# 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 Baaijens[29], Marleen Balvert[29,31], Buys de Barbanson[32,33,34], Antonio Cappuccio[35], Giacomo Corleone[36], Bas E. Dutilh[31,38], Maria Florescu[32,33,34], Victor Guryev[41], Rens Holmer[42], Katharina Jahn[12,13], Thamar Jessurun Lobo[41], Emma M. Keizer[45], Indu Khatri[46], Szymon M. Kiełbasa[47], Jan O. Korbel[48], Alexey M. Kozlov[24], Tzu-Hao Kuo[3], Boudewijn P.F. Lelieveldt[49,50], Ion I. Mandoiu[51], John C. Marioni[52,53,54], Tobias Marschall[55,56], Felix Mölder[1,59], Amir Niknejad[60,61], Łukasz Rączkowski[5], Marcel Reinders[18,19], Jeroen de Ridder[32,33], Antoine-Emmanuel Saliba[62], Antonios Somarakis[50], Oliver Stegle[48,54,63], Fabian J. Theis[67], Huan Yang[68], Alex Zelikovsky[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.
[1] Algorithms for Reproducible Bioinformatics, Genome Informatics, Institute of Human Genetics, University Hospital Essen, University of Duisburg-Essen, Germany
[2] Department of Paediatric Oncology, Haematology and Immunology, Medical Faculty, Heinrich Heine University, University Hospital, Düsseldorf, Germany
[3] Computational Biology of Infection Research Group, Helmholtz Centre for Infection Research, Braunschweig, Germany
[4] Medical Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, USA
[5] Institute of Informatics, Faculty of Mathematics, Informatics and Mechanics University of Warsaw, Poland
[6] Bioinformatics and Cellular Genomics, St Vincent's Institute of Medical Research, Fitzroy, Australia
[7] Melbourne Integrative Genomics, School of BioSciences — School of Mathematics & Statistics, Faculty of Science, University of Melbourne, Australia
[8] Department of Biostatistics, Johns Hopkins University, Baltimore, MD, USA
[9] Institute of Molecular Life Sciences and SIB Swiss Institute of Bioinformatics, University of Zurich, Switzerland
[10] MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, UK
[11] The Alan Turing Institute, British Library, London, UK
[12] Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland
[13] SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland
[15] Department of Statistics, University of British Columbia, Vancouver, Canada
[16] Department of Molecular Oncology, BC Cancer Agency, Vancouver, Canada
[17] Data Science Institute, University of British Columbia, Vancouver, Canada
[18] Leiden Computational Biology Center, Leiden University Medical Center, The Netherlands
[19] Delft Bioinformatics Lab, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, The Netherlands
[20] Molecular Pathology Unit and Center for Cancer Research, Massachusetts General Hospital Research Institute, Charlestown, USA
[21] Department of Pathology, Harvard Medical School, Boston, USA
[22] Broad Institute of Harvard and MIT, Cambridge, MA, USA
[23] Department of Computer Science, Georgia State University, Atlanta, USA
[24] Computational Molecular Evolution Group, Heidelberg Institute for Theoretical Studies, Germany
[25] Institute for Theoretical Informatics, Karlsruhe Institute of Technology, Germany
[26] Institute for Research in Biomedicine, The Barcelona Institute of Science and Technology, Spain
[27] Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, Canada
[29] Life Sciences and Health, Centrum Wiskunde & Informatica, Amsterdam, The Netherlands
[31] Theoretical Biology and Bioinformatics, Science for Life, Utrecht University, The Netherlands
[32] Center for Molecular Medicine, University Medical Center Utrecht, The Netherlands
[33] Oncode Institute, Utrecht, The Netherlands
[34] Quantitative biology, Hubrecht Institute, Utrecht, The Netherlands
[35] Institute for Advanced Study, University of Amsterdam, The Netherlands
[36] Department of Surgery and Cancer, The Imperial Centre for Translational and Experimental Medicine, Imperial College London, UK
[38] Centre for Molecular and Biomolecular Informatics, Radboud University Medical Center, Nijmegen, The Netherlands
[41] European Research Institute for the Biology of Ageing, University Medical Center Groningen, University of Groningen, The Netherlands
[42] Bioinformatics Group, Wageningen University, The Netherlands
[45] Biometris, Wageningen University & Research, The Netherlands
[46] Department of Immunohematology and Blood Transfusion, Leiden University Medical Center, The Netherlands
[47] Department of Biomedical Data Sciences, Leiden University Medical Center, The Netherlands
[48] Genome Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany
[49] PRB lab, Delft University of Technology, The Netherlands
[50] Division of Image Processing, Department of Radiology, Leiden University Medical Center, The Netherlands
[51] Computer Science & Engineering Department, University of Connecticut, Storrs, USA
[52] Cancer Research UK Cambridge Institute, Li Ka Shing Centre, University of Cambridge, UK
[53] Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, UK
[54] European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, UK
[55] Center for Bioinformatics, Saarland University, Saarbrücken, Germany
[56] Max Planck Institute for Informatics, Saarbrücken, Germany
[59] Institute of Pathology, University Hospital Essen, University of Duisburg-Essen, Germany.
[60] Computation molecular design, Zuse Institute Berlin, Germany
[61] Mathematics department, Mount Saint Vincent, New York, USA
[62] Helmholtz Institute for RNA-based Infection Research, Helmholtz-Center for Infection Research, Würzburg, Germany
[63] Division of Computational Genomics and Systems Genetics, German Cancer Research Center – DKFZ, Heidelberg, Germany
[67] Institute of Computational Biology, Helmholtz Zentrum München – German Research Center for Environmental Health, Neuherberg, Germany
[68] Division of Drug Discovery and Safety, Leiden Academic Center for Drug Research – LACDR — Leiden University, The Netherlands
[69] Department of Computer Science, Georgia State University, Atlanta, USA
[70] The Laboratory of Bioinformatics, I.M. Sechenov First Moscow State Medical University, Moscow, Russia
[71] Department of Computer Science, Princeton University, USA
[72] Computational Oncology, Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, USA

[+] Joint last authors, workshop organizers.

[@] Corresponding author: Alexander Schönhuth, as@cwi.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

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27885v3 | CC BY 4.0 Open Access | rec: 23 Aug 2019, publ: 23 Aug 2019

# 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, ?] 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 posit that we have entered the era of Single-

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27885v3 | CC BY 4.0 Open Access | rec: 23 Aug 2019, publ: 23 Aug 2019

Cell Data Science (SCDS).

While SCDS faces many of the data science issues arising in bulk sequencing, it also substantially adds to them and further compounds existing scientific challenges. Namely, limited amounts of material available per cell lead to exceptionally high levels of uncertainty about (possibly missed) observations, and where amplification is used to generate more material, technical noise is added to the resulting data. Further, a new level of resolution also means another—rapidly growing—dimension in data matrices, thus requiring scalable models and methods for data analysis. While the particular challenges can vary greatly by research goal, tissue analyzed, experimental setup or—last but not least—just by whether DNA or RNA is sequenced, further factoring into various protocols, assaying for example also the epigenome (bisulfite protocols), chromatin accessibility (e.g. ATAC-seq) or protein levels (CITE-seq), the common denominator is that the challenges are all rooted in data science, hence are computational or statistical in nature. Here, we propose the dozen data science challenges that we believe to be most relevant for bringing SCDS forward. We summarize and categorize them, providing a thorough review of the status of each challenge relative to existing approaches. From this foundation, we point to possible directions of research to tackle them. This catalog of SCDS challenges aims at focusing the development of data analysis methods and the directions of research in this rapidly evolving field—as a guideline for researchers looking for rewarding problems that match their personal 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 categorizing 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 particularly 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 sequencing 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 particular attention. The most driving of such elementary themes, not necessarily unique to sc-seq, are: (i) The need to quantify measurement uncertainty (see challenges in section 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 specific to sc-seq, exacerbated by the rapid advances 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 measured per cell (section 2.3); this often arises in combination with: (ii) The need to integrate 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). Finally, 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 illustration[1]. 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

[1] Figure 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/`).

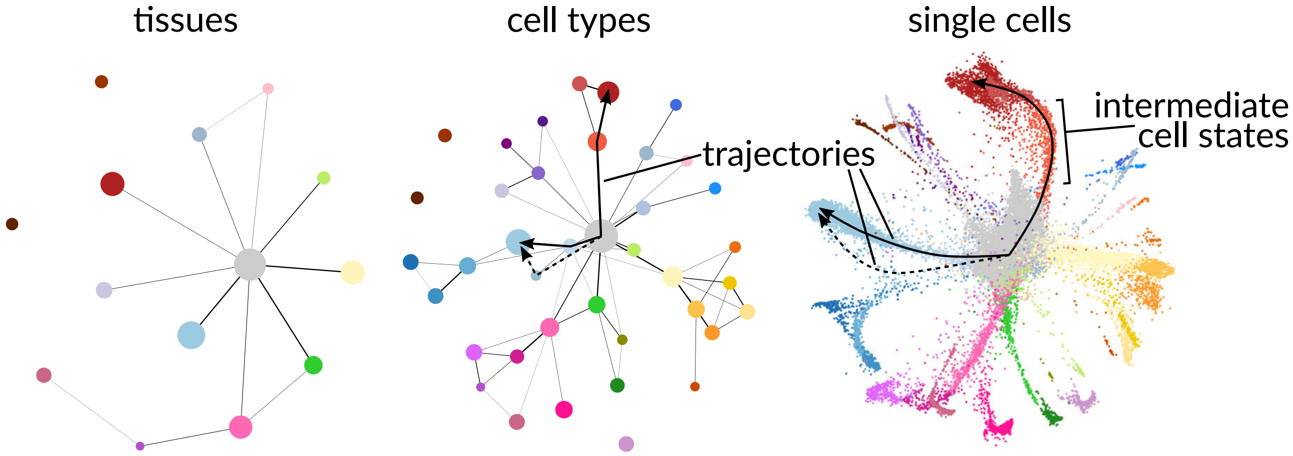

Figure 1: Different levels of resolution are of interest, depending on the research question and the data available. Thus, analysis tools and reference systems (such as cell atlases) will have to accommodate for multiple levels of resolution from whole organs and tissues over discrete cell types to continuously 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, BEARscc]).

Optimally, sc-seq analysis tools would accurately 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 Kendziorski [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 normalization. But whereas normalization 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 normalization, but do not entail all possible or useful approaches to normalization.

### 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 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 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

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27885v3 | CC BY 4.0 Open Access | rec: 23 Aug 2019, publ: 23 Aug 2019

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, visualization 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 neighbors 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 matrix 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-

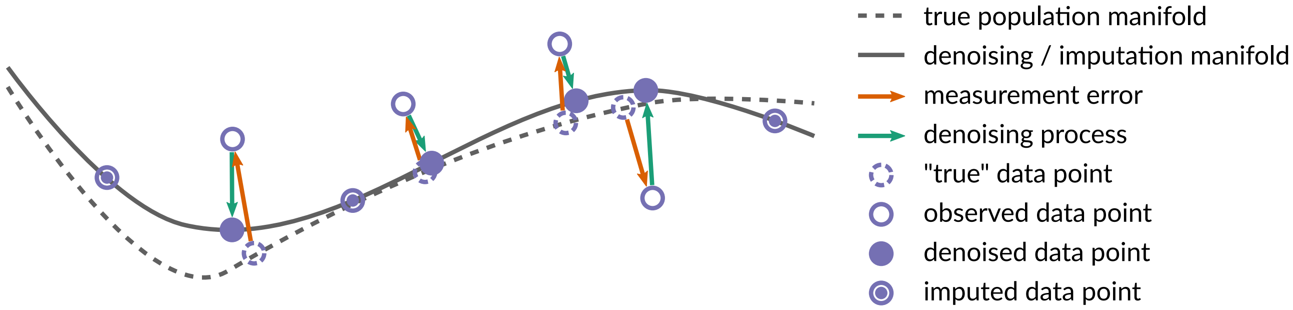

Figure 2: Measurement error requires denoising methods or approaches that quantify uncertainty and propagate it down analysis pipelines. Also, whenever methods cannot deal with the abundant missing values, imputation approaches are necessary. Whereas the true population manifold that generated data is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

proaches to handling sparse counts include:

- diffusion-based MAGIC [Dijk et al., 2018]
- k-nearest neighbor-based knn-smooth [Wagner et al., 2018b]
- network diffusion-based netSmooth [Jonathan Ronen, 2018]
- clustering-based DrImpute [Gong et al., 2018]
- locality sensitive imputation in LSImpute [Moussa and Măndoiu, 2019]

A major task in the analysis of high-dimensional single-cell data is to find low-dimensional representations of the data that capture the salient biological signals and render the data more interpretable and amenable to further analyses. As it happens, the matrix factorization and latent-space learning methods used for that task also provide another route for imputation through their ability to *reconstruct* the observed data matrix from simplified representations of it. Principal component analysis (PCA) is one such standard matrix factorization method that can be applied to scRNA-seq data (preferably after suitable data normalization) 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 loadings. Many matrix factorization methods with tweaks for single-cell data have been proposed 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]

- pCMF, probabilistic count matrix factorization with a Poisson model [Durif et al., 2019]

- SDA, sparse decomposition of arrays; another sparse Bayesian method [Jung et al., 2019].

Some data reconstruction approaches have been specifically proposed for imputation, including:

- ENHANCE, denoising PCA with an aggregation step [Wagner et al., 2019]

- ALRA, SVD with adaptive thresholding [Linderman et al., 2018]

- scRMD, robust matrix decomposition [Chen et al., 2018]

Recently, machine learning methods have emerged that apply autoencoders [AutoImpute, Talwar et al., 2018] and deep neural networks [DeepImpute, Arisdakessian et al., 2018]) or ensemble learning [EnImpute, Zhang et al., 2019c]) to impute expression values.

Additionally, many deep learning methods have been proposed for single-cell data analysis that can, but need not, use probabilistic data generative processes to capture low-dimensional or latent space representations of a dataset. Even if imputation is not a main focus, such methods can generate "imputed" expression values as an upshot of a model primarily focused on other tasks like learning latent spaces, clustering, batch correction, or visualization (and often several of these tasks simultaneously). The latter set includes tools such as:

- DCA, an autoencoder with a zero-inflated negative binomial distribution [Eraslan et al., 2019]

- scVI, a variational autoencoder with a zero-inflated negative binomial model [Lopez et al., 2018]

- LATE [Badsha et al., 2018]

- VASC [Wang and Gu, 2018]

- compscVAE [Grønbech et al., 2018]

- scScope [Deng et al., 2019]

- Tybalt [Way and Greene, 2018]

- SAUCIE [Amodio et al., 2019]

- scvis [Ding et al., 2018]

- net-SNE [Cho et al., 2018]

- BERMUDA, focused on batch correction [Wang et al., 2019]

- DUSC [Srinivasan et al., 2019]

- Expression Saliency [Kinalis et al., 2019]

- others [Lin et al., 2017a, Zhang, 2019]

Besides the three categories described above, a small number of scRNA-seq imputation methods have been developed to incorporate information external to the current dataset for imputation. These include: ADImpute [Leote et al., 2019], which uses gene regulatory network information from external sources; SAVER-X [Wang et al., 2018], a transfer learning method for denoising and imputation that can use information from atlas-type resources; and methods that borrow information from matched bulk RNA-seq 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 normalization 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]

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27885v3 | CC BY 4.0 Open Access | rec: 23 Aug 2019, publ: 23 Aug 2019

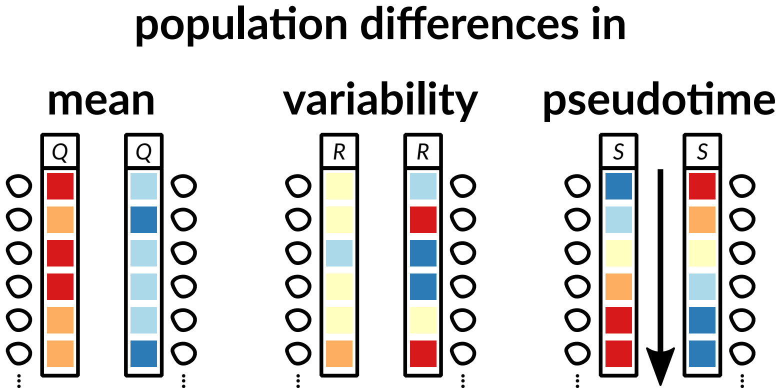

Figure 3: Differential expression of a gene or transcript between cell populations. The top row labels the specific gene or transcript, as is also done in Figure 6. A difference in **mean** gene expression manifests in a consistent difference of gene expression across all cells of a population (e.g. 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.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27885v3 | CC BY 4.0 Open Access | rec: 23 Aug 2019, publ: 23 Aug 2019

## 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 endeavors 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, Guilliams 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 Oudenaarden, 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<br>*Caenorhabditis elegans* | whole organism at larval stage L2 | [Cao et al., 2017] |
| planaria<br>*Schmidtea mediterranea* | whole organism of the adult animal | [Fincher et al., 2018, Plass et al., 2018] |
| fruit fly<br>*Drosophila melanogaster* | whole organism at embryonic stage | [Karaiskos et al., 2017b] |
| Zebrafish | whole organism at embryonic stage | [Farrell et al., 2018, Wagner et al., 2018a] |
| frog<br>*Xenopus tropicalis* | 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 trajectories 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 annotated 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 chromatin accessibility [Buenrostro et al., 2018, Chen et al., 2019, Pliner et al., 2018].

