNA-Seq Data. *Journal of*
33 *Computational Biology*, February 2019.
34 doi: 10.1089/cmb.2018.0236. URL
35 [https://www.liebertpub.com/doi/10.](https://www.liebertpub.com/doi/10.1089/cmb.2018.0236)
36 [1089/cmb.2018.0236](https://www.liebertpub.com/doi/10.1089/cmb.2018.0236).
- 37 Nature Methods, 2013. Method of the year
38 2013. *Nature Methods*, 11:1 EP –, 12
39 2013. URL [https://doi.org/10.1038/](https://doi.org/10.1038/nmeth.2801)
40 [nmeth.2801](https://doi.org/10.1038/nmeth.2801).
- Nicholas Navin, Jude Kendall, Jennifer Troge,
Peter Andrews, Linda Rodgers, Jeanne
McIndoo, Kerry Cook, Asya Stepan-
sky, Dan Levy, Diane Esposito, Lakshmi
Muthuswamy, Alex Krasnitz, W Richard
McCombie, James Hicks, and Michael
Wigler. Tumour evolution inferred by
single-cell sequencing. *Nature*, 472(7341):
90–94, April 2011.
- Richard A Neher, Colin A Russell, and Boris I
Shraiman. Predicting evolution from the
shape of genealogical trees. *Elife*, 3, Novem-
ber 2014.
- Malgorzata Nowicka, Carsten Krieg, Lukas M
Weber, Felix J Hartmann, Silvia Guglietta,
Burkhard Becher, Mitchell P Levesque,
and Mark D Robinson. CyTOF work-
flow: differential discovery in high-
throughput high-dimensional cytometry
datasets. *F1000Res.*, 6:748, May 2017.
- Huw A Ogilvie, Remco R Bouckaert, and
Alexei J Drummond. StarBEAST2 brings
faster species tree inference and accurate
estimates of substitution rates. *Mol. Biol.*
Evol., 34(8):2101–2114, August 2017.
- J Guillermo Paez, Ming Lin, Rameen
Beroukhim, Jeffrey C Lee, Xiaojun Zhao,
Daniel J Richter, Stacey Gabriel, Paula
Herman, Hidefumi Sasaki, David Alt-
shuler, Cheng Li, Matthew Meyerson, and
William R Sellers. Genome coverage and
sequence fidelity of phi29 polymerase-based
multiple strand displacement whole genome
amplification. *Nucleic Acids Res.*, 32(9):
e71, May 2004.
- Jong-Eun Park, Krzysztof Polanski, Kerstin
Meyer, and Sarah A. Teichmann. Fast
Batch Alignment of Single Cell Transcrip-
tomes Unifies Multiple Mouse Cell Atlases
into an Integrated Landscape. *bioRxiv*,

- 1 page 397042, August 2018. doi: 10.1101/
2 397042. URL [https://www.biorxiv.org/
3 content/10.1101/397042v2](https://www.biorxiv.org/content/10.1101/397042v2).
- 4 Anoop P Patel, Itay Tirosh, John J Trom-
5 betta, Alex K Shalek, Shawn M Gille-
6 spie, Hiroaki Wakimoto, Daniel P Cahill,
7 Brian V Nahed, William T Curry, Robert L
8 Martuza, David N Louis, Orit Rozenblatt-
9 Rosen, Mario L Suvà, Aviv Regev, and
10 Bradley E Bernstein. Single-cell RNA-
11 seq highlights intratumoral heterogeneity
12 in primary glioblastoma. *Science*, 344
13 (6190):1396–1401, June 2014.
- 14 Tao Peng, Qin Zhu, Penghang Yin, and
15 Kai Tan. SCRABBLE: single-cell RNA-
16 seq imputation constrained by bulk RNA-
17 seq data. *Genome biology*, 20(1):88, May
18 2019. ISSN 1465-6906. doi: 10.1186/
19 s13059-019-1681-8. URL [http://dx.doi.
20 org/10.1186/s13059-019-1681-8](http://dx.doi.org/10.1186/s13059-019-1681-8).
- 21 N. Pezzotti, T. Höllt, B. Lelieveldt, E. Eise-
22 mann, and A. Vilanova. Hierarchical
23 Stochastic Neighbor Embedding. *Com-
24 puter Graphics Forum*, 35(3):21–30, 2016.
25 ISSN 1467-8659. doi: 10.1111/cgf.
26 12878. URL [https://onlinelibrary.
27 wiley.com/doi/abs/10.1111/cgf.12878](https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.12878).
- 28 Ángel J Picher, Bettina Budeus, Oliver
29 Wafzig, Carola Krüger, Sara García-
30 Gómez, María I Martínez-Jiménez, Al-
31 berto Díaz-Talavera, Daniela Weber, Luis
32 Blanco, and Armin Schneider. TruePrime
33 is a novel method for whole-genome am-
34 plification from single cells based on Tth-
35 PrimPol. *Nat. Commun.*, 7:13296, Novem-
36 ber 2016.
- 37 Emma Pierson and Christopher Yau.
38 ZIFA: Dimensionality reduction for
39 zero-inflated single-cell gene expres-
40 sion analysis. *Genome Biology*, 16
(1):241, November 2015. ISSN 1474-
760X. doi: 10.1186/s13059-015-0805-z.
URL [https://doi.org/10.1186/
s13059-015-0805-z](https://doi.org/10.1186/s13059-015-0805-z).
- 41 Mireya Plass, Jordi Solana, F. Alexan-
42 der Wolf, Salah Ayoub, Aristotelis Mi-
43 sios, Petar Glazar, Benedikt Obermayer,
44 Fabian J. Theis, Christine Kocks, and Niko-
45 laus Rajewsky. Cell type atlas and lineage
46 tree of a whole complex animal by single-
47 cell transcriptomics. *Science*, 360(6391):
48 eaaq1723, May 2018. ISSN 0036-8075,
49 1095-9203. doi: 10.1126/science.aaq1723.
50 URL [http://science.sciencemag.org/
51 content/360/6391/eaaq1723](http://science.sciencemag.org/content/360/6391/eaaq1723).
- 52 Hannah A. Pliner, Jonathan S. Packer,
53 José L. McFaline-Figueroa, Darren A.
54 Cusanovich, Riza M. Daza, Delasa
55 Aghamirzaie, Sanjay Srivatsan, Xiaojie
56 Qiu, Dana Jackson, Anna Minkina, An-
57 drew C. Adey, Frank J. Steemers, Jay
58 Shendure, and Cole Trapnell. Cicero
59 Predicts cis-Regulatory DNA Interactions
60 from Single-Cell Chromatin Accessi-
61 bility Data. *Molecular Cell*, 71(5):
62 858–871.e8, 2018. ISSN 1097-4164. doi:
63 10.1016/j.molcel.2018.06.044.
- 64 Olivier Poirion, Xun Zhu, Travers Ching, and
65 Lana X. Garmire. Using single nucleotide
66 variations in single-cell RNA-seq to identify
67 subpopulations and genotype-phenotype
68 linkage. *Nature Communications*, 9(1):
69 4892, November 2018. ISSN 2041-1723.
70 doi: 10.1038/s41467-018-07170-5. URL
71 [https://www.nature.com/articles/
72 s41467-018-07170-5](https://www.nature.com/articles/s41467-018-07170-5).
- 73 David D Pollock, Derrick J Zwickl, Jimmy A
74 McGuire, and David M Hillis. Increased
75 taxon sampling is advantageous for phylo-
76 genetic inference. *Syst. Biol.*, 51(4):664–
77 671, August 2002.

- 1 Vladimir Potapov and Jennifer L Ong. Ex-
2 amining sources of error in PCR by Single-
3 Molecule sequencing. *PLoS One*, 12(1):
4 e0169774, January 2017.
- 5 Xiaojie Qiu, Qi Mao, Ying Tang, Li Wang,
6 Raghav Chawla, Hannah A. Pliner, and
7 Cole Trapnell. Reversed graph embed-
8 ding resolves complex single-cell trajecto-
9 ries. *Nature Methods*, 14(10):979–982, Oc-
10 tober 2017. ISSN 1548-7105. doi: 10.1038/
11 nmeth.4402. URL [https://www.nature.](https://www.nature.com/articles/nmeth.4402)
12 [com/articles/nmeth.4402](https://www.nature.com/articles/nmeth.4402).
- 13 Bruce Rannala and Ziheng Yang. Efficient
14 bayesian species tree inference under the
15 multispecies coalescent. *Syst. Biol.*, 66(5):
16 823–842, September 2017.
- 17 Benjamin Redelings. Erasing errors due to
18 alignment ambiguity when estimating posi-
19 tive selection. *Mol. Biol. Evol.*, 31(8):1979–
20 1993, August 2014.
- 21 Aviv Regev, Sarah A. Teichmann, Eric S.
22 Lander, Ido Amit, Christophe Benoist,
23 Ewan Birney, Bernd Bodenmiller, Peter
24 Campbell, Piero Carninci, Menna Clat-
25 worthy, Hans Clevers, Bart Deplancke,
26 Ian Dunham, James Eberwine, Roland
27 Eils, Wolfgang Enard, Andrew Farmer,
28 Lars Fugger, Berthold Göttgens, Nir Ha-
29 cohen, Muzlifah Haniffa, Martin Hem-
30 berg, Seung Kim, Paul Klennerman, Arnold
31 Kriegstein, Ed Lein, Sten Linnarsson,
32 Joakim Lundeberg, Partha Majumder,
33 John C. Marioni, Miriam Merad, Musa
34 Mhlanga, Martijn Nawijn, Mihai Netea,
35 Garry Nolan, Dana Pe’er, Anthony Philli-
36 pakis, Chris P. Ponting, Steve Quake,
37 Wolf Reik, Orit Rozenblatt-Rosen, Joshua
38 Sanes, Rahul Satija, Ton N. Schumacher,
39 Alex Shalek, Ehud Shapiro, Padmanee
40 Sharma, Jay W. Shin, Oliver Stegle,
Michael Stratton, Michael J. T. Stubbington,
Alexander van Oudenaarden, Allon
Wagner, Fiona Watt, Jonathan Weissman,
Barbara Wold, Ramnik Xavier, Nir Yosef,
and the Human Cell Atlas Meeting Partic-
ipants. The Human Cell Atlas. *bioRxiv*,
page 121202, May 2017. doi: 10.1101/
121202. URL [https://www.biorxiv.org/](https://www.biorxiv.org/content/10.1101/121202v1)
[content/10.1101/121202v1](https://www.biorxiv.org/content/10.1101/121202v1).
- John E. Reid and Lorenz Wernisch. Pseu-
dotime estimation: deconfounding single
cell time series. *Bioinformatics*, 32(19):
2973–2980, October 2016. ISSN 1367-
4803. doi: 10.1093/bioinformatics/btw372.
URL [https://academic.oup.com/](https://academic.oup.com/bioinformatics/article/32/19/2973/2196633)
[bioinformatics/article/32/19/2973/](https://academic.oup.com/bioinformatics/article/32/19/2973/2196633)
[2196633](https://academic.oup.com/bioinformatics/article/32/19/2973/2196633).
- Stephen Reid, Jonathan Taylor, and Robert
Tibshirani. A general framework for esti-
mation and inference from clusters of fea-
tures. *J. Am. Stat. Assoc.*, 113(521):280–
293, January 2018.
- Davide Risso, Fanny Perraudeau, Svet-
lana Gribkova, Sandrine Dudoit, and
Jean-Philippe Vert. A general and
flexible method for signal extraction
from single-cell RNA-seq data. *Nature*
Communications, 9(1):284, Jan-
uary 2018. ISSN 2041-1723. doi:
10.1038/s41467-017-02554-5. URL
[https://www.nature.com/articles/](https://www.nature.com/articles/s41467-017-02554-5)
[s41467-017-02554-5](https://www.nature.com/articles/s41467-017-02554-5).
- Elena Rivas and Sean R Eddy. Probabilis-
tic phylogenetic inference with insertions
and deletions. *PLoS Comput. Biol.*, 4(9):
e1000172, September 2008.
- Abbas H. Rizvi, Pablo G. Camara, Elena K.
Kandror, Thomas J. Roberts, Ira Schieren,
Tom Maniatis, and Raul Rabadan. Single-
cell topological RNA-seq analysis reveals

- 1 insights into cellular differentiation and de-
2 velopment. *Nature Biotechnology*, 35(6):
3 551–560, 2017. ISSN 1546-1696. doi: 10.
4 1038/nbt.3854.
- 5 Simone Rizzetto, Auda A Eltahla, Peijie Lin,
6 Rowena Bull, Andrew R Lloyd, Joshua
7 W K Ho, Vanessa Venturi, and Fabio Lu-
8 ciani. Impact of sequencing depth and read
9 length on single cell RNA sequencing data
10 of T cells. *Sci. Rep.*, 7(1):12781, October
11 2017.
- 12 S Roch. A short proof that phylogenetic tree
13 reconstruction by maximum likelihood is
14 hard. *IEEE/ACM Trans. Comput. Biol.*
15 *Bioinform.*, 3(1):92–94, 2006.
- 16 Samuel G. Rodriques, Robert R. Stick-
17 els, Aleksandrina Goeva, Carly A. Mar-
18 tin, Evan Murray, Charles R. Vander-
19 burg, Joshua Welch, Linlin M. Chen, Fei
20 Chen, and Evan Z. Macosko. Slide-
21 seq: A scalable technology for measur-
22 ing genome-wide expression at high spa-
23 tial resolution. *Science*, 363(6434):1463–
24 1467, March 2019. ISSN 0036-8075,
25 1095-9203. doi: 10.1126/science.aaw1219.
26 URL [https://science.sciencemag.org/](https://science.sciencemag.org/content/363/6434/1463)
27 [content/363/6434/1463](https://science.sciencemag.org/content/363/6434/1463).
- 28 Florian Rohart, Aida Eslami, Nicholas Mati-
29 gian, Stéphanie Bougeard, and Kim-
30 Anh Lê Cao. MINT: a multivari-
31 ate integrative method to identify re-
32 producible molecular signatures across
33 independent experiments and platforms.
34 *BMC Bioinformatics*, 18(1):128, February
35 2017a. ISSN 1471-2105. doi: 10.1186/
36 s12859-017-1553-8. URL [https://doi.](https://doi.org/10.1186/s12859-017-1553-8)
37 [org/10.1186/s12859-017-1553-8](https://doi.org/10.1186/s12859-017-1553-8).
- 38 Florian Rohart, Benoît Gautier, Amrit
39 Singh, and Kim-Anh Lê Cao. mixOmics:
40 An R package for ‘omics feature se-
41 lection and multiple data integration.
PLOS Computational Biology, 13(11):
e1005752, November 2017b. ISSN 1553-
7358. doi: 10.1371/journal.pcbi.1005752.
URL [https://journals.plos.org/](https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005752)
[ploscompbiol/article?id=10.1371/](https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005752)
[journal.pcbi.1005752](https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005752).
- Alexander B. Rosenberg, Charles M. Roco,
Richard A. Muscat, Anna Kuchina, Paul
Sample, Zizhen Yao, Lucas T. Graybuck,
David J. Peeler, Sumit Mukherjee, Wei
Chen, Suzie H. Pun, Drew L. Sellers,
Bosiljka Tasic, and Georg Seelig. Single-
cell profiling of the developing mouse brain
and spinal cord with split-pool barcoding.
Science, 360(6385):176–182, April 2018.
ISSN 0036-8075, 1095-9203. doi: 10.1126/
science.aam8999. URL [http://science.](http://science.sciencemag.org/content/360/6385/176)
[sciencemag.org/content/360/6385/176](http://science.sciencemag.org/content/360/6385/176).
- Edith M Ross and Florian Markowetz. On-
coNEM: inferring tumor evolution from
single-cell sequencing data. *Genome Biol.*,
17:69, April 2016.
- Andrew Roth, Andrew McPherson, Emma
Laks, Justina Biele, Damian Yap, Adrian
Wan, Maia A Smith, Cydney B Nielsen,
Jessica N McAlpine, Samuel Aparicio,
Alexandre Bouchard-Côté, and Sohrab P
Shah. Clonal genotype and population
structure inference from single-cell tumor
sequencing. *Nat. Methods*, 13(7):573–576,
July 2016.
- Adela Saco, Jose Ramírez, Natalia Rakislova,
Aurea Mira, and Jaume Ordi. Validation of
Whole-Slide imaging for histopathological
diagnosis: Current state. *Pathobiology*, 83
(2-3):89–98, April 2016.
- Wouter Saelens, Robrecht Cannoodt,
Helena Todorov, and Yvan Saeys. A
comparison of single-cell trajectory in-
ference methods. *Nature Biotechnology*,