Having defined robust methods to reconstruct trajectories from each data type, another future challenge is related to their comparison or alignment. Here, some ideas from recent methods used to align transcriptomic datasets may be extended [Butler et al., 2018b, Haghverdi et al., 2018, Welch et al., 2018]. A related unsolved problem is that of comparing different trajectories obtained from the same data type but across individuals or conditions to highlight unique and common aspects.

## 3.5 Challenge V: Finding patterns in spatially resolved measurements

Single-cell spatial transcriptomics or proteomics [Crosetto et al., 2015, Strell et al., 2018, Moffitt et al., 2018] technologies can obtain transcript abundance measurements while retaining spatial coordinates of cells or even transcripts within a tissue (this can be seen as an additional feature space to integrate, see Approach 3 in section 6.1, Figure 6 and Table 2). With such data, the question arises of how spatial information can best be leveraged to find patterns, infer cell types or functions and classify cells in a given tissue [Tanay and Regev, 2017].

### 3.5.1 Status

Experimental approaches have been tailored to either systematically extract foci of cells and analyze them with scRNA-seq, or to measure RNA and proteins *in-situ*. Histological sections can be projected in two dimensions while preserving spatial information using sequencing arrays [Ståhl et al., 2016]. Whole tissues can be decomposed using the Niche-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 sequence them with the respective barcodes [Rodriques et al., 2019]. Ultimately, one would like to sequence inside a tissue without dissociating the cells and without compromising on the unbiased nature of scRNA-seq. A preliminary approach has been proposed by Lee et al. [2015] coined FISSEQ (Fluorescent *in-situ* sequencing). Lubeck et al. [2014] have shown a first approach to iteratively apply fluorescence *in-situ* hybridization to measure hundreds of RNA species simultaneously, 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 signal 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 single cells simultaneously while retaining spatial coordinates [Moffitt et al., 2016]. Here, even the subcellular coordinates of each individual transcript are retained. In addition to the methods that provide *in-situ* measurements of RNA, Giesen et al. [2014] and Angelo et al. [2014] use mass cytometry technology to quantify the abundance of proteins while preserving subcellular resolution. Finally, 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 resolution.

For determining cell types, or clustering cells into groups that conduct a common function, 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 correlation 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, Edsgärd 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 non-parametric 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 heterogeneity 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 tumor 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 substantially 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 methods based on polymerase chain reaction (PCR) and multiple displacement amplification (MDA). The PCR-based technologies include degenerate oligonucleotide-primed PCR (DOP-PCR) [Telenius et al., 1992], linker-adapter PCR [Klein et al., 1999], primer extension preamplification 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 Φ29 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. How-

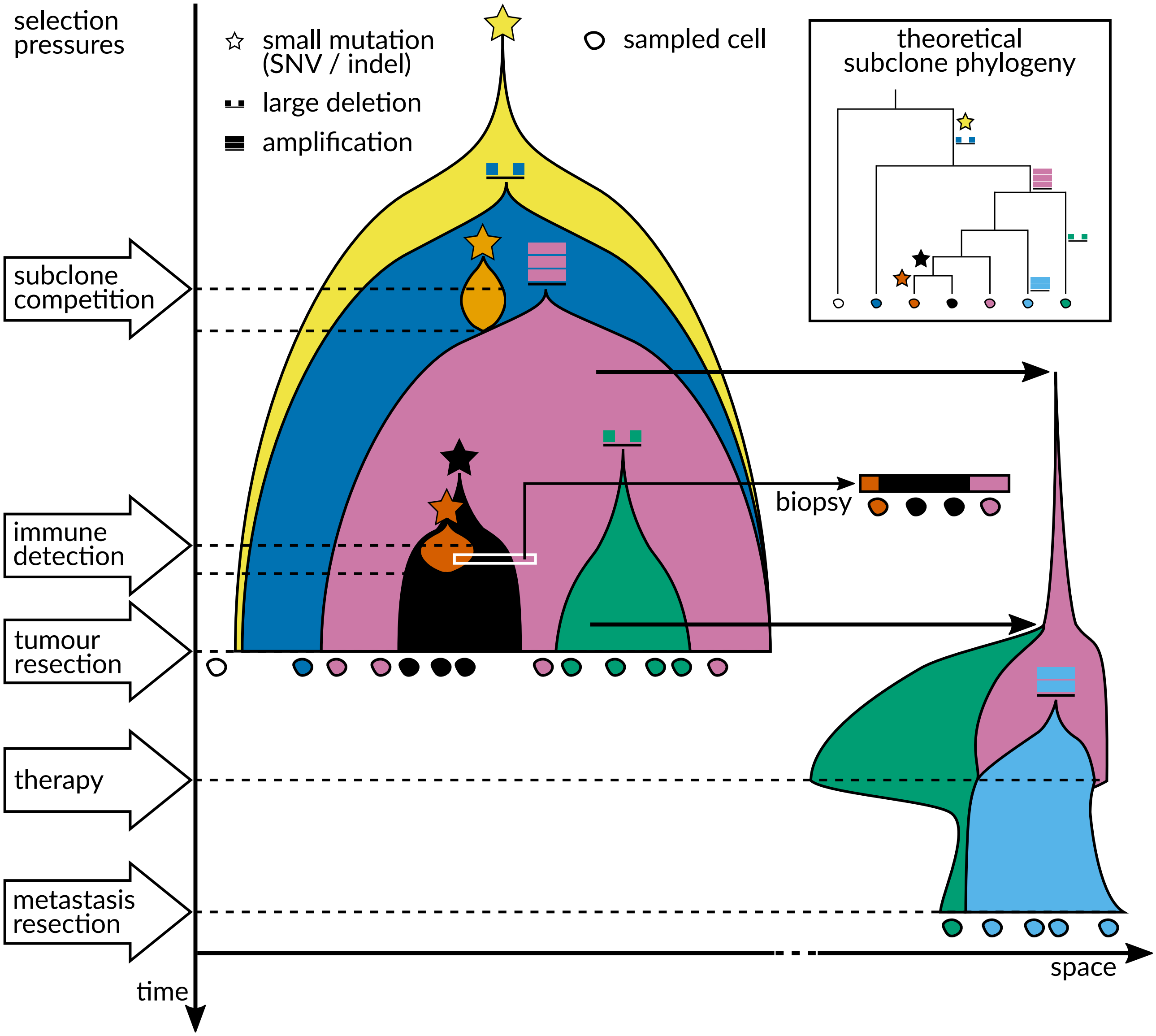

Figure 4: From initiation of a tumor to its detection, resection and possible metastasis, it will evolve somatically. New genomic mutations can confer a selective advantage to the resulting new subclone, that can allow it to outcompete other tumor subclones (subclone competition). At the same time, the 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.

ever, amplification biases and errors present a 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 Φ29 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 Φ29 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 Φ29 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 the identification of all types of genetic variation from SNVs over smaller insertions and deletions up to copy number variation and structural variants.

## 4.2 Challenge VII: Errors and missing data in the identification of features / variation from single-cell DNA sequencing data.

The aim of scDNA sequencing usually is to track somatic evolution at the cellular level, that is, at the finest resolution possible relative to the laws of reproduction (cell division, Figure 5). Examples refer to identifying heterogeneity and tracking evolution in cancer, as the likely most predominant use case (also see below in section 5), but also to monitoring the interaction of somatic mutation with developmental and differentiation processes. To track genetic drifts, selective pressures, or other phenomena inherent to the development of cell clones or types (Figure 4)—but also to stratify cancer patients for the presence of resistant subclones—it is instrumental to genotype and also phase genetic variants in single cells with sufficiently high confidence.

The major disturbing factor in scDNA-seq data is the WGA process (see section 4.1). All methodologies introduce amplification errors (false positive alternative alleles), but more drastic is the effect of amplification bias: the insufficient or complete failure of amplification, which leads to imbalanced proportions or complete lack of variant alleles. Overall, one can distinguish between three cases: (i) an imbalanced proportion of alleles, i.e. loci harboring heterozygous mutations where preferential amplification of one of the two alleles leads to read counts that are distorted, sometimes heavily; (ii) allele drop-out, i.e. loci harboring heterozygous mutations where only one of the alleles was amplified 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 uncertainties will yield massive improvements in geno- and haplotyping (phasing) somatic variants. This, in turn, is necessary to substantially improve the identification of subclonal genotypes and the tracking of evolutionary

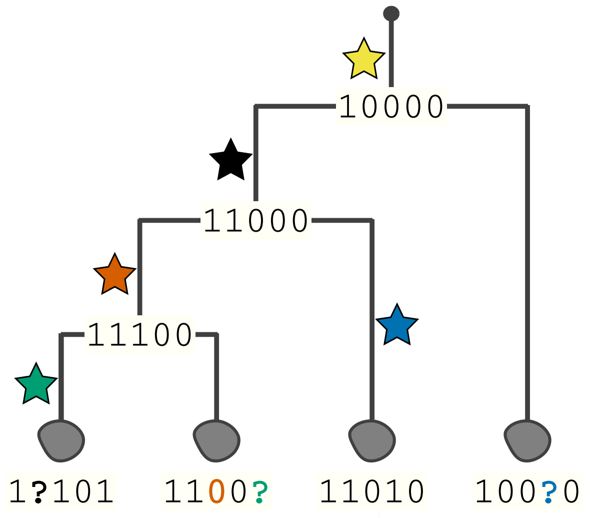

Figure 5: Mutations (colored stars) accumulate in cells during somatic cell divisions and can be used to reconstruct the developmental lineages of individual cells within an organism (leaf nodes of the tree with mutational presence / absence profiles attached). However, insufficient or unbalanced WGA can lead to the dropout of one or both alleles 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.

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 neighboring germline single-nucleotide polymorphisms. It exploits the fact that allele drop-out 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 unknown 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 SciΦ [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 Aneufinder [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 modeling 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 settings [Köster et al., 2019] 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 formalization 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 tumor 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 ancestors. 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 tumor 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 tumor 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 characterizing 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 tumor 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 destabilization 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 Markowetz, 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 SciΦ [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 modeling 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 modeling 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 tumor and metastases, and across different time points [Yang and Lin, 2017]. This imposes different selective pressures on different tumor cells, driving the formation of tumor 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 [Turajlic 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 tumor, with that subclone seeding multiple locations with a genotype closer to the late-stage primary tumor; 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 characterized 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 analyzed for copy number variation [Casasent et al., 2018]. For cell and tissue characterization 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, characterizing 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 tumors, lymph nodes and metastases upon surgery and tumors after relapse. These time-points already lend themselves 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 tumor subclones as populations of single cells, capturing some of the population structure seen in cancers. This phylogenetic context also lends itself to modeling differences in mutational rates and signatures between different cell populations, e.g. between normal somatic evolution before tumor initiation and cancer evolution after tumor 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 scDNA-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.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27885v3 | CC BY 4.0 Open Access | rec: 23 Aug 2019, publ: 23 Aug 2019

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 metastases are founded recurrently by the same subclone(s) and whether individual metastases are founded by individual or multiple subclones. A number of studies utilizing multiple region samples from the same tumor and from distant metastases already paved the way in investigating these questions [Turajlic and Swanton, 2016]. Still, only single-cell spatial resolution will allow identification of specific individual genotypes in specific locations and the drawing precise conclusions.

The second challenge will be to determine rates of proliferation and death per subclone. This could be achieved by measuring numbers of mitotic and apoptotic cells per subclone or by integrating subclone abundance profiles across time points. Good estimates of these basic parameters will greatly benefit models, e.g. for the detection of positive and negative selection in cancer.

A third challenge will be to determine subclone-specific rates of mutation. Here, integration of models from population genetics and phylogenetics holds promise.

A fourth challenge will be to devise ways to determine further relevant model parameters. For example, comparing expanded subclones in drug screens to determine subclone fitness under different treatment regimes can both help to predict subclone resistance (and thus expected treatment success) and further inform cancer evolution models.

A final step will then be to put all these parameters into context with further information about local microenvironments (such as vascular invasion and immune cell infiltration), to estimate the selection potential of such local factors for or against different subclones.

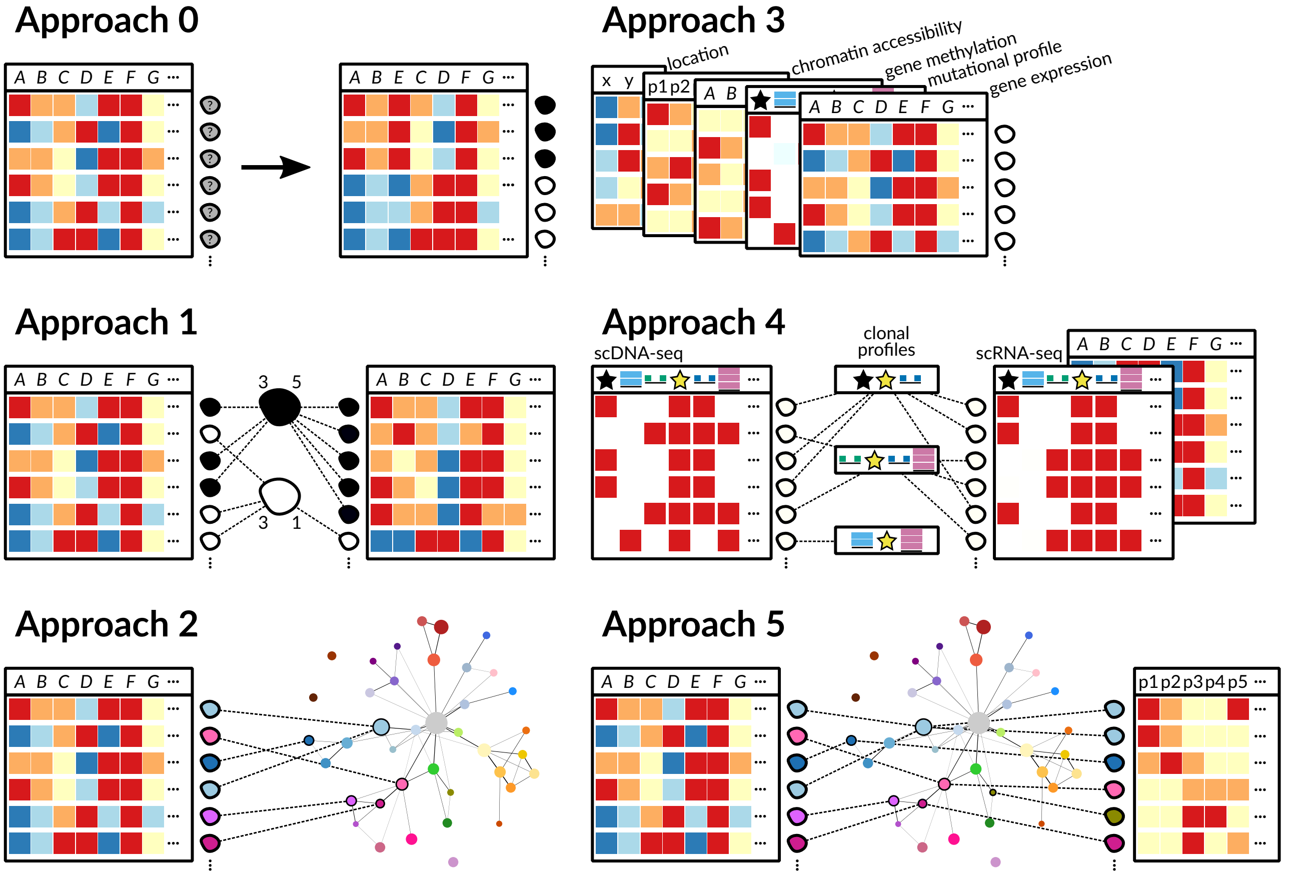

Figure 6: Approaches for integrating single-cell measurement datasets across measurement types, samples and experiments, as also described in Table 2.
**Approach 0** Clustering of cells from one sample from one experiment, no data integration is needed. **Approach 1** Cell populations / clusters from 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 observations | technical noise; batch effects; validate cell type assignments; standards across experimental centers |
| 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 experimental 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 combination | hypothetical: multi-omic HCA, single-cell TCGA | comprehensive characterizations of biological systems | all from approaches 2, 3 & 4; standards across experimental centers |

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

# 6 Overarching challenges

## 6.1 Challenge XI: Integration of single-cell data: across samples, experiments and types of measurement

Biological processes are complex and dynamic, varying across cells and organisms. To comprehensively analyze such processes, different types of measurements from multiple experiments need to be obtained and integrated. Depending on the actual research question, such experiments will refer to different time points, tissues or organisms. For different measurement types, we put particular emphasis on the combination of scRNA-seq and scDNA-seq data, although augmenting sequencing data with records on protein or metabolite levels is also possible.