- 1 page 1, April 2019. ISSN 1546-1696. 41
2 doi: 10.1038/s41587-019-0071-9. URL 42
3 [https://www.nature.com/articles/](https://www.nature.com/articles/s41587-019-0071-9)
4 [s41587-019-0071-9](https://www.nature.com/articles/s41587-019-0071-9).
- 5 Yvan Saeys, Sofie Van Gassen, and Bart N. 43
6 Lambrecht. Computational flow cytom- 44
7 etry: helping to make sense of high- 45
8 dimensional immunology data. *Nature* 46
9 *Reviews Immunology*, 16(7):449–462, July 47
10 2016. ISSN 1474-1741. doi: 10.1038/nri. 48
11 2016.56. URL [https://www.nature.com/](https://www.nature.com/articles/nri.2016.56) 49
12 [articles/nri.2016.56](https://www.nature.com/articles/nri.2016.56).
- 13 Stefano Santaguida, Amelia Richardson, 50
14 Divya Ramalingam Iyer, Ons M'Saad, 51
15 Lauren Zasadil, Kristin A. Knouse, 52
16 Yao Liang Wong, Nicholas Rhind, Arshad 53
17 Desai, and Angelika Amon. Chromosome 54
18 Mis-segregation Generates Cell-Cycle- 55
19 Arrested Cells with Complex Karyotypes 56
20 that Are Eliminated by the Immune 57
21 System. *Developmental Cell*, 41(6): 58
22 638–651.e5, June 2017. ISSN 15345807. 59
23 doi: 10.1016/j.devcel.2017.05.022. URL 60
24 [https://linkinghub.elsevier.com/](https://linkinghub.elsevier.com/retrieve/pii/S1534580717304306) 61
25 [retrieve/pii/S1534580717304306](https://linkinghub.elsevier.com/retrieve/pii/S1534580717304306).
- 26 Gryte Satas and Benjamin J Raphael. Haplo- 62
27 type phasing in single-cell DNA-sequencing 63
28 data. *Bioinformatics*, 34(13):i211–i217, 64
29 July 2018.
- 30 Rahul Satija, Jeffrey A Farrell, David Gen- 65
31 nert, Alexander F Schier, and Aviv Regev. 66
32 Spatial reconstruction of single-cell gene 67
33 expression data. *Nat. Biotechnol.*, 33(5):495– 68
34 502, May 2015.
- 35 Kenta Sato, Koki Tsuyuzaki, Kentaro 69
36 Shimizu, and Itoshi Nikaido. CellFish- 70
37 ing.jl: an ultrafast and scalable cell 71
38 search method for single-cell RNA sequenc- 72
39 ing. *Genome Biology*, 20(1):31, February 73
40 2019. ISSN 1474-760X. doi: 10.1186/ 74
s13059-019-1639-x. URL [https://doi.](https://doi.org/10.1186/s13059-019-1639-x) 75
41 [org/10.1186/s13059-019-1639-x](https://doi.org/10.1186/s13059-019-1639-x). 76
42
- 43 Arpiar Saunders, Evan Z. Macosko, Alec 77
44 Wysoker, Melissa Goldman, Fenna M. 78
45 Krienen, Heather de Rivera, Elizabeth 79
46 Bien, Matthew Baum, Laura Bortolin, 80
47 Shuyu Wang, Aleksandrina Goeva, James 81
48 Nemesh, Nolan Kamitaki, Sara Brum-
49 baugh, David Kulp, and Steven A.
50 McCarroll. Molecular Diversity and Spe-
51 cializations among the Cells of the Adult
52 Mouse Brain. *Cell*, 174(4):1015–1030.e16,
53 August 2018. ISSN 0092-8674. doi:
54 10.1016/j.cell.2018.07.028. URL [http:](http://www.sciencedirect.com/science/article/pii/S0092867418309553)
55 [//www.sciencedirect.com/science/](http://www.sciencedirect.com/science/article/pii/S0092867418309553)
56 [article/pii/S0092867418309553](http://www.sciencedirect.com/science/article/pii/S0092867418309553).
- 57 Denis Schapiro, Hartland W Jackson, Swetha
58 Raghuraman, Jana R Fischer, Vito R T
59 Zanotelli, Daniel Schulz, Charlotte Giesen,
60 Raúl Catena, Zsuzsanna Varga, and Bernd
61 Bodenmiller. histoCAT: analysis of cell
62 phenotypes and interactions in multiplex
63 image cytometry data. *Nat. Methods*, 14
64 (9):873–876, September 2017.
- 65 Geoffrey Schiebinger, Jian Shu, Marcin
66 Tabaka, Brian Cleary, Vidya Subrama-
67 nian, Aryeh Solomon, Siyan Liu, Sta-
68 cie Lin, Peter Berube, Lia Lee, Jenny
69 Chen, Justin Brumbaugh, Philippe Rigol-
70 let, Konrad Hochedlinger, Rudolf Jaenisch,
71 Aviv Regev, and Eric S. Lander. Re-
72 construction of developmental landscapes
73 by optimal-transport analysis of single-
74 cell gene expression sheds light on cel-
75 lular reprogramming. *bioRxiv*, page
76 191056, September 2017. doi: 10.1101/
77 191056. URL [https://www.biorxiv.org/](https://www.biorxiv.org/content/10.1101/191056v1)
78 [content/10.1101/191056v1](https://www.biorxiv.org/content/10.1101/191056v1).
- 79 Herbert B Schiller, Daniel T Montoro,
80 Lukas M Simon, Emma L Rawlins,
81 Kerstin B Meyer, Maximilian Strunz,

- 1 Felipe Vieira Braga, Wim Timens, Ger-
2 ard H Koppelman, G.R. Scott Budinger,
3 Janette K Burgess, Avinash Waghray,
4 Maarten van den Berge, Fabian J Theis,
5 Aviv Regev, Naftali Kaminski, Jayaraj
6 Rajagopal, Sarah A Teichmann, Alexan-
7 der V Misharin, and Martijn C Nawijn.
8 The Human Lung Cell Atlas - A high-
9 resolution reference map of the human
10 lung in health and disease. *American*
11 *Journal of Respiratory Cell and Molecular*
12 *Biology*, April 2019. ISSN 1044-1549.
13 doi: 10.1165/rcmb.2018-0416TR. URL
14 [https://www.atsjournals.org/doi/](https://www.atsjournals.org/doi/abs/10.1165/rcmb.2018-0416TR)
15 [abs/10.1165/rcmb.2018-0416TR](https://www.atsjournals.org/doi/abs/10.1165/rcmb.2018-0416TR).

16 Roland F. Schwarz, Anne Trinh, Botond
17 Sipos, James D. Brenton, Nick Gold-
18 man, and Florian Markowetz. Phyloge-
19 netic Quantification of Intra-tumour Het-
20 erogeneity. *PLoS Computational Biol-*
21 *ogy*, 10(4):e1003535, April 2014. ISSN
22 1553-7358. doi: 10.1371/journal.pcbi.
23 1003535. URL [https://dx.plos.org/10.](https://dx.plos.org/10.1371/journal.pcbi.1003535)
24 [1371/journal.pcbi.1003535](https://dx.plos.org/10.1371/journal.pcbi.1003535).

25 Roberto Semeraro, Valerio Orlandini, and Al-
26 berto Magi. Xome-Blender: A novel can-
27 cer genome simulator. *PLoS One*, 13(4):
28 e0194472, April 2018.

29 Debarka Sengupta, Nirmala Arul Rayan,
30 Michelle Lim, Bing Lim, and Shyam
31 Prabhakar. Fast, scalable and accu-
32 rate differential expression analysis for
33 single cells. *bioRxiv*, page 049734,
34 April 2016. doi: 10.1101/049734. URL
35 [https://www.biorxiv.org/content/10.](https://www.biorxiv.org/content/10.1101/049734v1)
36 [1101/049734v1](https://www.biorxiv.org/content/10.1101/049734v1).

37 Manu Setty, Michelle D. Tadmor, Shlomit
38 Reich-Zeliger, Omer Angel, Tomer Meir
39 Salame, Pooja Kathail, Kristy Choi, Sean
40 Bendall, Nir Friedman, and Dana Pe'er.

Wishbone identifies bifurcating develop-
mental trajectories from single-cell data.
Nature Biotechnology, 34(6):637–645, June
2016. ISSN 1546-1696. doi: 10.1038/nbt.
3569. URL [https://www.nature.com/](https://www.nature.com/articles/nbt.3569)
[articles/nbt.3569](https://www.nature.com/articles/nbt.3569).

D T Severson, R P Owen, M J White, X Lu,
and B Schuster-Böckler. BEARscc deter-
mines robustness of single-cell clusters us-
ing simulated technical replicates. *Nat.*
Commun., 9(1):1187, March 2018.

Sheel Shah, Eric Lubeck, Wen Zhou, and
Long Cai. In situ transcription profiling of
single cells reveals spatial organization of
cells in the mouse hippocampus. *Neuron*,
92(2):342–357, October 2016.

Arun Shivanandan, Jayakrishnan Unnikrish-
nan, and Aleksandra Radenovic. On char-
acterizing protein spatial clusters with cor-
relation approaches. *Sci. Rep.*, 6:31164,
August 2016.

Angus M Sidore, Freeman Lan, Shaun W
Lim, and Adam R Abate. Enhanced se-
quencing coverage with digital droplet mul-
tiple displacement amplification. *Nucleic*
Acids Res., 44(7):e66, April 2016.

Jochen Singer, Jack Kuipers, Katharina
Jahn, and Niko Beerenwinkel. Single-cell
mutation identification via phylogenetic
inference. *Nature Communications*, 9(1):
5144, December 2018. ISSN 2041-1723.
doi: 10.1038/s41467-018-07627-7. URL
[https://www.nature.com/articles/](https://www.nature.com/articles/s41467-018-07627-7)
[s41467-018-07627-7](https://www.nature.com/articles/s41467-018-07627-7).

Amrit Singh, Benoit Gautier, Casey P.
Shannon, Florian Rohart, Michael Vacher,
Scott J. Tebutt, and Kim-Anh Le Cao.
DIABLO: from multi-omics assays to
biomarker discovery, an integrative ap-
proach. *bioRxiv*, page 067611, March

- 1 2018. doi: 10.1101/067611. URL
2 [https://www.biorxiv.org/content/10.](https://www.biorxiv.org/content/10.1101/067611v2)
3 [1101/067611v2](https://www.biorxiv.org/content/10.1101/067611v2).
- 4 Debajyoti Sinha, Akhilesh Kumar, Himan-
5 shu Kumar, Sanghamitra Bandyopadhyay,
6 and Debarka Sengupta. dropclust: efficient
7 clustering of ultra-large scRNA-seq data.
8 *Nucleic Acids Res.*, 46(6):e36, April 2018.
- 9 Pavel Skums, Viachaslau Tsyvina, and Alex
10 Zelikovsky. Inference of clonal selec-
11 tion in cancer populations using single-
12 cell sequencing data. *bioRxiv*, page
13 465211, January 2019. doi: 10.1101/
14 465211. URL [https://www.biorxiv.org/](https://www.biorxiv.org/content/10.1101/465211v2)
15 [content/10.1101/465211v2](https://www.biorxiv.org/content/10.1101/465211v2).
- 16 Martin D Smith, Joel O Wertheim, Steven
17 Weaver, Ben Murrell, Konrad Scheffler, and
18 Sergei L Kosakovsky Pond. Less is more: an
19 adaptive branch-site random effects model
20 for efficient detection of episodic diversify-
21 ing selection. *Mol. Biol. Evol.*, 32(5):1342–
22 1353, May 2015.
- 23 Charlotte Soneson and Mark D Robinson. To-
24 wards unified quality verification of syn-
25 thetic count data with countsimQC. *Bioin-*
26 *formatics*, 34(4):691–692, 2017.
- 27 Charlotte Soneson and Mark D Robinson.
28 Bias, robustness and scalability in single-
29 cell differential expression analysis. *Nat.*
30 *Methods*, February 2018.
- 31 Bastiaan Spanjaard, Bo Hu, Nina Mitic,
32 Pedro Olivares-Chauvet, Sharan Janjuha,
33 Nikolay Ninov, and Jan Philipp Junker.
34 Simultaneous lineage tracing and cell-type
35 identification using CRISPR-Cas9-induced
36 genetic scars. *Nat. Biotechnol.*, 36(5):469–
37 473, June 2018.
- 38 C Spits, C Le Caignec, M De Rycke,
39 L Van Haute, A Van Steirteghem,
I Liebaers, and K Sermon. Optimization
and evaluation of single-cell whole-genome
multiple displacement amplification. *Hum.*
Mutat., 27(5):496–503, 2006a.
- Claudia Spits, Cédric Le Caignec, Martine
De Rycke, Lindsey Van Haute, André
Van Steirteghem, Inge Liebaers, and Karen
Sermon. Whole-genome multiple displace-
ment amplification from single cells. *Nat.*
Protoc., 1(4):1965–1970, November 2006b.
- S Srinivasan, N T Johnson, and D Ko-
rkin. A Hybrid Deep Clustering Ap-
proach for Robust Cell Type Profiling Us-
ing Single-cell RNA-seq Data. *bioRxiv*,
2019. URL [https://www.biorxiv.org/](https://www.biorxiv.org/content/10.1101/511626v1.abstract)
[content/10.1101/511626v1.abstract](https://www.biorxiv.org/content/10.1101/511626v1.abstract).
- Divyanshu Srivastava, Arvind Iyer, Vibhor
Kumar, and Debarka Sengupta. Cel-
lAtlasSearch: a scalable search engine
for single cells. *Nucleic Acids Re-*
search, 46(W1):W141–W147, July 2018.
ISSN 0305-1048. doi: 10.1093/nar/
gky421. URL [https://academic.oup.](https://academic.oup.com/nar/article/46/W1/W141/5000022)
[com/nar/article/46/W1/W141/5000022](https://academic.oup.com/nar/article/46/W1/W141/5000022).
- Oliver Stegle, Sarah A Teichmann, and
John C Marioni. Computational and an-
alytical challenges in single-cell transcrip-
tomics. *Nat. Rev. Genet.*, 16(3):133, Jan-
uary 2015.
- Genevieve L Stein-O’Brien, Brian S Clark,
Thomas Sherman, Cristina Zibetti, Qi-
wen Hu, Rachel Sealfon, Sheng Liu, Jiang
Qian, Carlo Colantuoni, Seth Blackshaw,
Loyal A Goff, and Elana J Fertig. De-
composing Cell Identity for Transfer Learn-
ing across Cellular Measurements, Plat-
forms, Tissues, and Species. *Cell sys-*
tems, 8(5):395–411.e8, May 2019. ISSN
2405-4720, 2405-4712. doi: 10.1016/j.cels.
2019.04.004. URL [http://dx.doi.org/](http://dx.doi.org/10.1016/j.cels.2019.04.004)
[10.1016/j.cels.2019.04.004](http://dx.doi.org/10.1016/j.cels.2019.04.004).