Since the exploration of complex, dynamic and variable processes requires the integration of data from multiple experiments, we need flexible but rigorous statistical and computational frameworks to support that integration. See Table 2 and Figure 6 for an overview of how the issues in creating such frameworks can vary relative to the particular problem[2].

When aiming at the identification of patterns of differential expression, so as to characterize variability across organisms, individuals, or location, data refers to the same (unique) measurement type (for example, only scRNA-seq), but stems from different time points, different locations (such as different tissues or sites in a tumor), or different organisms. See Approach 1 in Figure 6 and Table 2 for methodological challenges arising 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

[2] Graph 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/)

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 centers 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.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27885v3 | CC BY 4.0 Open Access | rec: 23 Aug 2019, publ: 23 Aug 2019

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, Boufea 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, Argelaguet 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 genomics [Zheng et al., 2017] for scRNA-seq and direct library preparation (DLP, Zahn et al. [2017b]) for scDNA-seq establish an experimental basis. As above-mentioned, the greater analytical challenge is to, upon having performed experiments, identify subpopulations that had hitherto remained invisible, 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 identical 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-number 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 expression patterns. (iii) MATCHER [Welch et al., 2017] uses manifold alignment to combine scM&T-seq [Angermueller et al., 2016] with sc-GEM [Cheow et al., 2016], leveraging the common set of loci. All of these methods are based on biologically coherent assumptions on how to summarize measurements across different types and samples in a reasonable way, despite their different physical origin.

### 6.1.2 Open problems

Experimental technologies that deal with taking measurements of different kinds on one single cell (Approach 3 in Figure 6 and Table 2) are on the rise and will allow to assay more cells at higher fidelity and reduced cost. Yet, however, many methods for evaluating combinations of different types of measurements performed on one single cell have not been in the focus. It is to be expected that the corresponding open problems will become more urgent. As an example, consider combined measurements of scDNA-seq and scRNA-seq, where one uses the transcripts derived 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 corresponding 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 fine-grained 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 suffer 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

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27885v3 | CC BY 4.0 Open Access | rec: 23 Aug 2019, publ: 23 Aug 2019

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 comparison 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-

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27885v3 | CC BY 4.0 Open Access | rec: 23 Aug 2019, publ: 23 Aug 2019

lenge. This approach has been very successful, e.g. in protein structure prediction[3] and metagenomic analyses[4]. A first step in this direction was the recent single-cell transcriptomics DREAM challenge[5].

# 7 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

**ICA** independent component analysis. 10

**MALBAC** multiple annealing and looping-based amplification cycles. 20, 22

**MDA** multiple displacement amplification. 20, 22

**MSA** multiple sequence alignment. 25, 26, 30

**NMF** non-negative matrix factorization. 10

**PCA** principal component analysis. 10, 11

**PCR** polymerase chain reaction. 20, 22, 39

**sc-seq** single-cell sequencing. 3–6, 29, 37, 39, 40

**scDNA-seq** single-cell DNA sequencing. 3, 6–8, 16, 20–25, 27–31, 34–38, 40

**SCDS** Single-Cell Data Science. 3, 4, 7

**scRNA-seq** single-cell RNA sequencing. 3, 5–14, 16–18, 22, 24, 34–38, 40

**SNV** single nucleotide variation. 6, 20, 22–28, 38, 40

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

# References

Andre J Aberer, Kassian Kobert, and Alexandros Stamatakis. ExaBayes: massively parallel bayesian tree inference for the whole-genome era. *Mol. Biol. Evol.*, 31(10):2553–2556, October 2014.

Ahmet Acar, Daniel Nichol, Javier Fernandez-Mateos, George D. Cresswell, Iros Barozzi, Sung Pil Hong, Inmaculada Spiteri, Mark Stubbs, Rosemary Burke, Adam Stewart, Georgios Vlachogiannis, Carlo C. Maley, Luca Magnani, Nicola Valeri, Udai Banerji, and

[3] http://predictioncenter.org/
[4] https://data.cami-challenge.org
[5] https://www.synapse.org/#!Synapse:syn15665609/wiki/582909

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27885v3 | CC BY 4.0 Open Access | rec: 23 Aug 2019, publ: 23 Aug 2019

Andrea Sottoriva. Exploiting evolutionary herding to control drug resistance in cancer. *bioRxiv*, page 566950, March 2019. doi: 10.1101/566950. URL `https://www.biorxiv.org/content/10.1101/566950v1`.

Sumon Ahmed, Magnus Rattray, and Alexis Boukouvalas. GrandPrix: scaling up the Bayesian GPLVM for single-cell data. *Bioinformatics*, 35(1):47–54, January 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty533. URL `https://academic.oup.com/bioinformatics/article/35/1/47/5047752`.

Philipp M Altrock, Lin L Liu, and Franziska Michor. The mathematics of cancer: integrating quantitative models. *Nat. Rev. Cancer*, 15(12):730–745, December 2015.

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

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

Tallulah S. Andrews and Martin Hemberg. False signals induced by single-cell imputation. *F1000Research*, 7:1740, March 2019. ISSN 2046-1402. doi: 10.12688/f1000research.16613.2. URL `https://f1000research.com/articles/7-1740/v2`.

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

Christof Angermueller, Stephen J. Clark, Heather J. Lee, Iain C. Macaulay, Mabel J. Teng, Tim Xiaoming Hu, Felix Krueger, Sebastien Smallwood, Chris P. Ponting, Thierry Voet, Gavin Kelsey, Oliver Stegle, and Wolf Reik. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nature Methods*, 13(3): 229–232, March 2016. ISSN 1548-7105. doi: 10.1038/nmeth.3728.

Christof Angermueller, Heather J. Lee, Wolf Reik, and Oliver Stegle. Deep-CpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biology*, 18(1):67, April 2017. ISSN 1474-760X. doi: 10.1186/s13059-017-1189-z. URL `https://doi.org/10.1186/s13059-017-1189-z`.

Ricard Argelaguet, Britta Velten, Damien Arnol, Sascha Dietrich, Thorsten Zenz, John C. Marioni, Florian Buettner, Wolfgang Huber, and Oliver Stegle. Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*, 14(6):e8124, June 2018. ISSN 1744-4292, 1744-4292. doi: 10.15252/msb.

20178124. URL http://msb.embopress.org/content/14/6/e8124.

C Arisdakessian, O Poirion, B Yunits, X Zhu, and L Garmire. DeepImpute: an accurate, fast and scalable deep neural network method to impute single-cell RNA-Seq data. *bioRxiv*, 2018. URL https://www.biorxiv.org/content/10.1101/353607v1.abstract.

Nona Arneson, Simon Hughes, Richard Houlston, and Susan Done. Whole-Genome amplification by improved primer extension preamplification PCR (I-PEP-PCR). *CSH Protoc.*, 2008:db.prot4921, January 2008.

Damien Arnol, Denis Schapiro, Bernd Bodenmiller, Julio Saez-Rodriguez, and Oliver Stegle. Modelling cell-cell interactions from spatial molecular data with spatial variance component analysis, 2018.

Daniel L. Ayres. *Research And Application Of Parallel Computing Algorithms For Statistical Phylogenetic Inference.* PhD thesis, University of Maryland, 2017. URL http://drum.lib.umd.edu/handle/1903/19951.

Elham Azizi, Sandhya Prabhakaran, Ambrose Carr, and Dana Pe'er. Bayesian Inference for Single-cell Clustering and Imputing. *Genomics and Computational Biology*, 3(1):46, January 2017. ISSN 2365-7154. doi: 10.18547/gcb.2017.vol3.iss1.e46. URL https://genomicscomputbiol.org/ojs/index.php/GCB/article/view/46.

Rhonda Bacher and Christina Kendziorski. Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biology*, 17(1):63, April 2016. ISSN 1474-760X. doi: 10.1186/s13059-016-0927-y. URL https://doi.org/10.1186/s13059-016-0927-y.

Md Bahadur Badsha, Rui Li, Boxiang Liu, Yang I Li, Min Xian, Nicholas E Banovich, and Audrey Qiuyan Fu. Imputation of single-cell gene expression with an autoencoder neural network. December 2018.

Bjorn Bakker, Aaron Taudt, Mirjam E Belderbos, David Porubsky, Diana C J Spierings, Tristan V de Jong, Nancy Halsema, Hinke G Kazemier, Karina Hoekstra-Wakker, Allan Bradley, Eveline S J M de Bont, Anke van den Berg, Victor Guryev, Peter M Lansdorp, Maria Colomé-Tatché, and Floris Foijer. Single-cell sequencing reveals karyotype heterogeneity in murine and human malignancies. *Genome Biol.*, 17(1):115, May 2016.

Nikolas Barkas, Viktor Petukhov, Daria Nikolaeva, Yaroslav Lozinsky, Samuel Demharter, Konstantin Khodosevich, and Peter V. Kharchenko. Wiring together large single-cell RNA-seq sample collections. *bioRxiv*, page 460246, November 2018. doi: 10.1101/460246. URL https://www.biorxiv.org/content/10.1101/460246v1.

Nico Battich, Thomas Stoeger, and Lucas Pelkmans. Control of transcript variability in single mammalian cells. *Cell*, 163(7):1596–1610, December 2015.

Benedikt Bauer, Reiner Siebert, and Arne Traulsen. Cancer initiation with epistatic interactions between driver and passenger mutations. *J. Theor. Biol.*, 358:52–60, October 2014.

Niko Beerenwinkel, Tibor Antal, David Dingli, Arne Traulsen, Kenneth W Kinzler, Victor E Velculescu, Bert Vogelstein, and Martin A Nowak. Genetic progression and the waiting time to cancer. *PLoS Comput. Biol.*, 3(11):e225, November 2007.

Graham R. Bignell, Thomas Santarius, Jessica C.M. Pole, Adam P. Butler, Janet Perry, Erin Pleasance, Chris Greenman, Andrew Menzies, Sheila Taylor, Sarah Edkins, Peter Campbell, Michael Quail, Bob Plumb, Lucy Matthews, Kirsten McLay, Paul A.W. Edwards, Jane Rogers, Richard Wooster, P. Andrew Futreal, and Michael R. Stratton. Architectures of somatic genomic rearrangement in human cancer amplicons at sequence-level resolution. *Genome Research*, 17 (9):1296–1303, 2007. doi: 10.1101/gr.6522707. URL http://genome.cshlp.org/content/17/9/1296.abstract.

Konstantin A Blagodatskikh, Vladimir M Kramarov, Ekaterina V Barsova, Alexey V Garkovenko, Dmitriy S Shcherbo, Andrew A Shelenkov, Vera V Ustinova, Maria R Tokarenko, Simon C Baker, Tatiana V Kramarova, and Konstantin B Ignatov. Improved DOP-PCR (iDOP-PCR): A robust and simple WGA method for efficient amplification of low copy number genomic DNA. *PLoS One*, 12(9):e0184507, September 2017.

L Blanco, A Bernad, J M Lázaro, G Martín, C Garmendia, and M Salas. Highly efficient 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. Bohrson, Alison R. Barton, Michael A. Lodato, Rachel E. Rodin, Lovelace J. Luquette, Vinay V. Viswanadham, Doga C. Gulhan, Isidro Cortés-Ciriano, Maxwell A. Sherman, Minseok 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/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 discriminant analysis. *bioRxiv*, page 470203, January 2019. doi: 10.1101/470203. URL https://www.biorxiv.org/content/10.1101/470203v2.

Ivana Bozic, Tibor Antal, Hisashi Ohtsuki, Hannah Carter, Dewey Kim, Sining Chen, Rachel Karchin, Kenneth W Kinzler, Bert Vogelstein, and Martin A Nowak. Accumulation of driver and passenger mutations during tumor progression. *Proc. Natl. Acad. Sci. U. S. A.*, 107(43):18545–18550, October 2010.

Ivana Bozic, Jeffrey M Gerold, and Martin A Nowak. Quantifying clonal and subclonal passenger mutations in cancer evolution. *PLoS Comput. Biol.*, 12(2):e1004731, February 2016.

James A. Briggs, Caleb Weinreb, Daniel E. Wagner, Sean Megason, Leonid Peshkin, Marc W. Kirschner, and Allon M. Klein. The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science*, 360(6392):eaar5780, June 2018. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aar5780. URL http://science.sciencemag.org/content/360/6392/eaar5780.

Jane Bromley, James W. Bentz, Léon Bottou, Isabelle Guyon, Yann Lecun, Cliff

Moore, Eduard Säckinger, and Roopak Shah. Signature verification using a “siamese” time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence*, 07(04): 669–688, August 1993. ISSN 0218-0014. doi: 10.1142/S0218001493000339. URL https://www.worldscientific.com/doi/10.1142/S0218001493000339.

Robert V Bruggner, Bernd Bodenmiller, David L Dill, Robert J Tibshirani, and Garry P Nolan. Automated identification of stratifying signatures in cellular subpopulations. *Proc. Natl. Acad. Sci. U. S. A.*, 111(26):E2770–7, July 2014.

Jason D. Buenrostro, Beijing Wu, Ulrike M. Litzenburger, Dave Ruff, Michael L. Gonzales, Michael P. Snyder, Howard Y. Chang, and William J. Greenleaf. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523 (7561):486–490, July 2015. ISSN 1476-4687. doi: 10.1038/nature14590. URL https://www.nature.com/articles/nature14590.

Jason D. Buenrostro, M. Ryan Corces, Caleb A. Lareau, Beijing Wu, Alicia N. Schep, Martin J. Aryee, Ravindra Majeti, Howard Y. Chang, and William J. Greenleaf. Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation. *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 versatile factor analysis for single-cell RNA-seq. *Genome biology*, 18(1):212, November 2017. ISSN 1465-6906. doi: 10.1186/s13059-017-1334-8. URL http://dx.doi.org/10.1186/s13059-017-1334-8.

Andrew Butler, Paul Hoffman, Peter Smibert, 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 Smibert, 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://www.nature.com/articles/nbt.4096.

Kieran R. Campbell and Christopher Yau. Order Under Uncertainty: Robust Differential Expression Analysis Using Probabilistic Models for Pseudotime Inference. *PLOS Computational Biology*, 12 (11):e1005212, November 2016. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1005212. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005212.

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

Kieran R. Campbell, Adi Steif, Emma Laks, Hans Zahn, Daniel Lai, Andrew McPherson, Hossein Farahani, Farhia Kabeer, Ciara O’Flanagan, Justina Biele, Jazmine Brimhall, Beixi Wang, Pascale Walters, IMAXT Consortium, Alexandre Bouchard-Côté, Samuel Aparicio, and Sohrab P.