- 1 Lars Steinbrück and Alice Carolyn McHardy.
2 Allele dynamics plots for the study of evolu-
3 tionary dynamics in viral populations. *Nu-*
4 *cleic Acids Res.*, 39(1):e4, January 2011. 41
- 5 Carina Strell, Markus M Hilscher, Navya
6 Laxman, Jessica Svedlund, Chenglin Wu,
7 Chika Yokota, and Mats Nilsson. Placing
8 RNA in context and space - methods for
9 spatially resolved transcriptomics. *FEBS*
10 *J.*, March 2018. 45
- 11 Tim Stuart, Andrew Butler, Paul Hoffman,
12 Christoph Hafemeister, Efthymia Papalexi,
13 William M. Mauck, Marlon Stoeckius, Pe-
14 ter Smibert, and Rahul Satija. Comprehen-
15 sive integration of single cell data. *bioRxiv*,
16 page 460147, November 2018. doi: 10.1101/
17 460147. URL [https://www.biorxiv.org/](https://www.biorxiv.org/content/10.1101/460147v1)
18 [content/10.1101/460147v1](https://www.biorxiv.org/content/10.1101/460147v1). 46
- 19 Michael J T Stubbington, Orit Rozenblatt-
20 Rosen, Aviv Regev, and Sarah A Teich-
21 mann. Single-cell transcriptomics to ex-
22 plore the immune system in health and dis-
23 ease. *Science*, 358(6359):58–63, October
24 2017. 47
- 25 Patrik L. Ståhl, Fredrik Salmén, Sanja Vick-
26 ovic, Anna Lundmark, José Fernández
27 Navarro, Jens Magnusson, Stefania Gia-
28 comello, Michaela Asp, Jakub O. West-
29 holm, Mikael Huss, Annelie Mollbrink,
30 Sten Linnarsson, Simone Codeluppi, Åke
31 Borg, Fredrik Pontén, Paul Igor Costea,
32 Pelin Sahlén, Jan Mulder, Olaf Bergmann,
33 Joakim Lundeberg, and Jonas Frisén. Vi-
34 sualization and analysis of gene expres-
35 sion in tissue sections by spatial transcrip-
36 tomics. *Science (New York, N.Y.)*, 353
37 (6294):78–82, July 2016. ISSN 1095-9203.
38 doi: 10.1126/science.aaf2403. 48
- 39 S Sun, J Zhu, Y Ma, and X Zhou. Ac-
40 curacy, Robustness and Scalability of Di-
49 mensionality Reduction Methods for Sin-
50 gle Cell RNAseq Analysis. *bioRxiv*,
51 2019. URL [https://www.biorxiv.org/](https://www.biorxiv.org/content/10.1101/641142v1.abstract)
52 [content/10.1101/641142v1.abstract](https://www.biorxiv.org/content/10.1101/641142v1.abstract). 49
- 53 Valentine Svensson, Sarah A Teichmann, and
54 Oliver Stegle. SpatialDE: identification of
55 spatially variable genes. *Nat. Methods*, 15
(5):343–346, May 2018a. 50
- 56 Valentine Svensson, Roser Vento-Tormo, and
57 Sarah A. Teichmann. Exponential scal-
58 ing of single-cell RNA-seq in the past
59 decade. *Nature Protocols*, 13(4):599–604,
60 April 2018b. ISSN 1750-2799. doi: 10.
61 1038/nprot.2017.149. URL [https://www.](https://www.nature.com/articles/nprot.2017.149)
62 [nature.com/articles/nprot.2017.149](https://www.nature.com/articles/nprot.2017.149). 51
- 63 Charles Swanton. Intratumor heterogeneity:
64 evolution through space and time. *Cancer*
65 *Res.*, 72(19):4875–4882, October 2012. 52
- 66 Ewa Szczurek, Navodit Misra, and Martin
67 Vingron. Synthetic sickness or lethality
68 points at candidate combination therapy
69 targets in glioblastoma. *Int. J. Cancer*, 133
70 (9):2123–2132, November 2013. 53
- 71 The Tabula Muris Consortium. Single-cell
72 transcriptomics of 20 mouse organs cre-
73 ates a Tabula Muris. *Nature*, 562(7727):
74 367, October 2018. ISSN 1476-4687.
75 doi: 10.1038/s41586-018-0590-4. URL
76 [https://www.nature.com/articles/](https://www.nature.com/articles/s41586-018-0590-4)
77 [s41586-018-0590-4](https://www.nature.com/articles/s41586-018-0590-4). 54
- 78 Divyanshu Talwar, Aanchal Mongia, De-
79 barka Sengupta, and Angshul Majum-
80 dar. AutoImpute: Autoencoder based
81 imputation of single-cell RNA-seq data.
82 *Scientific reports*, 8(1):16329, November
83 2018. ISSN 2045-2322. doi: 10.1038/
84 s41598-018-34688-x. URL [http://dx.](http://dx.doi.org/10.1038/s41598-018-34688-x)
85 [doi.org/10.1038/s41598-018-34688-x](http://dx.doi.org/10.1038/s41598-018-34688-x). 55

- 1 Amos Tanay and Aviv Regev. Scaling
2 single-cell genomics from phenomenology
3 to mechanism. *Nature*, 541(7637):331–338,
4 January 2017.
- 5 W Tang, F Bertaux, P Thomas, C Ste-
6 fanelli, M Saint, and others. bayNorm:
7 Bayesian gene expression recovery, im-
8 putation and normalisation for single
9 cell RNA-sequencing data. *bioRxiv*,
10 2018. URL [https://www.biorxiv.org/
11 content/10.1101/384586v2.abstract](https://www.biorxiv.org/content/10.1101/384586v2.abstract).
- 12 H Telenius, N P Carter, C E Bebb, M Norden-
13 skjöld, B A Ponder, and A Tunnacliffe. De-
14 generate oligonucleotide-primed PCR: gen-
15 eral amplification of target DNA by a single
16 degenerate primer. *Genomics*, 13(3):718–
17 725, July 1992.
- 18 Luyi Tian, Xueyi Dong, Saskia Freytag,
19 Kim-Anh Lê Cao, Shian Su, Abolfazl
20 JalalAbadi, Daniela Amann-Zalcenstein,
21 Tom S. Weber, Azadeh Seidi, Jafar S.
22 Jabbari, Shalin H. Naik, and Matthew E.
23 Ritchie. Benchmarking single cell RNA-
24 sequencing analysis pipelines using mixture
25 control experiments. *Nature Methods*, 16
26 (6):479, June 2019. ISSN 1548-7105.
27 doi: 10.1038/s41592-019-0425-8. URL
28 [https://www.nature.com/articles/
29 s41592-019-0425-8](https://www.nature.com/articles/s41592-019-0425-8).
- 30 F William Townes, Stephanie C Hicks,
31 Martin J Aryee, and Rafael A Irizarry.
32 Feature Selection and Dimension Reduc-
33 tion for Single Cell RNA-Seq based on a
34 Multinomial Model. March 2019. URL
35 [https://www.biorxiv.org/content/10.
36 1101/574574v1](https://www.biorxiv.org/content/10.1101/574574v1).
- 37 Cole Trapnell, Davide Cacchiarelli, Jonna
38 Grimsby, Prapti Pokharel, Shuqiang Li,
39 Michael Morse, Niall J. Lennon, Kenneth J.
40 Livak, Tarjei S. Mikkelsen, and John L.
Rinn. The dynamics and regulators of cell
fate decisions are revealed by pseudotempo-
ral ordering of single cells. *Nature Biotech-*
nology, 32(4):381–386, April 2014. ISSN
1546-1696. doi: 10.1038/nbt.2859.
- Samra Turajlic and Charles Swanton. Metas-
tasis as an evolutionary process. *Science*,
352(6282):169–175, April 2016.
- Vincent van Unen, Thomas Höllt, Nicola
Pezzotti, Na Li, Marcel J. T. Rein-
ders, Elmar Eisemann, Frits Koning,
Anna Vilanova, and Boudewijn P. F.
Lelieveldt. Visual analysis of mass cy-
tometry data by hierarchical stochastic
neighbour embedding reveals rare cell
types. *Nature Communications*, 8(1):
1740, November 2017. ISSN 2041-1723.
doi: 10.1038/s41467-017-01689-9. URL
[https://www.nature.com/articles/
s41467-017-01689-9](https://www.nature.com/articles/s41467-017-01689-9).
- Catalina A Vallejos, John C Marioni, and
Sylvia Richardson. BASiCS: Bayesian anal-
ysis of Single-Cell sequencing data. *PLoS
Comput. Biol.*, 11(6):e1004333, June 2015.
- Trieu My Van and Christian U. Blank.
A user’s perspective on GeoMxTM
digital spatial profiling. *Immuno-
Oncology Technology*, 1:11–18, July
2019. ISSN 2590-0188, 2590-0188.
doi: 10.1016/j.iotech.2019.05.001. URL
[https://www.esmoitech.org/article/
S2590-0188\(19\)30002-4/abstract](https://www.esmoitech.org/article/S2590-0188(19)30002-4/abstract).
- Koen van den Berge, Hector Roux de
Bezieux, Kelly Street, Wouter Saelens, Ro-
brecht Cannoodt, Yvan Saeys, Sandrine
Dudoit, and Lieven Clement. Trajectory-
based differential expression analysis for
single-cell sequencing data. *bioRxiv*,
page 623397, May 2019. doi: 10.1101/
623397. URL [https://www.biorxiv.org/
content/10.1101/623397v1](https://www.biorxiv.org/content/10.1101/623397v1).

- 1 Dimitrios V Vavoulis, Margherita
2 Francescatto, Peter Heutink, and Ju-
3 lian Gough. DGEclust: differential
4 expression analysis of clustered count data.
5 *Genome Biol.*, 16:39, February 2015. 41
- 6 A Verma and B Engelhardt. A robust nonlin-
7 ear low-dimensional manifold for single cell
8 RNA-seq data. *bioRxiv*, 2018. 42
- 9 Beate Vieth, Christoph Ziegenhain, Swati
10 Parekh, Wolfgang Enard, and Ines Hell-
11 mann. powsimr: power analysis for
12 bulk and single cell RNA-seq experiments.
13 *Bioinformatics*, 33(21):3486–3488, Novem-
14 ber 2017. 43
- 15 Irma Virant-Klun, Stefan Leicht, Christopher
16 Hughes, and Jeroen Krijgsveld. Identifi-
17 cation of Maturation-Specific Proteins by
18 Single-Cell Proteomics of Human Oocytes.
19 *Molecular & cellular proteomics: MCP*, 15
20 (8):2616–2627, 2016. ISSN 1535-9484. doi:
21 10.1074/mcp.M115.056887. 44
- 22 Sarah A. Vitak, Kristof A. Torkenczy, Jimi L.
23 Rosenkrantz, Andrew J. Fields, Lena
24 Christiansen, Melissa H. Wong, Lucia Car-
25 bone, Frank J. Steemers, and Andrew
26 Adey. Sequencing thousands of single-cell
27 genomes with combinatorial indexing. *Na-
28 ture Methods*, 14(3):302–308, March 2017.
29 ISSN 1548-7105. doi: 10.1038/nmeth.
30 4154. URL [https://www.nature.com/
31 articles/nmeth.4154](https://www.nature.com/articles/nmeth.4154). 45
- 32 Bartlomiej Wacław, Ivana Bozic, Meredith E
33 Pittman, Ralph H Hruban, Bert Vogelstein,
34 and Martin A Nowak. A spatial model
35 predicts that dispersal and cell turnover
36 limit intratumour heterogeneity. *Nature*,
37 525(7568):261–264, September 2015. 46
- 38 Daniel E. Wagner, Caleb Weinreb, Zach M.
39 Collins, James A. Briggs, Sean G. Megason,
40 and Allon M. Klein. Single-cell mapping of
gene expression landscapes and lineage in
the zebrafish embryo. *Science*, 360(6392):
981–987, June 2018a. ISSN 0036-8075,
1095-9203. doi: 10.1126/science.aar4362.
URL [http://science.sciencemag.org/
content/360/6392/981](http://science.sciencemag.org/content/360/6392/981). 47
- F Wagner, D Barkley, and I Yanai. Accurate
denoising of single-cell RNA-Seq data us-
ing unbiased principal component analysis.
bioRxiv, 2019. 48
- Florian Wagner and Itai Yanai. Moana: A
robust and scalable cell type classifica-
tion framework for single-cell RNA-Seq
data. *bioRxiv*, page 456129, Octo-
ber 2018. doi: 10.1101/456129. URL
[https://www.biorxiv.org/content/10.
1101/456129v1](https://www.biorxiv.org/content/10.1101/456129v1). 49
- Florian Wagner, Yun Yan, and Itai
Yanai. K-nearest neighbor smooth-
ing for high-throughput single-cell
RNA-Seq data. January 2018b. URL
[https://www.biorxiv.org/content/
early/2018/01/24/217737?rss=1](https://www.biorxiv.org/content/early/2018/01/24/217737?rss=1). 50
- Dongfang Wang and Jin Gu. VASC: Dimen-
sion reduction and visualization of single-
cell RNA-seq data by deep variational au-
toencoder. *Genomics Proteomics Bioinfor-
matics*, 16(5):320–331, October 2018. 51
- Jian Wang and Yuanlin Song. Single cell se-
quencing: a distinct new field. *Clin. Transl.
Med.*, 6(1):10, December 2017. 52
- Jingshu Wang, Divyansh Agarwal,
Mo Huang, Gang Hu, Zilu Zhou, Vin-
cent B Conley, Hugh MacMullan, and
Nancy R Zhang. Transfer learning in
single-cell transcriptomics improves data
denoising and pattern discovery. November
2018. 53