Shah. clonealign: statistical integration of independent single-cell RNA and DNA sequencing data from human cancers. *Genome Biology*, 20(1):54, March 2019. ISSN 1474-760X. doi: 10.1186/s13059-019-1645-z. URL `https://doi.org/10.1186/s13059-019-1645-z`.

Junyue Cao, Jonathan S. Packer, Vijay Ramani, Darren A. Cusanovich, Chau Huynh, Riza Daza, Xiaojie Qiu, Choli Lee, Scott N. Furlan, Frank J. Steemers, Andrew Adey, Robert H. Waterston, Cole Trapnell, and Jay Shendure. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*, 357(6352): 661–667, August 2017. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aam8940. URL `http://science.sciencemag.org/content/357/6352/661`.

Junyue Cao, Darren A. Cusanovich, Vijay Ramani, Delasa Aghamirzaie, Hannah A. Pliner, Andrew J. Hill, Riza M. Daza, Jose L. McFaline-Figueroa, Jonathan S. Packer, Lena Christiansen, Frank J. Steemers, Andrew C. Adey, Cole Trapnell, and Jay Shendure. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science*, 361(6409):1380–1385, September 2018. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aau0730. URL `https://science.sciencemag.org/content/361/6409/1380`.

Junyue Cao, Malte Spielmann, Xiaojie Qiu, Xingfan Huang, Daniel M. Ibrahim, Andrew J. Hill, Fan Zhang, Stefan Mundlos, Lena Christiansen, Frank J. Steemers, Cole Trapnell, and Jay Shendure. The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, 566(7745): 496, February 2019a. ISSN 1476-4687. doi: 10.1038/s41586-019-0969-x. URL `https://www.nature.com/articles/s41586-019-0969-x`.

Zhi-Jie Cao, Lin Wei, Shen Lu, De-Chang Yang, and Ge Gao. Cell BLAST: Searching large-scale scRNA-seq database via unbiased cell embedding. *bioRxiv*, page 587360, March 2019b. doi: 10.1101/587360. URL `https://www.biorxiv.org/content/10.1101/587360v1`.

Anna K Casasent, Aislyn Schalck, Ruli Gao, Emi Sei, Annalyssa Long, William Pangburn, Tod Casasent, Funda Meric-Bernstam, Mary E Edgerton, and Nicholas E Navin. Multiclonal invasion in breast tumors identified by topographic single cell sequencing. *Cell*, 172(1-2): 205–217.e12, January 2018.

Chong Chen, Changjing Wu, Linjie Wu, Yishu Wang, Minghua Deng, and Ruibin Xi. scRMD: Imputation for single cell RNA-seq data via robust matrix decomposition. November 2018. URL `https://www.biorxiv.org/content/10.1101/459404v2`.

Chongyi Chen, Dong Xing, Longzhi Tan, Heng Li, Guangyu Zhou, Lei Huang, and X Sunney Xie. Single-cell whole-genome analyses by linear amplification via transposon insertion (LIANTI). *Science*, 356 (6334):189–194, April 2017.

Huidong Chen, Luca Albergante, Jonathan Y. Hsu, Caleb A. Lareau, Giosuè Lo Bosco, Jihong Guan, Shuigeng Zhou, Alexander N. Gorban, Daniel E. Bauer, Martin J. Aryee, David M. Langenau, Andrei Zinovyev, Jason D. Buenrostro, Guo-Cheng Yuan, and Luca Pinello. Single-cell trajectories reconstruction, exploration and mapping of omics data with STREAM. *Nature Communications*, 10

(1):1903, April 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-09670-4. URL https://www.nature.com/articles/s41467-019-09670-4.

Kok Hao Chen, Alistair N Boettiger, Jeffrey R Moffitt, Siyuan Wang, and Xiaowei Zhuang. RNA imaging. spatially resolved, highly multiplexed RNA profiling in single cells. *Science*, 348(6233):aaa6090, April 2015.

Mengjie Chen and Xiang Zhou. VIPER: variability-preserving imputation for accurate gene expression recovery in single-cell RNA sequencing studies. *Genome Biol.*, 19 (1):196, November 2018.

Lih Feng Cheow, Elise T. Courtois, Yuliana Tan, Ramya Viswanathan, Qiaorui Xing, Rui Zhen Tan, Daniel S. W. Tan, Paul Robson, Yuin-Han Loh, Stephen R. Quake, and William F. Burkholder. Single-cell multimodal profiling reveals cellular epigenetic heterogeneity. *Nature Methods*, 13(10):833–836, October 2016. ISSN 1548-7105. doi: 10.1038/nmeth.3961. URL https://www.nature.com/articles/nmeth.3961.

Cariad Chester and Holden T Maecker. Algorithmic 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 visualization of Single-Cell data using neural networks. *Cell Syst*, 7(2):185–191.e4, August 2018.

Simone Ciccolella, Mauricio Soto Gomez, Murray Patterson, Gianluca Della Vedova, Iman Hajirasouliha, and Paola Bonizzoni. Inferring Cancer Progression from Single-cell Sequencing while Allowing Mutation Losses. *bioRxiv*, page 268243, April 2018. doi: 10.1101/268243. URL https://www.biorxiv.org/content/10.1101/268243v2.

Stephen J. Clark, Ricard Argelaguet, Chantriolnt-Andreas Kapourani, Thomas M. Stubbs, Heather J. Lee, Celia Alda-Catalinas, Felix Krueger, Guido Sanguinetti, Gavin Kelsey, John C. Marioni, Oliver Stegle, and Wolf Reik. scNMT-seq enables joint profiling of chromatin accessibility 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/s41467-018-03149-4.

Germán Corredor, Xiangxue Wang, Yu Zhou, Cheng Lu, Pingfu Fu, Konstantinos Syrigos, David L Rimm, Michael Yang, Eduardo Romero, Kurt A Schalper, Vamsidhar Velcheti, and Anant Madabhushi. Spatial architecture and arrangement of Tumor-Infiltrating lymphocytes for predicting likelihood of recurrence in Early-Stage Non-Small cell lung cancer. *Clin. Cancer Res.*, September 2018.

Alexandra Cretu and Peter C Brooks. Impact of the non-cellular tumor microenvironment on metastasis: potential therapeutic and imaging opportunities. *J. Cell. Physiol.*, 213(2):391–402, November 2007.

Nicola Crosetto, Magda Bienko, and Alexander van Oudenaarden. Spatially resolved transcriptomics and beyond. *Nat. Rev. Genet.*, 16(1):57–66, January 2015.

Darren A. Cusanovich, Riza Daza, Andrew Adey, Hannah A. Pliner, Lena Christiansen, Kevin L. Gunderson, Frank J. Steemers, Cole Trapnell, and Jay Shendure. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science (New York, N.Y.)*, 348(6237):910–914, May 2015. ISSN 1095-9203. doi: 10.1126/science.aab1601.

Darren A. Cusanovich, James P. Reddington, David A. Garfield, Riza M. Daza, Delasa Aghamirzaie, Raquel Marco-Ferreres, Hannah A. Pliner, Lena Christiansen, Xiaojie Qiu, Frank J. Steemers, Cole Trapnell, Jay Shendure, and Eileen E. M. Furlong. The cis-regulatory dynamics of embryonic development at single-cell resolution. *Nature*, 555(7697):538–542, March 2018. ISSN 1476-4687. doi: 10.1038/nature25981. URL `https://www.nature.com/articles/nature25981`.

Sayantan Das, Gonçalo R Abecasis, and Brian L Browning. Genotype Imputation from Large Reference Panels. *Annual review of genomics and human 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/annurev-genom-083117-021602`.

Soma Datta, Lavina Malhotra, Ryan Dickerson, Scott Chaffee, Chandan K Sen, and Sashwati Roy. Laser capture microdissection: Big data from small samples. *Histol. Histopathol.*, 30(11):1255–1269, November 2015.

Alexander Davis, Ruli Gao, and Nicholas Navin. Tumor evolution: Linear, branching, neutral or punctuated? *Biochim. Biophys. Acta*, 1867(2):151–161, April 2017.

Carl G. de Boer and Aviv Regev. BROCKMAN: deciphering variance in epigenomic regulators by k-mer factorization. *BMC Bioinformatics*, 19(1):253, July 2018. ISSN 1471-2105. doi: 10.1186/s12859-018-2255-6. URL `https://doi.org/10.1186/s12859-018-2255-6`.

Charles F A de Bourcy, Iwijn De Vlaminck, Jad N Kanbar, Jianbin Wang, Charles Gawad, and Stephen R Quake. A quantitative 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, Wanmin Song, Stephen F Kingsmore, Michael Egholm, and Roger S Lasken. Comprehensive human genome amplification using multiple displacement amplification. *Proc. Natl. Acad. Sci. U. S. A.*, 99(8):5261–5266, April 2002.

Yue Deng, Feng Bao, Qionghai Dai, Lani F Wu, and Steven J Altschuler. Scalable analysis of cell-type composition from single-cell transcriptomics using deep recurrent learning. *Nature methods*, March 2019. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-019-0353-7. URL `https://doi.org/10.1038/s41592-019-0353-7`.

Erica AK DePasquale, Kyle Ferchen, Stuart Hay, H. Leighton Grimes, and Nathan Salomonis. cellHarmony: Cell-level matching and comparison of single-cell transcriptomes. *bioRxiv*, page 412080, January 2019. doi: 10.1101/412080. URL `https://www.biorxiv.org/content/10.1101/412080v4`.

Siddharth S. Dey, Lennart Kester, Bastiaan Spanjaard, Magda Bienko, and Alexander van Oudenaarden. Integrated genome and transcriptome sequencing of the same cell. *Nature Biotechnology*, 33(3):285–289, March 2015. ISSN 1546-1696. doi: 10.1038/

nbt.3129. URL https://www.nature.com/articles/nbt.3129.

David van Dijk, Roshan Sharma, Juozas Nainys, Kristina Yim, Pooja Kathail, Ambrose J. Carr, Cassandra Burdziak, Kevin R. Moon, Christine L. Chaffer, Diwakar Pattabiraman, Brian Bierie, Linas Mazutis, Guy Wolf, Smita Krishnaswamy, and Dana Pe'er. Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell*, 174(3):716–729.e27, July 2018. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2018.05.061. URL https://www.cell.com/cell/abstract/S0092-8674(18)30724-4.

Jiarui Ding, Anne Condon, and Sohrab P Shah. Interpretable dimensionality reduction 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 identification of single-nucleotide variants in whole-genome-amplified single cells. *Nature Methods*, 14(5):491–493, May 2017. ISSN 1548-7105. doi: 10.1038/nmeth.4227. URL https://www.nature.com/articles/nmeth.4227.

Angelo Duò, Mark D Robinson, and Charlotte 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. *Bioinformatics*, March 2019. ISSN 1367-4803, 1367-4811. doi: 10.1093/bioinformatics/btz177. URL http://dx.doi.org/10.1093/bioinformatics/btz177.

Daniel Edsgärd, Per Johnsson, and Rickard Sandberg. Identification of spatial expression 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 Farahani, 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 Bruna, Jose L. Sandoval, Teresa Algara, Wendy Greenwood, Kaston Leung, Hongwei Cheng, Hui Xue, Yuzhuo Wang, Dong Lin, Andrew J. Mungall, Richard Moore, Yongjun Zhao, Julie Lorette, Long Nguyen, David Huntsman, Connie J. Eaves, Carl Hansen, Marco A. Marra, Carlos Caldas, Sohrab P. Shah, and Samuel Aparicio. Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. *Nature*, 518(7539): 422–426, February 2015. ISSN 0028-0836. doi: 10.1038/nature13952. URL http://www.nature.com/nature/journal/v518/n7539/full/nature13952.html.

Mohammed El-Kebir. SPhyR: tumor phylogeny estimation from single-cell sequencing data under loss and error. *Bioinformatics*, 34(17):i671–i679, September 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty589. URL https://academic.oup.com/bioinformatics/article/34/17/i671/5093218.

Nils Eling, Arianne C. Richard, Sylvia Richardson, John C. Marioni, and Catalina A. Vallejos. Correcting the Mean-Variance Dependency for Differential Variability Testing Using Single-Cell RNA Sequencing Data. *Cell Systems*, 7(3): 284–294.e12, September 2018. ISSN 2405-4712. doi: 10.1016/j.cels.2018.06.011. URL `https://www.cell.com/cell-systems/abstract/S2405-4712(18)30278-3`.

Chee-Huat Linus Eng, Michael Lawson, Qian Zhu, Ruben Dries, Noushin Koulena, Yodai Takei, Jina Yun, Christopher Cronin, Christoph Karp, Guo-Cheng Yuan, and Long Cai. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature*, 568(7751): 235, April 2019. ISSN 1476-4687. doi: 10.1038/s41586-019-1049-y. URL `https://www.nature.com/articles/s41586-019-1049-y`.

Gökcen Eraslan, Lukas M. Simon, Maria Mircea, Nikola S. Mueller, and Fabian J. Theis. Single-cell RNA-seq denoising 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/s41467-018-07931-2`.

Nuria Estévez-Gómez, Tamara Prieto, Amy Guillaumet-Adkins, Holger Heyn, Sonia Prado-López, and David Posada. Comparison of single-cell whole-genome amplification strategies. *bioRxiv*, page 443754, October 2018. doi: 10.1101/443754. URL `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 transcriptional 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, Samantha J. Riesenfeld, Karthik Shekhar, Aviv Regev, and Alexander F. Schier. Single-cell reconstruction of developmental 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/content/360/6392/eaar3131`.

Joseph Felsenstein. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.*, 17(6):368–376, 1981.

Greg Finak, Andrew McDavid, Masanao Yajima, Jingyuan Deng, Vivian Gersuk, Alex K. Shalek, Chloe K. Slichter, Hannah W. Miller, M. Juliana McElrath, Martin Prlic, Peter S. Linsley, and Raphael Gottardo. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology*, 16, 2015. ISSN 1474-7596. doi: 10.1186/s13059-015-0844-5. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4676162/`.

Christopher T. Fincher, Omri Wurtzel, Thom de Hoog, Kellie M. Kravarik, and Peter W. Reddien. Cell type transcriptome atlas for the planarian Schmidtea mediterranea. *Science*, 360(6391):eaaq1736, May 2018. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aaq1736.

URL http://science.sciencemag.org/content/360/6391/eaaq1736.

William Fletcher and Ziheng Yang. The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. *Mol. Biol. Evol.*, 27(10):2257–2267, October 2010.

Jasmine Foo, Kevin Leder, and Franziska Michor. Stochastic dynamics of cancer initiation. *Phys. Biol.*, 8(1):015002, February 2011.

Joshua M. Francis, Cheng-Zhong Zhang, Cecile L. Maire, Joonil Jung, Veronica E. Manzo, Viktor A. Adalsteinsson, Heather Homer, Sam Haidar, Brendan Blumenstiel, Chandra Sekhar Pedamallu, Azra H. Ligon, J. Christopher Love, Matthew Meyerson, and Keith L. Ligon. EGFR Variant Heterogeneity in Glioblastoma Resolved through Single-Nucleus Sequencing. *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.

Saskia Freytag, Luyi Tian, Ingrid Lönnstedt, 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/.

Wolf H Fridman, Jérome 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, Heidelberg, 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 manipulation and biochemical analysis on microfluidics. *Analyst*, 144(3):766–781, January 2019.

Xin Gao, Deqing Hu, Madelaine Gogol, and Hua Li. ClusterMap: Comparing analyses across multiple Single Cell RNA-Seq profiles. *bioRxiv*, page 331330, June 2018. doi: 10.1101/331330. URL https://www.biorxiv.org/content/10.1101/331330v2.

Tyler Garvin, Robert Aboukhalil, Jude Kendall, Timour Baslan, Gurinder S Atwal, James Hicks, Michael Wigler, and Michael C Schatz. Interactive analysis and assessment of single-cell copy-number variations. *Nat. Methods*, 12(11):1058–1060, November 2015.

Charles Gawad, Winston Koh, and Stephen R Quake. Single-cell genome sequencing: current state of the science. *Nat. Rev. Genet.*, 17(3):175–188, March 2016.