- 1 T Wang, T S Johnson, W Shao, J Zhang, and
2 K Huang. BURMUDA: A novel deep trans-
3 fer learning method for single-cell RNA se-
4 quencing batch correction reveals hidden
5 high-resolution cellular subtypes. *bioRxiv*,
6 2019.
- 7 Gregory P Way and Casey S Greene. Extract-
8 ing a biologically relevant latent space from
9 cancer transcriptomes with variational au-
10 toencoders. *Pacific Symposium on Bio-*
11 *computing. Pacific Symposium on Biocom-*
12 *puting*, 23:80–91, 2018. ISSN 2335-6936,
13 2335-6928. doi: 10.1142/9789813235533\
14 _0008. URL [https://www.ncbi.nlm.](https://www.ncbi.nlm.nih.gov/pubmed/29218871)
15 [nih.gov/pubmed/29218871](https://www.ncbi.nlm.nih.gov/pubmed/29218871).
- 16 Lukas M. Weber and Mark D. Robinson.
17 Comparison of clustering methods for
18 high-dimensional single-cell flow and mass
19 cytometry data. *Cytometry Part A*, 89
20 (12):1084–1096, December 2016. ISSN
21 1552-4922. doi: 10.1002/cyto.a.23030.
22 URL [https://onlinelibrary.wiley.](https://onlinelibrary.wiley.com/doi/full/10.1002/cyto.a.23030)
23 [com/doi/full/10.1002/cyto.a.23030](https://onlinelibrary.wiley.com/doi/full/10.1002/cyto.a.23030).
- 24 Lukas M. Weber, Malgorzata Nowicka, Char-
25 lotte Soneson, and Mark D. Robin-
26 son. diffcyt: Differential discovery
27 in high-dimensional cytometry via high-
28 resolution clustering. *bioRxiv*, page
29 349738, November 2018. doi: 10.1101/
30 349738. URL [https://www.biorxiv.org/](https://www.biorxiv.org/content/10.1101/349738v2)
31 [content/10.1101/349738v2](https://www.biorxiv.org/content/10.1101/349738v2).
- 32 Lukas M. Weber, Wouter Saelens, Robrecht
33 Cannoodt, Charlotte Soneson, Alexander
34 Hapfelmeier, Paul P. Gardner, Anne-Laure
35 Boulesteix, Yvan Saeys, and Mark D.
36 Robinson. Essential guidelines for compu-
37 tational method benchmarking. *Genome*
38 *Biology*, 20(1):125, June 2019. ISSN 1474-
39 760X. doi: 10.1186/s13059-019-1738-8.
40 URL [https://doi.org/10.1186/](https://doi.org/10.1186/s13059-019-1738-8)
41 [s13059-019-1738-8](https://doi.org/10.1186/s13059-019-1738-8).
- Caleb Weinreb, Samuel Wolock, Betsabeh K.
Tusi, Merav Socolovsky, and Allon M.
Klein. Fundamental limits on dynamic in-
ference from single-cell snapshots. *Proceed-*
ings of the National Academy of Sciences,
115(10):E2467–E2476, March 2018. ISSN
0027-8424, 1091-6490. doi: 10.1073/pnas.
1714723115. URL [https://www.pnas.](https://www.pnas.org/content/115/10/E2467)
org/content/115/10/E2467.
- Joshua Welch, Velina Kozareva, Ashley Fer-
reira, Charles Vanderburg, Carly Martin,
and Evan Macosko. Integrative inference
of brain cell similarities and differences
from single-cell genomics. *bioRxiv*, page
459891, November 2018. doi: 10.1101/
459891. URL [https://www.biorxiv.org/](https://www.biorxiv.org/content/10.1101/459891v1)
content/10.1101/459891v1.
- Joshua D Welch, Alexander J Hartemink, and
Jan F Prins. MATCHER: manifold align-
ment reveals correspondence between single
cell transcriptome and epigenome dynam-
ics. *Genome Biol.*, 18(1):138, July 2017.
- F Alexander Wolf, Philipp Angerer, and
Fabian J Theis. SCANPY: large-scale
single-cell gene expression data analysis.
Genome Biol., 19(1):15, February 2018.
- F. Alexander Wolf, Fiona K. Hamey,
Mireya Plass, Jordi Solana, Joakim S.
Dahlin, Berthold Göttgens, Nikolaus Ra-
jewsky, Lukas Simon, and Fabian J.
Theis. PAGA: graph abstraction recon-
ciles clustering with trajectory inference
through a topology preserving map of sin-
gle cells. *Genome Biology*, 20(1):59, March
2019. ISSN 1474-760X. doi: 10.1186/
s13059-019-1663-x. URL [https://doi.](https://doi.org/10.1186/s13059-019-1663-x)
org/10.1186/s13059-019-1663-x.
- Larry Xi, Alexander Belyaev, Sandra Spur-
geon, Xiaohui Wang, Haibiao Gong, Robert
Aboukhalil, and Richard Fekete. New li-
brary construction method for single-cell

- 1 genomes. *PLoS One*, 12(7):e0181163, July
2 2017.
- 3 Li Charlie Xia, Dongmei Ai, Hojoon Lee,
4 Noemi Andor, Chao Li, Nancy R Zhang,
5 and Hanlee P Ji. SVEngine: an efficient
6 and versatile simulator of genome struc-
7 tural variations with features of cancer
8 clonal evolution. *Gigascience*, 7(7), July
9 2018.
- 10 Li Yang and P Charles Lin. Mechanisms that
11 drive inflammatory tumor microenviron-
12 ment, tumor heterogeneity, and metastatic
13 progression. *Semin. Cancer Biol.*, 47:185–
14 195, December 2017.
- 15 Z Yang. Maximum likelihood phylogenetic es-
16 timation from DNA sequences with variable
17 rates over sites: approximate methods. *J.*
18 *Mol. Evol.*, 39(3):306–314, September 1994.
- 19 Yinyin Yuan. Spatial heterogeneity in the tu-
20 mor microenvironment. *Cold Spring Harb.*
21 *Perspect. Med.*, 6(8), August 2016.
- 22 Simone Zaccaria, Mohammed El-Kebir, Gun-
23 nar W. Klau, and Benjamin J. Raphael.
24 The Copy-Number Tree Mixture Deconvolu-
25 tion Problem and Applications to Multi-
26 sample Bulk Sequencing Tumor Data. In
27 S. Cenk Sahinalp, editor, *Research in*
28 *Computational Molecular Biology*, Lecture
29 Notes in Computer Science, pages 318–335.
30 Springer International Publishing, 2017.
31 ISBN 978-3-319-56970-3.
- 32 H Zafar, N Navin, K Chen, and L Nakhleh.
33 SiCloneFit: Bayesian inference of popula-
34 tion structure, genotype, and phylogeny of
35 tumor clones from single-cell genome se-
36 quencing data. *bioRxiv*, 2018.
- 37 Hamim Zafar, Yong Wang, Luay Nakhleh,
38 Nicholas Navin, and Ken Chen. Mono-
39 var: single-nucleotide variant detection in
single cells. *Nature Methods*, 13(6):505–
507, June 2016. ISSN 1548-7105. doi:
10.1038/nmeth.3835. URL [https://www.](https://www.nature.com/articles/nmeth.3835)
[nature.com/articles/nmeth.3835](https://www.nature.com/articles/nmeth.3835).
- Hamim Zafar, Anthony Tzen, Nicholas Navin,
Ken Chen, and Luay Nakhleh. SiFit: infer-
ring tumor trees from single-cell sequenc-
ing data under finite-sites models. *Genome*
Biol., 18(1):178, September 2017.
- Hans Zahn, Adi Steif, Emma Laks, Peter
Eirew, Michael VanInsberghe, Sohrab P
Shah, Samuel Aparicio, and Carl L Hansen.
Scalable whole-genome single-cell library
preparation without preamplification. *Nat.*
Methods, 14(2):167–173, February 2017a.
- Hans Zahn, Adi Steif, Emma Laks, Peter
Eirew, Michael VanInsberghe, Sohrab P.
Shah, Samuel Aparicio, and Carl L.
Hansen. Scalable whole-genome single-
cell library preparation without preampli-
fication. *Nature Methods*, 14(2):167–173,
February 2017b. ISSN 1548-7105. doi:
10.1038/nmeth.4140. URL [https://www.](https://www.nature.com/articles/nmeth.4140)
[nature.com/articles/nmeth.4140](https://www.nature.com/articles/nmeth.4140).
- Luke Zappia, Belinda Phipson, and Alicia
Oshlack. Splatter: simulation of single-cell
RNA sequencing data. *Genome Biol.*, 18
(1):174, September 2017.
- Ron Zeira and Ron Shamir. Genome
Rearrangement Problems with Sin-
gle and Multiple Gene Copies : A
Review. Not clear where this was
initially published and whether it is
peer-reviewed., 2018. URL [https:](https://pdfs.semanticscholar.org/85e6/7eb03d1b3d004c60a12df08c1f937fbba974.pdf)
[/pdfs.semanticscholar.org/85e6/](https://pdfs.semanticscholar.org/85e6/7eb03d1b3d004c60a12df08c1f937fbba974.pdf)
[7eb03d1b3d004c60a12df08c1f937fbba974.](https://pdfs.semanticscholar.org/85e6/7eb03d1b3d004c60a12df08c1f937fbba974.pdf)
[pdf](https://pdfs.semanticscholar.org/85e6/7eb03d1b3d004c60a12df08c1f937fbba974.pdf).
- Amit Zeisel, Hannah Hochgerner, Peter
Lönnerberg, Anna Johnsson, Fatima
Memic, Job van der Zwan, Martin Häring,