Farzad Ghaznavi, Andrew Evans, Anant Madabhushi, and Michael Feldman. Digital imaging in pathology: whole-slide imaging and beyond. *Annu. Rev. Pathol.*, 8:331–359, January 2013.

Charlotte Giesen, Hao A. O. Wang, Denis Schapiro, Nevena Zivanovic, Andrea Jacobs, Bodo Hattendorf, Peter J. Schüffler, Daniel Grolimund, Joachim M. Buhmann, Simone Brandt, Zsuzsanna Varga, Peter J. Wild, Detlef Günther, and Bernd Bodenmiller. Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nature Methods*, 11(4): 417–422, April 2014. ISSN 1548-7105. doi: 10.1038/nmeth.2869. URL `https://www.nature.com/articles/nmeth.2869`.

Yury Goltsev, Nikolay Samusik, Julia Kennedy-Darling, Salil Bhate, Matthew Hale, Gustavo Vazquez, Sarah Black, and Garry P. Nolan. Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed Imaging. *Cell*, 174(4):968–981.e15, August 2018. ISSN 0092-8674. doi: 10.1016/j.cell.2018.07.010. URL `http://www.sciencedirect.com/science/article/pii/S0092867418309048`.

Wuming Gong, Il-Youp Kwak, Pruthvi Pota, Naoko Koyano-Nakagawa, and Daniel J. Garry. DrImpute: imputing dropout events in single cell RNA sequencing data. *BMC Bioinformatics*, 19(1):220, June 2018. ISSN 1471-2105. doi: 10.1186/s12859-018-2226-y. URL `https://doi.org/10.1186/s12859-018-2226-y`.

R R Gray, O G Pybus, and M Salemi. Measuring the temporal structure in Serially-Sampled phylogenies. *Methods Ecol. Evol.*, 2(5):437–445, October 2011.

Anna Graybeal. Is it better to add taxa or characters to a difficult phylogenetic problem? *Syst. Biol.*, 47(1):9–17, 1998.

Christopher Heje Grønbech, Maximillian Fornitz Vording, Pascal N Timshel, Casper Kaae Sønderby, Tune Hannes Pers, and Ole Winther. scVAE: Variational auto-encoders for single-cell gene expression data. May 2018. URL `https://www.biorxiv.org/content/early/2018/05/16/318295`.

Dominic Grün, Lennart Kester, and Alexander van Oudenaarden. Validation of noise models for single-cell transcriptomics. *Nature Methods*, 11(6):637–640, June 2014. ISSN 1548-7105. doi: 10.1038/nmeth.2930. URL `https://www.nature.com/articles/nmeth.2930`.

Martin Guilliams, Charles-Antoine Dutertre, Charlotte L. Scott, Naomi McGovern, Dorine Sichien, Svetoslav Chakarov, Sofie Van Gassen, Jinmiao Chen, Michael Poidinger, Sofie De Prijck, Simon J. Tavernier, Ivy Low, Sergio Erdal Irac, Citra Nurfarah Mattar, Hermi Rizal Sumatoh, Gillian Hui Ling Low, Tam John Kit Chung, Dedrick Kok Hong Chan, Ker Kan Tan, Tony Lim Kiat Hon, Even Fossum, Bjarne Bogen, Mahesh Choolani, Jerry Kok Yen Chan, Anis Larbi, Hervé Luche, Sandrine Henri, Yvan Saeys, Evan William Newell, Bart N. Lambrecht, Bernard Malissen, and Florent Ginhoux. Unsupervised High-Dimensional Analysis Aligns Dendritic Cells across Tissues and Species. *Immunity*, 45(3):669–684, September 2016. ISSN 1074-7613. doi: 10.1016/j.immuni.2016.08.015. URL `http://www.sciencedirect.com/science/article/pii/S1074761316303399`.

Metin N Gurcan, Laura Boucheron, Ali Can, Anant Madabhushi, Nasir Rajpoot, and Bulent Yener. Histopathological image analysis: A review. *IEEE Rev. Biomed. Eng.*, 2:147, 2009.

Hiroshi Haeno, Mithat Gonen, Meghan B Davis, Joseph M Herman, Christine A

Iacobuzio-Donahue, and Franziska Michor. Computational modeling of pancreatic cancer reveals kinetics of metastasis suggesting optimum treatment strategies. *Cell*, 148(1-2):362–375, January 2012.

Christoph Hafemeister and Rahul Satija. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. March 2019. URL `https://www.biorxiv.org/content/10.1101/576827v2`.

Laleh Haghverdi, Maren Büttner, F. Alexander Wolf, Florian Buettner, and Fabian J. Theis. Diffusion pseudotime robustly reconstructs lineage branching. *Nature Methods*, 13(10):845–848, October 2016. ISSN 1548-7105. doi: 10.1038/nmeth.3971. URL `https://www.nature.com/articles/nmeth.3971`.

Laleh Haghverdi, Aaron T. L. Lun, Michael D. Morgan, and John C. Marioni. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*, 36(5):421–427, 2018. ISSN 1546-1696. doi: 10.1038/nbt.4091.

Xiaoping Han, Renying Wang, Yincong Zhou, Lijiang Fei, Huiyu Sun, Shujing Lai, Assieh Saadatpour, Ziming Zhou, Haide Chen, Fang Ye, Daosheng Huang, Yang Xu, Wentao Huang, Mengmeng Jiang, Xinyi Jiang, Jie Mao, Yao Chen, Chenyu Lu, Jin Xie, Qun Fang, Yibin Wang, Rui Yue, Tiefeng Li, He Huang, Stuart H Orkin, Guo-Cheng Yuan, Ming Chen, and Guoji Guo. Mapping the mouse cell atlas by Microwell-Seq. *Cell*, 172(5):1091–1107.e17, February 2018.

Andreas Heindl, Sidra Nawaz, and Yinyin Yuan. Mapping spatial heterogeneity in the tumor microenvironment: a new era for digital pathology. *Lab. Invest.*, 95(4):377–384, April 2015.

Stephanie C. Hicks and Roger D. Peng. Elements and Principles of Data Analysis. *arXiv:1903.07639 [stat]*, March 2019. URL `http://arxiv.org/abs/1903.07639`. arXiv: 1903.07639.

Stephanie C. Hicks, F. William Townes, Mingxiang Teng, and Rafael A. Irizarry. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics*, 19(4):562–578, October 2018. ISSN 1465-4644. doi: 10.1093/biostatistics/kxx053. URL `https://academic.oup.com/biostatistics/article/19/4/562/4599254`.

Elad Hoffer and Nir Ailon. Deep Metric Learning Using Triplet Network. In Aasa Feragen, Marcello Pelillo, and Marco Loog, editors, *Similarity-Based Pattern Recognition*, Lecture Notes in Computer Science, pages 84–92. Springer International Publishing, 2015. ISBN 978-3-319-24261-3.

Ian H Holmes. Solving the master equation for indels. *BMC Bioinformatics*, 18(1):255, May 2017.

Chung-Chau Hon, Jay W. Shin, Piero Carninci, and Michael J. T. Stubbington. The Human Cell Atlas: Technical approaches and challenges. *Briefings in Functional Genomics*, 17(4):283–294, July 2018. ISSN 2041-2649. doi: 10.1093/bfgp/elx029. URL `https://academic.oup.com/bfg/article/17/4/283/4571849`.

Masahito Hosokawa, Yohei Nishikawa, Masato Kogawa, and Haruko Takeyama. Massively parallel whole genome amplification for single-cell sequencing using droplet microfluidics. *Sci. Rep.*, 7(1):5199, July 2017.

Yong Hou, Kui Wu, Xulian Shi, Fuqiang Li, Luting Song, Hanjie Wu, Michael Dean, Guibo Li, Shirley Tsang, Runze Jiang, Xiaolong Zhang, Bo Li, Geng Liu, Niharika Bedekar, Na Lu, Guoyun Xie, Han Liang, Liao Chang, Ting Wang, Jianghao Chen, Yingrui Li, Xiuqing Zhang, Huanming Yang, Xun Xu, Ling Wang, and Jun Wang. Comparison of variations detection between whole-genome amplification methods used in single-cell resequencing. *Gigascience*, 4:37, August 2015.

Qiwen Hu and Casey S Greene. Parameter tuning is a key part of dimensionality reduction via deep variational autoencoders for single cell RNA transcriptomics. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 24:362–373, 2019. ISSN 2335-6936, 2335-6928. URL https://www.ncbi.nlm.nih.gov/pubmed/30963075.

Lei Huang, Fei Ma, Alec Chapman, Sijia Lu, and Xiaoliang Sunney Xie. Single-Cell Whole-Genome amplification and sequencing: Methodology and applications. *Annu. Rev. Genomics Hum. Genet.*, 16:79–102, June 2015.

Mo Huang, Jingshu Wang, Eduardo Torre, Hannah Dueck, Sydney Shaffer, Roberto Bonasio, John I. Murray, Arjun Raj, Mingyao Li, and Nancy R. Zhang. SAVER: gene expression recovery for single-cell RNA sequencing. *Nature Methods*, 15(7):539, July 2018. ISSN 1548-7105. doi: 10.1038/s41592-018-0033-z. URL https://www.nature.com/articles/s41592-018-0033-z.

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 somatic 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.org/10.1186/s13059-019-1673-8.

T. Höllt, N. Pezzotti, V. van Unen, F. Koning, B. P. F. Lelieveldt, and A. Vilanova. CyteGuide: Visual Guidance for Hierarchical Single-Cell Analysis. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):739–748, January 2018. ISSN 1077-2626. doi: 10.1109/TVCG.2017.2744318.

Giovanni Iacono, Elisabetta Mereu, Amy Guillaumet-Adkins, Roser Corominas, Ivon Cuscó, Gustavo Rodríguez-Esteban, Marta Gut, Luis Alberto Pérez-Jurado, Ivo Gut, and Holger Heyn. bigSCale: an analytical framework for big-scale single-cell data. *Genome Res.*, 28(6):878–890, June 2018.

Humayun Irshad, Antoine Veillard, Ludovic Roux, and Daniel Racoceanu. Methods for nuclei detection, segmentation, and classification in digital histopathology: A Review—Current status and future potential, 2014.

Martin Jacobsen. *Point Process Theory and Applications: Marked Point and Piecewise Deterministic Processes*. Springer Science & Business Media, December 2005.

Katharina Jahn, Jack Kuipers, and Niko Beerenwinkel. Tree inference for single-cell data. *Genome Biol.*, 17:86, May 2016.

Livnat Jerby-Arnon, Nadja Pfetzer, Yedael Y Waldman, Lynn McGarry, Daniel James, Emma Shanks, Brinton Seashore-Ludlow, Adam Weinstock, Tamar Geiger, Paul A

Clemons, Eyal Gottlieb, and Eytan Ruppin. Predicting cancer-specific vulnerability via data-driven detection of synthetic lethality. *Cell*, 158(5):1199–1209, August 2014.

Zhicheng Ji and Hongkai Ji. TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Research*, 44(13):e117, 2016. ISSN 1362-4962. doi: 10.1093/nar/gkw430.

Nelson Johansen and Gerald Quon. scAlign: a tool for alignment, integration and rare cell identification from scRNA-seq data. *bioRxiv*, page 504944, March 2019. doi: 10.1101/504944. URL 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 Yamamoto, 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 analysis 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/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, Simon R Myers, and Donald F Conrad. Unified single-cell analysis of testis gene regulation and pathology in five mouse strains. *eLife*, 8, June 2019. ISSN 2050-084X. doi: 10.7554/eLife.43966. URL http://dx.doi.org/10.7554/eLife.43966.

Melissa R Junttila and Frederic J de Sauvage. Influence of tumour micro-environment heterogeneity on therapeutic response. *Nature*, 501(7467):346–354, September 2013.

Hyun Min Kang, Meena Subramaniam, Sasha Targ, Michelle Nguyen, Lenka Maliskova, Elizabeth McCarthy, Eunice Wan, Simon Wong, Lauren Byrnes, Cristina Lanata, Rachel Gate, Sara Mostafavi, Alexander Marson, Noah Zaitlen, Lindsey A Criswell, and Chun Jimmie Ye. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nature biotechnology*, 36(1):89–94, January 2018a. ISSN 1087-0156. doi: 10.1038/nbt.4042. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5784859/.

Hyun Min Kang, Meena Subramaniam, Sasha Targ, Michelle Nguyen, Lenka Maliskova, Elizabeth McCarthy, Eunice Wan, Simon Wong, Lauren Byrnes, Cristina M Lanata, Rachel E Gate, Sara Mostafavi, Alexander Marson, Noah Zaitlen, Lindsey A Criswell, and Chun Jimmie Ye. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.*, 36(1): 89–94, January 2018b.

Jurrian Kornelis de Kanter, Philip Lijnzaad, Tito Candelli, Thanasis Margaritis, and Frank Holstege. CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing. *bioRxiv*, page 558908, February 2019. doi: 10.1101/558908. URL https://www.biorxiv.org/content/10.1101/558908v1.

Nikos Karaiskos, Philipp Wahle, Jonathan Alles, Anastasiya Boltengagen, Salah Ayoub, Claudia Kipar, Christine Kocks, Nikolaus Rajewsky, and Robert P Zinzen. The drosophila embryo at single-cell transcriptome resolution. *Science*, 358(6360):194–199, October 2017a.

Nikos Karaiskos, Philipp Wahle, Jonathan Alles, Anastasiya Boltengagen, Salah Ayoub, Claudia Kipar, Christine Kocks, Nikolaus Rajewsky, and Robert P. Zinzen. The Drosophila embryo at single-cell transcriptome resolution. *Science*, 358(6360):194–199, October 2017b. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aan3235. URL http://science.sciencemag.org/content/358/6360/194.

Ino D. Karemaker and Michiel Vermeulen. Single-Cell DNA Methylation Profiling: Technologies and Biological Applications. *Trends in Biotechnology*, 36(9):952–965, September 2018. ISSN 0167-7799, 1879-3096. doi: 10.1016/j.tibtech.2018.04.002. URL https://www.cell.com/trends/biotechnology/abstract/S0167-7799(18)30115-X.

Lennart Kester and Alexander van Oudenaarden. Single-Cell transcriptomics meets lineage tracing. *Cell Stem Cell*, May 2018.

Peter V Kharchenko, Lev Silberstein, and David T Scadden. Bayesian approach to single-cell differential expression analysis. *Nature Methods*, 11(7):740–742, July 2014. ISSN 1548-7091. doi: 10.1038/nmeth.2967. URL http://www.nature.com/doifinder/10.1038/nmeth.2967.

Kyu-Tae Kim, Hye Won Lee, Hae-Ock Lee, Sang Cheol Kim, Yun Jee Seo, Woosung Chung, Hye Hyeon Eum, Do-Hyun Nam, Junhyong Kim, Kyeung Min Joo, and Woong-Yang Park. Single-cell mRNA sequencing identifies subclonal heterogeneity in anti-cancer drug responses of lung adenocarcinoma cells. *Genome Biol.*, 16:127, June 2015.

Tae-Min Kim, Ruibin Xi, Lovelace J. Luquette, Richard W. Park, Mark D. Johnson, and Peter J. Park. Functional genomic analysis of chromosomal aberrations in a compendium of 8000 cancer genomes. *Genome Research*, 23(2): 217–227, 2013. doi: 10.1101/gr.140301.112. URL http://genome.cshlp.org/content/23/2/217.abstract.

Marek Kimmel and David Axelrod. *Branching Processes in Biology.* Interdisciplinary Applied Mathematics. Springer-Verlag, New York, 2 edition, 2015. ISBN 978-1-4939-1558-3. URL https://www.springer.com/gp/book/9781493915583.

Savvas Kinalis, Finn Cilius Nielsen, Ole Winther, and Frederik Otzen Bagger. Deconvolution of autoencoders to learn biological regulatory modules from single cell mRNA sequencing data. *BMC bioinformatics*, 20(1):379, July 2019. ISSN 1471-2105. doi: 10.1186/s12859-019-2952-9. URL http://dx.doi.org/10.1186/s12859-019-2952-9.

Vladimir Yu Kiselev, Andrew Yiu, and Martin Hemberg. scmap: projection of single-cell RNA-seq data across data sets. *Nature Methods*, 15(5):359–362, May 2018. ISSN 1548-7105. doi: 10.1038/nmeth.4644. URL `https://www.nature.com/articles/nmeth.4644`.

Vladimir Yu Kiselev, Tallulah S. Andrews, and Martin Hemberg. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature Reviews Genetics*, page 1, January 2019. ISSN 1471-0064. doi: 10.1038/s41576-018-0088-9. URL `https://www.nature.com/articles/s41576-018-0088-9`.

Allon M. Klein, Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A. Weitz, and Marc W. Kirschner. Droplet barcoding 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 Riethmüller. Comparative genomic hybridization, loss of heterozygosity, and DNA sequence analysis of single cells. *Proc. Natl. Acad. Sci. U. S. A.*, 96(8):4494–4499, April 1999.

Sergey Knyazev, Viachaslau Tsyvina, Andrew Melnyk, Alexander Artyomenko, Tatiana 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 Kolaczkowski and Joseph W Thornton. A mixed branch length model of heterotachy improves phylogenetic accuracy. *Mol. Biol. Evol.*, 25(6):1054–1066, June 2008.

Daisuke Komura and Shumpei Ishikawa. Machine learning methods for histopathological 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 Prabhakar, Dave Ruff, and Axel M Hillmer. Concurrent Single-Cell RNA and targeted DNA sequencing on an automated platform for comeasurement of genomic and transcriptomic signatures. *Clin. Chem.*, 65(2):272–281, February 2019.

Hazal Koptagel, Seong-Hwan Jun, and Jens Lagergren. SCuPhr: A Probabilistic Framework for Cell Lineage Tree Reconstruction. *bioRxiv*, page 357442, June 2018. doi: 10.1101/357442. URL `https://www.biorxiv.org/content/early/2018/06/29/357442`.

Keegan D Korthauer, Li-Fang Chu, Michael A Newton, Yuan Li, James Thomson, Ron Stewart, and Christina Kendziorski. A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biol.*, 17(1):222, October 2016a.

Keegan D. Korthauer, Li-Fang Chu, Michael A. Newton, Yuan Li, James Thomson, Ron Stewart, and Christina Kendziorski. A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biology*, 17(1):222, 2016b. ISSN 1474-760X. doi: 10.1186/s13059-016-1077-y.

Johannes Köster, Myles Brown, and Xiaole Shirley Liu. A bayesian model for single cell transcript expression analysis on MERFISH data, September 2017.

Dylan Kotliar, Adrian Veres, M Aurel Nagy, Shervin Tabrizi, Eran Hodis, Douglas A Melton, and Pardis C Sabeti. Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. *Elife*, 8:e43803, July 2019.

Alexey M. Kozlov, Diego Darriba, Tomáš Flouri, Benoit Morel, and Alexandros Stamatakis. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*, May 2019. doi: 10.1093/bioinformatics/btz305. URL https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btz305/5487384.

O Kozlov. *Models, Optimizations, and Tools for Large-Scale Phylogenetic Inference, Handling Sequence Uncertainty, and Taxonomic Validation*. PhD thesis, Karlsruhe Institute of Technology (KIT), October 2018.

Sergey Kryazhimskiy and Joshua B Plotkin. The population genetics of dN/dS. *PLoS Genet.*, 4(12):e1000304, December 2008.

Jack Kuipers, Katharina Jahn, Benjamin J Raphael, and Niko Beerenwinkel. Single-cell sequencing data reveal widespread recurrence and loss of mutational hits in the life histories of tumors. *Genome Res.*, 27 (11):1885–1894, November 2017.

Johannes Köster, Louis Dijkstra, Tobias Marschall, and Alexander Schönhuth. Enhancing sensitivity and controlling false discovery rate in somatic indel discovery. *bioRxiv*, page 741256, August 2019. doi: 10.1101/741256. URL https://www.biorxiv.org/content/10.1101/741256v1.

Emma Laks, Hans Zahn, Daniel Lai, Andrew McPherson, Adi Steif, Jazmine Brimhall, Justina Biele, Beixi Wang, Tehmina Masud, Diljot Grewal, Cydney Nielsen, Samantha Leung, Viktoria Bojilova, Maia Smith, Oleg Golovko, Steven Poon, Peter Eirew, Farhia Kabeer, Teresa Ruiz de Algara, So Ra Lee, M. Jafar Taghiyar, Curtis Huebner, Jessica Ngo, Tim Chan, Spencer Vatrt-Watts, Pascale Walters, Nafis Abrar, Sophia Chan, Matt Wiens, Lauren Martin, R. Wilder Scott, Michael T. Underhill, Elizabeth Chavez, Christian Steidl, Daniel Da Costa, Yusanne Ma, Robin J. N. Coope, Richard Corbett, Stephen Pleasance, Richard Moore, Andy J. Mungall, Cruk Imaxt Consortium, Marco A. Marra, Carl Hansen, Sohrab Shah, and Samuel Aparicio. Resource: Scalable whole genome sequencing of 40,000 single cells identifies stochastic aneuploidies, genome replication states and clonal repertoires. *bioRxiv*, page 411058, September 2018. doi: 10.1101/411058. URL https://www.biorxiv.org/content/early/2018/09/13/411058.

Ruben T H M Larue, Gilles Defraene, Dirk De Ruysscher, Philippe Lambin, and Wouter van Elmpt. Quantitative radiomics studies for tissue characterization: a review of technology and methodological procedures. *Br. J. Radiol.*, 90(1070):20160665, February 2017.

Si Quang Le, Cuong Cao Dang, and Olivier Gascuel. Modeling protein evolution with several amino acid replacement matrices depending on site rates. *Mol. Biol. Evol.*, 29(10):2921–2936, October 2012.

Adam D Leaché, Barbara L Banbury, Joseph Felsenstein, Adrián Nieto-Montes de Oca, and Alexandros Stamatakis. Short tree, long tree, right tree, wrong tree: New acquisition bias corrections for inferring SNP phylogenies. *Syst. Biol.*, 64(6):1032–1047, 2015.

Je Hyuk Lee, Evan R Daugharthy, Jonathan Scheiman, Reza Kalhor, Thomas C Ferrante, Richard Terry, Brian M Turczyk, Joyce L Yang, Ho Suk Lee, John Aach, Kun Zhang, and George M Church. Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. *Nat. Protoc.*, 10(3):442–458, March 2015.

Jeffrey T. Leek, Robert B. Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W. Evan Johnson, Donald 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 accurate and robust imputation method scimpute 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.

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 clustering 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.

G C Linderman, J Zhao, and Y Kluger. Zero-preserving imputation of scRNA-seq data using low-rank approximation. *bioRxiv*, 2018. URL https://www.biorxiv.org/content/10.1101/397588v1.abstract.

Liang Liu, Zhenxiang Xi, Shaoyuan Wu, Charles C Davis, and Scott V Edwards. Estimating phylogenetic trees from genome-scale data. *Ann. N. Y. Acad. Sci.*, 1360: 36–53, December 2015.

Jackson Loper, Trygve Bakken, Uygar Sumbul, Gabe Murphy, Hongkui Zeng, David Blei, and Liam Paninski. The Markov link method: a nonparametric approach to combine observations from multiple experiments. *bioRxiv*, page 457283, January 2019. doi: 10.1101/457283. URL https://www.biorxiv.org/content/10.1101/457283v3.

Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15 (12):1053–1058, December 2018. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-018-0229-2. URL http://dx.doi.org/10.1038/s41592-018-0229-2.

Eric Lubeck, Ahmet F Coskun, Timur Zhiyentayev, Mubhij Ahmad, and Long Cai. Single-cell in situ RNA profiling by sequential hybridization. *Nat. Methods*, 11 (4):360–361, April 2014.

Aaron T L Lun and John C Marioni. Overcoming confounding plate effects in differential expression analyses of single-cell RNA-seq data. *Biostatistics*, 18(3):451–464, July 2017.

Aaron T L Lun, Karsten Bach, and John C Marioni. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.*, 17:75, April 2016.

Aaron T L Lun, Arianne C Richard, and John C Marioni. Testing for differential abundance in mass cytometry data. *Nat. Methods*, 14(7):707–709, July 2017.

Tao Luo, Lei Fan, Rong Zhu, and Dong Sun. Microfluidic Single-Cell manipulation and analysis: Methods and applications. *Micromachines (Basel)*, 10(2), February 2019.

Iain C. Macaulay, Mabel J. Teng, Wilfried Haerty, Parveen Kumar, Chris P. Ponting, and Thierry Voet. Separation and parallel sequencing of the genomes and transcriptomes of single cells using G&amp;T-seq. *Nature Protocols*, 11(11):2081–2103, November 2016. ISSN 1750-2799. doi: 10.1038/nprot.2016.138. URL https://www.nature.com/articles/nprot.2016.138.

Iain C. Macaulay, Chris P. Ponting, and Thierry Voet. Single-Cell Multiomics: Multiple Measurements from Single Cells. *Trends in Genetics*, 33(2):155–168, February 2017. ISSN 0168-9525. doi: 10.1016/j.tig.2016.12.003. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5303816/.

Evan Z. Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R. Bialas, Nolan Kamitaki, Emily M. Martersteck, John J. Trombetta, David A. Weitz, Joshua R. Sanes, Alex K. Shalek, Aviv Regev, and Steven A. McCarroll. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, 161(5):1202–1214, May 2015. ISSN 1097-4172. doi: 10.1016/j.cell.2015.05.002.

Serghei Mangul, Lana S. Martin, Brian L. Hill, Angela Ka-Mei Lam, Margaret G. Distler, Alex Zelikovsky, Eleazar Eskin, and Jonathan Flint. Systematic benchmarking of omics computational tools. *Nature Communications*, 10(1): 1393, March 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-09406-4. URL https://www.nature.com/articles/s41467-019-09406-4.

Gioele La Manno, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochgerner, Viktor Petukhov, Katja Lidschreiber, Maria E. Kastriti, Peter Lönnerberg, Alessandro Furlan, Jean Fan, Lars E. Borm, Zehua Liu, David van Bruggen, Jimin Guo, Xiaoling He, Roger Barker, Erik Sundström, Gonçalo Castelo-Branco, Patrick Cramer, Igor Adameyko, Sten Linnarsson, and Peter V. Kharchenko. RNA velocity of single cells. *Nature*, 560 (7719):494, August 2018. ISSN 1476-4687. doi: 10.1038/s41586-018-0414-6. URL https://www.nature.com/articles/s41586-018-0414-6.

Erik A Martens, Rumen Kostadinov, Carlo C Maley, and Oskar Hallatschek. Spatial structure increases the waiting time for cancer. *New J. Phys.*, 13, November 2011.

Dariusz Matlak and Ewa Szczurek. Epistasis in genomic and survival data of cancer patients. *PLoS Comput. Biol.*, 13(7): e1005626, July 2017.

Davis James McCarthy, Raghd Rostom, Yuanhua Huang, Daniel J. Kunz, Petr Danecek, Marc Jan Bonder, Tzachi Hagai, HipSci Consortium, Wenyi Wang, Daniel J. Gaffney, Benjamin D. Simons, Oliver Stegle, and Sarah A. Teichmann. Cardelino: Integrating whole exomes and single-cell transcriptomes to reveal phenotypic impact of somatic variants. *bioRxiv*, page 413047, November 2018. doi: 10.1101/413047. URL https://www.biorxiv.org/content/10.1101/413047v2.

Nicholas McGranahan and Charles Swanton. Clonal heterogeneity and tumor evolution: Past, present, and the future. *Cell*, 168(4): 613–628, February 2017.

Chiara Medaglia, Amir Giladi, Liat Stoler-Barak, Marco De Giovanni, Tomer Meir Salame, Adi Biram, Eyal David, Hanjie Li, Matteo Iannacone, Ziv Shulman, and Ido Amit. Spatial reconstruction of immune niches by combining photoactivatable reporters and scRNA-seq. *Science*, 358(6370):1622–1626, December 2017. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aao4277. URL http://science.sciencemag.org/content/358/6370/1622.

Jing Meng and Yi-Ping Phoebe Chen. A database of simulated tumor genomes towards accurate detection of somatic small variants in cancer. *PLoS One*, 13(8): e0202982, August 2018.

Christopher R. Merritt, Giang T. Ong, Sarah Church, Kristi Barker, Gary Geiss, Margaret Hoang, Jaemyeong Jung, Yan Liang, Jill McKay-Fleisch, Karen Nguyen, Kristina Sorg, Isaac Sprague, Charles Warren, Sarah Warren, Zoey Zhou, Daniel R. Zollinger, Dwayne L. Dunaway, Gordon B. Mills, and Joseph M. Beechem. High multiplex, digital spatial profiling of proteins and RNA in fixed tissue using genomic detection methods. *bioRxiv*, page 559021, February 2019. doi: 10.1101/559021. URL https://www.biorxiv.org/content/10.1101/559021v2.

Zhun Miao, Jiaqi Li, and Xuegong Zhang. screcover: Discriminating true and false zeros in single-cell RNA-seq data for imputation. June 2019.

Franziska Michor, Yoh Iwasa, and Martin A Nowak. Dynamics of cancer progression. *Nat. Rev. Cancer*, 4(3):197–205, March 2004.

Jeffrey R Moffitt, Junjie Hao, Dhananjay Bambah-Mukku, Tian Lu, Catherine Dulac, and Xiaowei Zhuang. High-performance multiplexed fluorescence in situ hybridization in culture and tissue with matrix imprinting and clearing. *Proc. Natl. Acad. Sci. U. S. A.*, 113(50):14456–14461, December 2016.

Jeffrey R. Moffitt, Dhananjay Bambah-Mukku, Stephen W. Eichhorn, Eric Vaughn, Karthik Shekhar, Julio D. Perez, Nimrod D. Rubinstein, Junjie Hao, Aviv Regev, Catherine Dulac, and Xiaowei Zhuang. Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science*, 362(6416):eaau5324, November 2018. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aau5324. URL http://science.sciencemag.org/content/362/6416/eaau5324.

Kevin R Moon, Jay S Stanley, Daniel Burkhardt, David van Dijk, Guy Wolf, and Smita Krishnaswamy. Manifold learning-based methods for analyzing single-cell RNA-sequencing data. *Current Opinion in Systems Biology*, 7:36–46, 2018.

Marmar Moussa and Ion I. Măndoiu. Locality Sensitive Imputation for Single Cell RNA-Seq Data. *Journal of Computational Biology*, February 2019. doi: 10.1089/cmb.2018.0236. URL https://www.liebertpub.com/doi/10.1089/cmb.2018.0236.

Nature Methods, 2013. Method of the year 2013. *Nature Methods*, 11:1 EP –, 12 2013. URL https://doi.org/10.1038/nmeth.2801.

Nicholas Navin, Jude Kendall, Jennifer Troge, Peter Andrews, Linda Rodgers, Jeanne McIndoo, Kerry Cook, Asya Stepansky, 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, November 2014.

Malgorzata Nowicka, Carsten Krieg, Lukas M Weber, Felix J Hartmann, Silvia Guglietta, Burkhard Becher, Mitchell P Levesque, and Mark D Robinson. CyTOF workflow: 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 Altshuler, 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 Transcriptomes Unifies Multiple Mouse Cell Atlases into an Integrated Landscape. *bioRxiv*, page 397042, August 2018. doi: 10.1101/397042. URL https://www.biorxiv.org/content/10.1101/397042v2.

Anoop P Patel, Itay Tirosh, John J Trombetta, Alex K Shalek, Shawn M Gillespie, Hiroaki Wakimoto, Daniel P Cahill, Brian V Nahed, William T Curry, Robert L Martuza, David N Louis, Orit Rozenblatt-Rosen, Mario L Suvà, Aviv Regev, and Bradley E Bernstein. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, 344 (6190):1396–1401, June 2014.

Tao Peng, Qin Zhu, Penghang Yin, and Kai Tan. SCRABBLE: single-cell RNA-seq imputation constrained by bulk RNA-seq data. *Genome biology*, 20(1):88, May 2019. ISSN 1465-6906. doi: 10.1186/s13059-019-1681-8. URL http://dx.doi.org/10.1186/s13059-019-1681-8.

N. Pezzotti, T. Höllt, B. Lelieveldt, E. Eisemann, and A. Vilanova. Hierarchical Stochastic Neighbor Embedding. *Computer Graphics Forum*, 35(3):21–30, 2016.