- 1 Emelie Braun, Lars E. Borm, Gioele
2 La Manno, Simone Codeluppi, Alessan-
3 dro Furlan, Kawai Lee, Nathan Skene,
4 Kenneth D. Harris, Jens Hjerling-Leffler,
5 Ernest Arenas, Patrik Ernfors, Ulrika
6 Marklund, and Sten Linnarsson. Molec-
7 ular Architecture of the Mouse Nervous
8 System. *Cell*, 174(4):999–1014.e22,
9 August 2018. ISSN 0092-8674. doi:
10 10.1016/j.cell.2018.06.021. URL [http:](http://www.sciencedirect.com/science/article/pii/S009286741830789X)
11 [//www.sciencedirect.com/science/](http://www.sciencedirect.com/science/article/pii/S009286741830789X)
12 [article/pii/S009286741830789X](http://www.sciencedirect.com/science/article/pii/S009286741830789X).
- 13 Allen W. Zhang, Ciara O’Flanagan, Eliz-
14 abeth Chavez, Jamie LP Lim, Andrew
15 McPherson, Matt Wiens, Pascale Wal-
16 ters, Tim Chan, Brittany Hewitson, Daniel
17 Lai, Anja Mottok, Clementine Sarkozy,
18 Lauren Chong, Tomohiro Aoki, Xue-
19 hai Wang, Andrew P. Weng, Jessica N.
20 McAlpine, Samuel Aparicio, Christian
21 Steidl, Kieran R. Campbell, and Sohrab P.
22 Shah. Probabilistic cell type assignment
23 of single-cell transcriptomic data reveals
24 spatiotemporal microenvironment dynam-
25 ics in human cancers. *bioRxiv*, page
26 521914, January 2019a. doi: 10.1101/
27 521914. URL [https://www.biorxiv.org/](https://www.biorxiv.org/content/10.1101/521914v1)
28 [content/10.1101/521914v1](https://www.biorxiv.org/content/10.1101/521914v1).
- 29 Chao Zhang. Single-Cell Data Analysis Us-
30 ing MMD Variational Autoencoder. April
31 2019. URL [https://www.biorxiv.org/](https://www.biorxiv.org/content/10.1101/613414v1.abstract)
32 [content/10.1101/613414v1.abstract](https://www.biorxiv.org/content/10.1101/613414v1.abstract).
- 33 Huanan Zhang, Catherine A A Lee, Zhuliu Li,
34 John R Garbe, Cindy R Eide, Raphael Pe-
35 tegrosso, Rui Kuang, and Jakub Tolar. A
36 multitask clustering approach for single-cell
37 RNA-seq analysis in recessive dystrophic
38 epidermolysis bullosa. *PLoS Comput. Biol.*,
39 14(4):e1006053, April 2018.
- 40 Jesse M. Zhang, Govinda M. Kamath,
41 and David N. Tse. Valid post-clustering
differential analysis for single-cell RNA-
Seq. *bioRxiv*, page 463265, June
2019b. doi: 10.1101/463265. URL
[https://www.biorxiv.org/content/10.](https://www.biorxiv.org/content/10.1101/463265v3)
1101/463265v3.
- Jianzhi Zhang, Rasmus Nielsen, and Ziheng
Yang. Evaluation of an improved branch-
site likelihood method for detecting posi-
tive selection at the molecular level. *Mol.*
Biol. Evol., 22(12):2472–2479, December
2005.
- Jingsong Zhang, Jessica J. Cunningham,
Joel S. Brown, and Robert A. Gatenby.
Integrating evolutionary dynamics into
treatment of metastatic castrate-resistant
prostate cancer. *Nature Communications*, 8
(1):1816, November 2017. ISSN 2041-1723.
doi: 10.1038/s41467-017-01968-5. URL
[https://www.nature.com/articles/](https://www.nature.com/articles/s41467-017-01968-5)
s41467-017-01968-5.
- L. Zhang and S. Zhang. Comparison of com-
putational methods for imputing single-
cell RNA-sequencing data. *IEEE/ACM*
Transactions on Computational Biology
and Bioinformatics, pages 1–1, 2018. ISSN
1545-5963. doi: 10.1109/TCBB.2018.
2848633.
- L Zhang, X Cui, K Schmitt, R Hubert, W Na-
vidi, and N Arnheim. Whole genome am-
plification from a single cell: implications
for genetic analysis. *Proc. Natl. Acad. Sci.*
U. S. A., 89(13):5847–5851, July 1992.
- Xiao-Fei Zhang, Le Ou-Yang, Shuo Yang,
Xing-Ming Zhao, Xiaohua Hu, and Hong
Yan. EnImpute: imputing dropout
events in single cell RNA sequencing
data via ensemble learning. *Bioinform-*
atics, May 2019c. ISSN 1367-4803,
1367-4811. doi: 10.1093/bioinformatics/
btz435. URL [http://dx.doi.org/10.](http://dx.doi.org/10.1093/bioinformatics/btz435)
1093/bioinformatics/btz435.

- 1 Xiaoyan Zhang, Sadie L Marjani, Zhaoyang
2 Hu, Sherman M Weissman, Xinghua Pan,
3 and Shixiu Wu. Single-Cell sequencing
4 for precise cancer research: Progress and
5 prospects. *Cancer Res.*, 76(6):1305–1312,
6 March 2016.
- 7 Xiuwei Zhang, Chenling Xu, and Nir Yosef.
8 SymSim: simulating multi-faceted variabil-
9 ity in single cell RNA sequencing. *bioRxiv*,
10 page 378646, April 2019d. doi: 10.1101/
11 378646. URL [https://www.biorxiv.org/
12 content/10.1101/378646v3](https://www.biorxiv.org/content/10.1101/378646v3).
- 13 Yifan Zhang and Feng Liu. Multidimensional
14 Single-Cell analyses in organ development
15 and maintenance. *Trends Cell Biol.*, March
16 2019.
- 17 Grace X. Y. Zheng, Jessica M. Terry, Phillip
18 Belgrader, Paul Ryvkin, Zachary W.
19 Bent, Ryan Wilson, Solongo B. Ziraldo,
20 Tobias D. Wheeler, Geoff P. McDer-
21 mott, Junjie Zhu, Mark T. Gregory, Joe
22 Shuga, Luz Montesclaros, Jason G. Under-
23 wood, Donald A. Masquelier, Stefanie Y.
24 Nishimura, Michael Schnall-Levin, Paul W.
25 Wyatt, Christopher M. Hindson, Rajiv
26 Bharadwaj, Alexander Wong, Kevin D.
27 Ness, Lan W. Beppu, H. Joachim Deeg,
28 Christopher McFarland, Keith R. Loeb,
29 William J. Valente, Nolan G. Ericson,
30 Emily A. Stevens, Jerald P. Radich, Tar-
31 jei S. Mikkelsen, Benjamin J. Hindson,
32 and Jason H. Bielas. Massively paral-
33 lel digital transcriptional profiling of sin-
34 gle cells. *Nature Communications*, 8:14049,
35 January 2017. ISSN 2041-1723. doi: 10.
36 1038/ncomms14049. URL [https://www.
37 nature.com/articles/ncomms14049](https://www.nature.com/articles/ncomms14049).
- 38 Lingxue Zhu, Jing Lei, Bernie Devlin, and
39 Kathryn Roeder. A UNIFIED STATISTI-
40 CAL FRAMEWORK FOR SINGLE CELL
41 AND BULK RNA SEQUENCING DATA.
The annals of applied statistics, 12(1):609–
632, March 2018. ISSN 1932-6157. doi:
10.1214/17-AOAS1110. URL [http://dx.
doi.org/10.1214/17-AOAS1110](http://dx.doi.org/10.1214/17-AOAS1110).
- Rapolas Zilionis, Juozas Nainys, Adrian
Veres, Virginia Savova, David Zemmour,
Allon M Klein, and Linas Mazutis. Single-
cell barcoding and sequencing using droplet
microfluidics. *Nat. Protoc.*, 12(1):44–73,
January 2017.
- Chenghang Zong, Sijia Lu, Alec R Chapman,
and X Sunney Xie. Genome-wide detection
of single-nucleotide and copy-number vari-
ations of a single human cell. *Science*, 338
(6114):1622–1626, December 2012.