ISSN 1467-8659. doi: 10.1111/cgf.12878. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.12878.

Ángel J Picher, Bettina Budeus, Oliver Wafzig, Carola Krüger, Sara García-Gómez, María I Martínez-Jiménez, Alberto Díaz-Talavera, Daniela Weber, Luis Blanco, and Armin Schneider. TruePrime is a novel method for whole-genome amplification from single cells based on TthPrimPol. *Nat. Commun.*, 7:13296, November 2016.

Emma Pierson and Christopher Yau. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression 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.

Mireya Plass, Jordi Solana, F. Alexander Wolf, Salah Ayoub, Aristotelis Misios, Petar Glažar, Benedikt Obermayer, Fabian J. Theis, Christine Kocks, and Nikolaus Rajewsky. Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science*, 360(6391): eaaq1723, May 2018. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aaq1723. URL http://science.sciencemag.org/content/360/6391/eaaq1723.

Hannah A. Pliner, Jonathan S. Packer, José L. McFaline-Figueroa, Darren A. Cusanovich, Riza M. Daza, Delasa Aghamirzaie, Sanjay Srivatsan, Xiaojie Qiu, Dana Jackson, Anna Minkina, Andrew C. Adey, Frank J. Steemers, Jay Shendure, and Cole Trapnell. Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. *Molecular Cell*, 71(5): 858–871.e8, 2018. ISSN 1097-4164. doi: 10.1016/j.molcel.2018.06.044.

Olivier Poirion, Xun Zhu, Travers Ching, and Lana X. Garmire. Using single nucleotide variations in single-cell RNA-seq to identify subpopulations and genotype-phenotype linkage. *Nature Communications*, 9(1): 4892, November 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-07170-5. URL https://www.nature.com/articles/s41467-018-07170-5.

David D Pollock, Derrick J Zwickl, Jimmy A McGuire, and David M Hillis. Increased taxon sampling is advantageous for phylogenetic inference. *Syst. Biol.*, 51(4):664–671, August 2002.

Vladimir Potapov and Jennifer L Ong. Examining sources of error in PCR by Single-Molecule sequencing. *PLoS One*, 12(1): e0169774, January 2017.

Xiaojie Qiu, Qi Mao, Ying Tang, Li Wang, Raghav Chawla, Hannah A. Pliner, and Cole Trapnell. Reversed graph embedding resolves complex single-cell trajectories. *Nature Methods*, 14(10):979–982, October 2017. ISSN 1548-7105. doi: 10.1038/nmeth.4402. URL https://www.nature.com/articles/nmeth.4402.

Bruce Rannala and Ziheng Yang. Efficient bayesian species tree inference under the multispecies coalescent. *Syst. Biol.*, 66(5): 823–842, September 2017.

Benjamin Redelings. Erasing errors due to alignment ambiguity when estimating positive selection. *Mol. Biol. Evol.*, 31(8):1979–1993, August 2014.

Aviv Regev, Sarah A. Teichmann, Eric S. Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter

Campbell, Piero Carninci, Menna Clatworthy, Hans Clevers, Bart Deplancke, Ian Dunham, James Eberwine, Roland Eils, Wolfgang Enard, Andrew Farmer, Lars Fugger, Berthold Göttgens, Nir Hacohen, Muzlifah Haniffa, Martin Hemberg, Seung Kim, Paul Klenerman, Arnold Kriegstein, Ed Lein, Sten Linnarsson, Joakim Lundeberg, Partha Majumder, John C. Marioni, Miriam Merad, Musa Mhlanga, Martijn Nawijn, Mihai Netea, Garry Nolan, Dana Pe'er, Anthony Philippakis, Chris P. Ponting, Steve Quake, Wolf Reik, Orit Rozenblatt-Rosen, Joshua Sanes, Rahul Satija, Ton N. Schumacher, Alex Shalek, Ehud Shapiro, Padmanee 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 Participants. The Human Cell Atlas. *bioRxiv*, page 121202, May 2017. doi: 10.1101/121202. URL https://www.biorxiv.org/content/10.1101/121202v1.

John E. Reid and Lorenz Wernisch. Pseudotime 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/bioinformatics/article/32/19/2973/2196633.

Stephen Reid, Jonathan Taylor, and Robert Tibshirani. A general framework for estimation and inference from clusters of features. *J. Am. Stat. Assoc.*, 113(521):280–293, January 2018.

Davide Risso, Fanny Perraudeau, Svetlana 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, January 2018. ISSN 2041-1723. doi: 10.1038/s41467-017-02554-5. URL https://www.nature.com/articles/s41467-017-02554-5.

Elena Rivas and Sean R Eddy. Probabilistic 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 insights into cellular differentiation and development. *Nature Biotechnology*, 35(6): 551–560, 2017. ISSN 1546-1696. doi: 10.1038/nbt.3854.

Simone Rizzetto, Auda A Eltahla, Peijie Lin, Rowena Bull, Andrew R Lloyd, Joshua W K Ho, Vanessa Venturi, and Fabio Luciani. Impact of sequencing depth and read length on single cell RNA sequencing data of T cells. *Sci. Rep.*, 7(1):12781, October 2017.

S Roch. A short proof that phylogenetic tree reconstruction by maximum likelihood is hard. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 3(1):92–94, 2006.

Samuel G. Rodriques, Robert R. Stickels, Aleksandrina Goeva, Carly A. Martin, Evan Murray, Charles R. Vanderburg, Joshua Welch, Linlin M. Chen, Fei Chen, and Evan Z. Macosko. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, 363(6434):1463–1467, March 2019. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aaw1219.

URL https://science.sciencemag.org/content/363/6434/1463.

Florian Rohart, Aida Eslami, Nicholas Matigian, Stéphanie Bougeard, and Kim-Anh Lê Cao. MINT: a multivariate integrative method to identify reproducible molecular signatures across independent experiments and platforms. *BMC Bioinformatics*, 18(1):128, February 2017a. ISSN 1471-2105. doi: 10.1186/s12859-017-1553-8. URL https://doi.org/10.1186/s12859-017-1553-8.

Florian Rohart, Benoît Gautier, Amrit Singh, and Kim-Anh Lê Cao. mixOmics: An R package for 'omics feature selection 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/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.sciencemag.org/content/360/6385/176.

Edith M Ross and Florian Markowetz. OncoNEM: 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 histolopathogical 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 inference methods. *Nature Biotechnology*, page 1, April 2019. ISSN 1546-1696. doi: 10.1038/s41587-019-0071-9. URL https://www.nature.com/articles/s41587-019-0071-9.

Yvan Saeys, Sofie Van Gassen, and Bart N. Lambrecht. Computational flow cytometry: helping to make sense of high-dimensional immunology data. *Nature Reviews Immunology*, 16(7):449–462, July 2016. ISSN 1474-1741. doi: 10.1038/nri.2016.56. URL https://www.nature.com/articles/nri.2016.56.

Stefano Santaguida, Amelia Richardson, Divya Ramalingam Iyer, Ons M'Saad, Lauren Zasadil, Kristin A. Knouse, Yao Liang Wong, Nicholas Rhind, Arshad Desai, and Angelika Amon. Chromosome Mis-segregation Generates Cell-Cycle-Arrested Cells with Complex Karyotypes that Are Eliminated by the Immune System. *Developmental Cell*, 41(6): 638–651.e5, June 2017. ISSN 15345807. doi: 10.1016/j.devcel.2017.05.022. URL https://linkinghub.elsevier.com/retrieve/pii/S1534580717304306.

Gryte Satas and Benjamin J Raphael. Haplotype phasing in single-cell DNA-sequencing data. *Bioinformatics*, 34(13):i211–i217, July 2018.

Rahul Satija, Jeffrey A Farrell, David Gennert, Alexander F Schier, and Aviv Regev. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.*, 33(5):495–502, May 2015.

Kenta Sato, Koki Tsuyuzaki, Kentaro Shimizu, and Itoshi Nikaido. CellFishing.jl: an ultrafast and scalable cell search method for single-cell RNA sequencing. *Genome Biology*, 20(1):31, February 2019. ISSN 1474-760X. doi: 10.1186/s13059-019-1639-x. URL https://doi.org/10.1186/s13059-019-1639-x.

Arpiar Saunders, Evan Z. Macosko, Alec Wysoker, Melissa Goldman, Fenna M. Krienen, Heather de Rivera, Elizabeth Bien, Matthew Baum, Laura Bortolin, Shuyu Wang, Aleksandrina Goeva, James Nemesh, Nolan Kamitaki, Sara Brumbaugh, David Kulp, and Steven A. McCarroll. Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain. *Cell*, 174(4):1015–1030.e16, August 2018. ISSN 0092-8674. doi: 10.1016/j.cell.2018.07.028. URL http://www.sciencedirect.com/science/article/pii/S0092867418309553.

Denis Schapiro, Hartland W Jackson, Swetha Raghuraman, Jana R Fischer, Vito R T Zanotelli, Daniel Schulz, Charlotte Giesen, Raúl Catena, Zsuzsanna Varga, and Bernd Bodenmiller. histoCAT: analysis of cell phenotypes and interactions in multiplex image cytometry data. *Nat. Methods*, 14 (9):873–876, September 2017.

Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Siyan Liu, Stacie Lin, Peter Berube, Lia Lee, Jenny Chen, Justin Brumbaugh, Philippe Rigollet, Konrad Hochedlinger, Rudolf Jaenisch, Aviv Regev, and Eric S. Lander. Reconstruction of developmental landscapes by optimal-transport analysis of single-cell gene expression sheds light on cellular reprogramming. *bioRxiv*, page 191056, September 2017. doi: 10.1101/191056. URL https://www.biorxiv.org/content/10.1101/191056v1.

Herbert B Schiller, Daniel T Montoro, Lukas M Simon, Emma L Rawlins, Kerstin B Meyer, Maximilian Strunz, Felipe Vieira Braga, Wim Timens, Gerard H Koppelman, G.R. Scott Budinger, Janette K Burgess, Avinash Waghray, Maarten van den Berge, Fabian J Theis, Aviv Regev, Naftali Kaminski, Jayaraj Rajagopal, Sarah A Teichmann, Alexander V Misharin, and Martijn C Nawijn. The Human Lung Cell Atlas - A high-resolution reference map of the human lung in health and disease. *American Journal of Respiratory Cell and Molecular Biology*, April 2019. ISSN 1044-1549. doi: 10.1165/rcmb.2018-0416TR. URL https://www.atsjournals.org/doi/abs/10.1165/rcmb.2018-0416TR.

Roland F. Schwarz, Anne Trinh, Botond Sipos, James D. Brenton, Nick Goldman, and Florian Markowetz. Phylogenetic Quantification of Intra-tumour Heterogeneity. *PLoS Computational Biology*, 10(4):e1003535, April 2014. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1003535. URL https://dx.plos.org/10.1371/journal.pcbi.1003535.

Roberto Semeraro, Valerio Orlandini, and Alberto Magi. Xome-Blender: A novel cancer genome simulator. *PLoS One*, 13(4): e0194472, April 2018.

Debarka Sengupta, Nirmala Arul Rayan, Michelle Lim, Bing Lim, and Shyam Prabhakar. Fast, scalable and accurate differential expression analysis for single cells. *bioRxiv*, page 049734, April 2016. doi: 10.1101/049734. URL `https://www.biorxiv.org/content/10.1101/049734v1`.

Manu Setty, Michelle D. Tadmor, Shlomit Reich-Zeliger, Omer Angel, Tomer Meir Salame, Pooja Kathail, Kristy Choi, Sean Bendall, Nir Friedman, and Dana Pe'er. Wishbone identifies bifurcating developmental 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/articles/nbt.3569`.

D T Severson, R P Owen, M J White, X Lu, and B Schuster-Böckler. BEARscc determines robustness of single-cell clusters using 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 Unnikrishnan, and Aleksandra Radenovic. On characterizing protein spatial clusters with correlation approaches. *Sci. Rep.*, 6:31164, August 2016.

Angus M Sidore, Freeman Lan, Shaun W Lim, and Adam R Abate. Enhanced sequencing coverage with digital droplet multiple 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/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 approach. *bioRxiv*, page 067611, March 2018. doi: 10.1101/067611. URL `https://www.biorxiv.org/content/10.1101/067611v2`.

Debajyoti Sinha, Akhilesh Kumar, Himanshu Kumar, Sanghamitra Bandyopadhyay, and Debarka Sengupta. dropclust: efficient clustering of ultra-large scRNA-seq data. *Nucleic Acids Res.*, 46(6):e36, April 2018.

Pavel Skums, Viachaslau Tsyvina, and Alex Zelikovsky. Inference of clonal selection in cancer populations using single-cell sequencing data. *bioRxiv*, page 465211, January 2019. doi: 10.1101/465211. URL `https://www.biorxiv.org/content/10.1101/465211v2`.

Martin D Smith, Joel O Wertheim, Steven Weaver, Ben Murrell, Konrad Scheffler, and Sergei L Kosakovsky Pond. Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Mol. Biol. Evol.*, 32(5):1342–1353, May 2015.

Charlotte Soneson and Mark D Robinson. Towards unified quality verification of synthetic count data with countsimQC. *Bioinformatics*, 34(4):691–692, 2017.

Charlotte Soneson and Mark D Robinson. Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods*, February 2018.

Bastiaan Spanjaard, Bo Hu, Nina Mitic, Pedro Olivares-Chauvet, Sharan Janjuha, Nikolay Ninov, and Jan Philipp Junker. Simultaneous lineage tracing and cell-type identification using CRISPR-Cas9-induced genetic scars. *Nat. Biotechnol.*, 36(5):469–473, June 2018.

C Spits, C Le Caignec, M De Rycke, 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 displacement amplification from single cells. *Nat. Protoc.*, 1(4):1965–1970, November 2006b.

S Srinivasan, N T Johnson, and D Korkin. A Hybrid Deep Clustering Approach for Robust Cell Type Profiling Using Single-cell RNA-seq Data. *bioRxiv*, 2019. URL https://www.biorxiv.org/content/10.1101/511626v1.abstract.

Divyanshu Srivastava, Arvind Iyer, Vibhor Kumar, and Debarka Sengupta. CellAtlasSearch: a scalable search engine for single cells. *Nucleic Acids Research*, 46(W1):W141–W147, July 2018. ISSN 0305-1048. doi: 10.1093/nar/gky421. URL https://academic.oup.com/nar/article/46/W1/W141/5000022.

Oliver Stegle, Sarah A Teichmann, and John C Marioni. Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.*, 16(3):133, January 2015.

Genevieve L Stein-O'Brien, Brian S Clark, Thomas Sherman, Cristina Zibetti, Qiwen Hu, Rachel Sealfon, Sheng Liu, Jiang Qian, Carlo Colantuoni, Seth Blackshaw, Loyal A Goff, and Elana J Fertig. Decomposing Cell Identity for Transfer Learning across Cellular Measurements, Platforms, Tissues, and Species. *Cell systems*, 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/10.1016/j.cels.2019.04.004.

Lars Steinbrück and Alice Carolyn McHardy. Allele dynamics plots for the study of evolutionary dynamics in viral populations. *Nucleic Acids Res.*, 39(1):e4, January 2011.

Carina Strell, Markus M Hilscher, Navya Laxman, Jessica Svedlund, Chenglin Wu, Chika Yokota, and Mats Nilsson. Placing RNA in context and space - methods for spatially resolved transcriptomics. *FEBS J.*, March 2018.

Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M. Mauck, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single cell data. *bioRxiv*, page 460147, November 2018. doi: 10.1101/460147. URL https://www.biorxiv.org/content/10.1101/460147v1.

Michael J T Stubbington, Orit Rozenblatt-Rosen, Aviv Regev, and Sarah A Teichmann. Single-cell transcriptomics to explore the immune system in health and disease. *Science*, 358(6359):58–63, October 2017.

Patrik L. Ståhl, Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández

Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub O. Westholm, Mikael Huss, Annelie Mollbrink, Sten Linnarsson, Simone Codeluppi, Åke Borg, Fredrik Pontén, Paul Igor Costea, Pelin Sahlén, Jan Mulder, Olaf Bergmann, Joakim Lundeberg, and Jonas Frisén. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science (New York, N.Y.)*, 353 (6294):78–82, July 2016. ISSN 1095-9203. doi: 10.1126/science.aaf2403.

S Sun, J Zhu, Y Ma, and X Zhou. Accuracy, Robustness and Scalability of Dimensionality Reduction Methods for Single Cell RNAseq Analysis. *bioRxiv*, 2019. URL https://www.biorxiv.org/content/10.1101/641142v1.abstract.

Valentine Svensson, Sarah A Teichmann, and Oliver Stegle. SpatialDE: identification of spatially variable genes. *Nat. Methods*, 15 (5):343–346, May 2018a.

Valentine Svensson, Roser Vento-Tormo, and Sarah A. Teichmann. Exponential scaling of single-cell RNA-seq in the past decade. *Nature Protocols*, 13(4):599–604, April 2018b. ISSN 1750-2799. doi: 10.1038/nprot.2017.149. URL https://www.nature.com/articles/nprot.2017.149.

Charles Swanton. Intratumor heterogeneity: evolution through space and time. *Cancer Res.*, 72(19):4875–4882, October 2012.

Ewa Szczurek, Navodit Misra, and Martin Vingron. Synthetic sickness or lethality points at candidate combination therapy targets in glioblastoma. *Int. J. Cancer*, 133 (9):2123–2132, November 2013.

The Tabula Muris Consortium. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature*, 562(7727): 367, October 2018. ISSN 1476-4687. doi: 10.1038/s41586-018-0590-4. URL https://www.nature.com/articles/s41586-018-0590-4.

Divyanshu Talwar, Aanchal Mongia, Debarka Sengupta, and Angshul Majumdar. AutoImpute: Autoencoder based imputation of single-cell RNA-seq data. *Scientific reports*, 8(1):16329, November 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-34688-x. URL http://dx.doi.org/10.1038/s41598-018-34688-x.

Amos Tanay and Aviv Regev. Scaling single-cell genomics from phenomenology to mechanism. *Nature*, 541(7637):331–338, January 2017.

W Tang, F Bertaux, P Thomas, C Stefanelli, M Saint, and others. bayNorm: Bayesian gene expression recovery, imputation and normalisation for single cell RNA-sequencing data. *bioRxiv*, 2018. URL https://www.biorxiv.org/content/10.1101/384586v2.abstract.

H Telenius, N P Carter, C E Bebb, M Nordenskjöld, B A Ponder, and A Tunnacliffe. Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer. *Genomics*, 13(3):718–725, July 1992.

Luyi Tian, Xueyi Dong, Saskia Freytag, Kim-Anh Lê Cao, Shian Su, Abolfazl JalalAbadi, Daniela Amann-Zalcenstein, Tom S. Weber, Azadeh Seidi, Jafar S. Jabbari, Shalin H. Naik, and Matthew E. Ritchie. Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. *Nature Methods*, 16 (6):479, June 2019. ISSN 1548-7105. doi: 10.1038/s41592-019-0425-8. URL https://www.nature.com/articles/s41592-019-0425-8.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27885v3 | CC BY 4.0 Open Access | rec: 23 Aug 2019, publ: 23 Aug 2019

F William Townes, Stephanie C Hicks, Martin J Aryee, and Rafael A Irizarry. Feature Selection and Dimension Reduction for Single Cell RNA-Seq based on a Multinomial Model. March 2019. URL https://www.biorxiv.org/content/10.1101/574574v1.

Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J. Lennon, Kenneth J. Livak, Tarjei S. Mikkelsen, and John L. Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, 32(4):381–386, April 2014. ISSN 1546-1696. doi: 10.1038/nbt.2859.

Samra Turajlic and Charles Swanton. Metastasis as an evolutionary process. *Science*, 352(6282):169–175, April 2016.

Vincent van Unen, Thomas Höllt, Nicola Pezzotti, Na Li, Marcel J. T. Reinders, Elmar Eisemann, Frits Koning, Anna Vilanova, and Boudewijn P. F. Lelieveldt. Visual analysis of mass cytometry 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.

Catalina A Vallejos, John C Marioni, and Sylvia Richardson. BASiCS: Bayesian analysis 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.esmoiotech.org/article/S2590-0188(19)30002-4/abstract.

Koen van den Berge, Hector Roux de Bezieux, Kelly Street, Wouter Saelens, Robrecht 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.

Dimitrios V Vavoulis, Margherita Francescatto, Peter Heutink, and Julian Gough. DGEclust: differential expression analysis of clustered count data. *Genome Biol.*, 16:39, February 2015.

A Verma and B Engelhardt. A robust nonlinear low-dimensional manifold for single cell RNA-seq data. *bioRxiv*, 2018.

Beate Vieth, Christoph Ziegenhain, Swati Parekh, Wolfgang Enard, and Ines Hellmann. powsimr: power analysis for bulk and single cell RNA-seq experiments. *Bioinformatics*, 33(21):3486–3488, November 2017.

Irma Virant-Klun, Stefan Leicht, Christopher Hughes, and Jeroen Krijgsveld. Identification of Maturation-Specific Proteins by Single-Cell Proteomics of Human Oocytes. *Molecular & cellular proteomics: MCP*, 15 (8):2616–2627, 2016. ISSN 1535-9484. doi: 10.1074/mcp.M115.056887.

Sarah A. Vitak, Kristof A. Torkenczy, Jimi L. Rosenkrantz, Andrew J. Fields, Lena Christiansen, Melissa H. Wong, Lucia Carbone, Frank J. Steemers, and Andrew Adey. Sequencing thousands of single-cell genomes with combinatorial indexing. *Nature Methods*, 14(3):302–308, March 2017.

ISSN 1548-7105. doi: 10.1038/nmeth.4154. URL https://www.nature.com/articles/nmeth.4154.

Bartlomiej Waclaw, Ivana Bozic, Meredith E Pittman, Ralph H Hruban, Bert Vogelstein, and Martin A Nowak. A spatial model predicts that dispersal and cell turnover limit intratumour heterogeneity. *Nature*, 525(7568):261–264, September 2015.

Daniel E. Wagner, Caleb Weinreb, Zach M. Collins, James A. Briggs, Sean G. Megason, 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.

F Wagner, D Barkley, and I Yanai. Accurate denoising of single-cell RNA-Seq data using unbiased principal component analysis. *bioRxiv*, 2019.

Florian Wagner and Itai Yanai. Moana: A robust and scalable cell type classification framework for single-cell RNA-Seq data. *bioRxiv*, page 456129, October 2018. doi: 10.1101/456129. URL https://www.biorxiv.org/content/10.1101/456129v1.

Florian Wagner, Yun Yan, and Itai Yanai. K-nearest neighbor smoothing for high-throughput single-cell RNA-Seq data. January 2018b. URL https://www.biorxiv.org/content/early/2018/01/24/217737?rss=1.

Dongfang Wang and Jin Gu. VASC: Dimension reduction and visualization of single-cell RNA-seq data by deep variational autoencoder. *Genomics Proteomics Bioinformatics*, 16(5):320–331, October 2018.

Jian Wang and Yuanlin Song. Single cell sequencing: a distinct new field. *Clin. Transl. Med.*, 6(1):10, December 2017.

Jingshu Wang, Divyansh Agarwal, Mo Huang, Gang Hu, Zilu Zhou, Vincent B Conley, Hugh MacMullan, and Nancy R Zhang. Transfer learning in single-cell transcriptomics improves data denoising and pattern discovery. November 2018.

T Wang, T S Johnson, W Shao, J Zhang, and K Huang. BURMUDA: A novel deep transfer learning method for single-cell RNA sequencing batch correction reveals hidden high-resolution cellular subtypes. *bioRxiv*, 2019.

Gregory P Way and Casey S Greene. Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 23:80–91, 2018. ISSN 2335-6936, 2335-6928. doi: 10.1142/9789813235533_0008. URL https://www.ncbi.nlm.nih.gov/pubmed/29218871.

Lukas M. Weber and Mark D. Robinson. Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data. *Cytometry Part A*, 89 (12):1084–1096, December 2016. ISSN 1552-4922. doi: 10.1002/cyto.a.23030. URL https://onlinelibrary.wiley.com/doi/full/10.1002/cyto.a.23030.

Lukas M. Weber, Malgorzata Nowicka, Charlotte Soneson, and Mark D. Robinson. diffcyt: Differential discovery in high-dimensional cytometry via high-resolution clustering. *bioRxiv*, page 349738, November 2018. doi: 10.1101/349738. URL https://www.biorxiv.org/content/10.1101/349738v2.

Lukas M. Weber, Wouter Saelens, Robrecht Cannoodt, Charlotte Soneson, Alexander Hapfelmeier, Paul P. Gardner, Anne-Laure Boulesteix, Yvan Saeys, and Mark D. Robinson. Essential guidelines for computational method benchmarking. *Genome Biology*, 20(1):125, June 2019. ISSN 1474-760X. doi: 10.1186/s13059-019-1738-8. URL `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 inference from single-cell snapshots. *Proceedings 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.org/content/115/10/E2467`.

Joshua Welch, Velina Kozareva, Ashley Ferreira, 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/content/10.1101/459891v1`.

Joshua D Welch, Alexander J Hartemink, and Jan F Prins. MATCHER: manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics. *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 Rajewsky, Lukas Simon, and Fabian J. Theis. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biology*, 20(1):59, March 2019. ISSN 1474-760X. doi: 10.1186/s13059-019-1663-x. URL `https://doi.org/10.1186/s13059-019-1663-x`.

Larry Xi, Alexander Belyaev, Sandra Spurgeon, Xiaohui Wang, Haibiao Gong, Robert Aboukhalil, and Richard Fekete. New library construction method for single-cell genomes. *PLoS One*, 12(7):e0181163, July 2017.

Li Charlie Xia, Dongmei Ai, Hojoon Lee, Noemi Andor, Chao Li, Nancy R Zhang, and Hanlee P Ji. SVEngine: an efficient and versatile simulator of genome structural variations with features of cancer clonal evolution. *Gigascience*, 7(7), July 2018.

Li Yang and P Charles Lin. Mechanisms that drive inflammatory tumor microenvironment, tumor heterogeneity, and metastatic progression. *Semin. Cancer Biol.*, 47:185–195, December 2017.

Z Yang. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.*, 39(3):306–314, September 1994.

Yinyin Yuan. Spatial heterogeneity in the tumor microenvironment. *Cold Spring Harb. Perspect. Med.*, 6(8), August 2016.

Simone Zaccaria, Mohammed El-Kebir, Gunnar W. Klau, and Benjamin J. Raphael. The Copy-Number Tree Mixture Deconvolution Problem and Applications to Multi-sample Bulk Sequencing Tumor Data. In S. Cenk Sahinalp, editor, *Research in Computational Molecular Biology*, Lecture Notes in Computer Science, pages 318–335.

Springer International Publishing, 2017. ISBN 978-3-319-56970-3.

H Zafar, N Navin, K Chen, and L Nakhleh. SiCloneFit: Bayesian inference of population structure, genotype, and phylogeny of tumor clones from single-cell genome sequencing data. *bioRxiv*, 2018.

Hamim Zafar, Yong Wang, Luay Nakhleh, Nicholas Navin, and Ken Chen. Monovar: 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.nature.com/articles/nmeth.3835.

Hamim Zafar, Anthony Tzen, Nicholas Navin, Ken Chen, and Luay Nakhleh. SiFit: inferring tumor trees from single-cell sequencing 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 preamplification. *Nature Methods*, 14(2):167–173, February 2017b. ISSN 1548-7105. doi: 10.1038/nmeth.4140. URL 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 Single and Multiple Gene Copies : A Review. Not clear where this was initially published and whether it is peer-reviewed., 2018. URL https://pdfs.semanticscholar.org/85e6/7eb03d1b3d004c60a12df08c1f937fbaa974.pdf.

Amit Zeisel, Hannah Hochgerner, Peter Lönnerberg, Anna Johnsson, Fatima Memic, Job van der Zwan, Martin Häring, Emelie Braun, Lars E. Borm, Gioele La Manno, Simone Codeluppi, Alessandro Furlan, Kawai Lee, Nathan Skene, Kenneth D. Harris, Jens Hjerling-Leffler, Ernest Arenas, Patrik Ernfors, Ulrika Marklund, and Sten Linnarsson. Molecular Architecture of the Mouse Nervous System. *Cell*, 174(4):999–1014.e22, August 2018. ISSN 0092-8674. doi: 10.1016/j.cell.2018.06.021. URL http://www.sciencedirect.com/science/article/pii/S009286741830789X.

Allen W. Zhang, Ciara O'Flanagan, Elizabeth Chavez, Jamie LP Lim, Andrew McPherson, Matt Wiens, Pascale Walters, Tim Chan, Brittany Hewitson, Daniel Lai, Anja Mottok, Clementine Sarkozy, Lauren Chong, Tomohiro Aoki, Xuehai Wang, Andrew P. Weng, Jessica N. McAlpine, Samuel Aparicio, Christian Steidl, Kieran R. Campbell, and Sohrab P. Shah. Probabilistic cell type assignment of single-cell transcriptomic data reveals spatiotemporal microenvironment dynamics in human cancers. *bioRxiv*, page 521914, January 2019a. doi: 10.1101/521914. URL https://www.biorxiv.org/content/10.1101/521914v1.

Chao Zhang. Single-Cell Data Analysis Using MMD Variational Autoencoder. April

2019. URL https://www.biorxiv.org/content/10.1101/613414v1.abstract.

Huanan Zhang, Catherine A A Lee, Zhuliu Li, John R Garbe, Cindy R Eide, Raphael Petegrosso, Rui Kuang, and Jakub Tolar. A multitask clustering approach for single-cell RNA-seq analysis in recessive dystrophic epidermolysis bullosa. *PLoS Comput. Biol.*, 14(4):e1006053, April 2018.

Jesse M. Zhang, Govinda M. Kamath, 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/10.1101/463265v3.

Jianzhi Zhang, Rasmus Nielsen, and Ziheng Yang. Evaluation of an improved branch-site likelihood method for detecting positive 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/s41467-017-01968-5.

L. Zhang and S. Zhang. Comparison of computational 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 Navidi, and N Arnheim. Whole genome amplification 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. *Bioinformatics*, May 2019c. ISSN 1367-4803, 1367-4811. doi: 10.1093/bioinformatics/btz435. URL http://dx.doi.org/10.1093/bioinformatics/btz435.

Xiaoyan Zhang, Sadie L Marjani, Zhaoyang Hu, Sherman M Weissman, Xinghua Pan, and Shixiu Wu. Single-Cell sequencing for precise cancer research: Progress and prospects. *Cancer Res.*, 76(6):1305–1312, March 2016.

Xiuwei Zhang, Chenling Xu, and Nir Yosef. SymSim: simulating multi-faceted variability in single cell RNA sequencing. *bioRxiv*, page 378646, April 2019d. doi: 10.1101/378646. URL https://www.biorxiv.org/content/10.1101/378646v3.

Yifan Zhang and Feng Liu. Multidimensional Single-Cell analyses in organ development and maintenance. *Trends Cell Biol.*, March 2019.

Grace X. Y. Zheng, Jessica M. Terry, Phillip Belgrader, Paul Ryvkin, Zachary W. Bent, Ryan Wilson, Solongo B. Ziraldo, Tobias D. Wheeler, Geoff P. McDermott, Junjie Zhu, Mark T. Gregory, Joe Shuga, Luz Montesclaros, Jason G. Underwood, Donald A. Masquelier, Stefanie Y. Nishimura, Michael Schnall-Levin, Paul W. Wyatt, Christopher M. Hindson, Rajiv Bharadwaj, Alexander Wong, Kevin D. Ness, Lan W. Beppu, H. Joachim Deeg, Christopher McFarland, Keith R. Loeb, William J. Valente, Nolan G. Ericson, Emily A. Stevens, Jerald P. Radich, Tarjei S. Mikkelsen, Benjamin J. Hindson, and Jason H. Bielas. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8:14049, January 2017. ISSN 2041-1723. doi: 10.1038/ncomms14049. URL https://www.nature.com/articles/ncomms14049.

Lingxue Zhu, Jing Lei, Bernie Devlin, and Kathryn Roeder. A UNIFIED STATISTICAL FRAMEWORK FOR SINGLE CELL 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.

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 variations of a single human cell. *Science*, 338 (6114):1622–1626, December 2012.