12 Grand Challenges in Single-Cell Data Science

David Lähnemann*1,2,3, Johannes Küster*1,2,3, Ewa Szczurek*1,4, Davis J. McCarthy*5,6,7, Stephanie C. Hicks*8, Mark D. Robinson*9, Catalina A. Vallesjo*10,11, Niko Beerenwinkel*12,13, Kieran R. Campbell*14,15,16,17, Ahmed Mahfouz*18,19, Luca Pinello*20,21,22, Pavel Skurnik*23, Alexandros Stamatakis*24,25, Camille Stephan-Otto Attolini*26,27, Samuel Aparicio*18,27, Jasmijn Baaijens*28, Marleen Balvert*29,31, buys de Barbanson*32,33, Antonio Cappuccio*34, Giacomo Corleone*36, Bas E. Dutilh*37,38, Maria Florescu*39,40, Victor Gyurev*41, Rens Holmer*42, Katharina Jahn*12,13, Thamar Jesuorun Lobo*43, Emma M. Keizer*44, Indu Khatri*45, Syzmon M. Kibbassa*46, Jan O. Korben*47, Alexey M. Kozlov*48, Tzu-Hao Kuo*49, Boudewijn P.F. Leelievd50,51, John C. Marion*52,53,54, Tobias Marschall55,56, Felix Mölder*57,58, Amir Niknejad*59,60, Lukasz Raczkowski*61,62, Marcel Reinders18,19, Jeroen de Ridder32,33, Antoine-Emmanuel Saliba63, Antonios Sonarakis*64, Oliver Stegle48,54,63, Fabian J. Theis65, Huan Yang*66, Alex Zelikovsky67,70, Alice C. McHardy1,71, Benjamin J. Raphael1,71, Sohrob P. Shah1,72, and Alexander Schönthut16,17,29,31

*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
14 Department of Statistics, University of British Columbia, Vancouver, Canada
15 Department of Molecular Oncology, BC Cancer Agency, Vancouver, Canada
16 Data Science Institute, University of British Columbia, Vancouver, Canada
17 Leiden Computational Biology Center, Leiden University Medical Center, The Netherlands
18 Delft Bioinformatics Lab, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, The Netherlands
19 Molecular Pathology Unit and Center for Cancer Research, Massachusetts General Hospital Research Institute, Charlestown, USA
20 Department of Pathology, Harvard Medical School, Boston, USA
21 Broad Institute of Harvard and MIT, Cambridge, MA, USA
22 Department of Computer Science, Georgia State University, Atlanta, USA
23 Computational Molecular Evolution Group, Heidelberg Institute for Theoretical Studies, Germany
24 Institute for Theoretical Informatics, Karlsruhe Institute of Technology, Germany
25 Institute for Research in Biomedicine, The Barcelona Institute of Science and Technology, Spain
26 Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, Canada
27 Life Science Institute, University of Warwick, Coventry, UK
28 Theoretical Biology and Bioinformatics, Science for Life, Umeå University, Sweden
29 Center for Molecular Medicine, University Medical Center Utrecht, The Netherlands
30 OncoCode Institute, Utrecht, The Netherlands
31 Quantitative biology, Hubrecht Institute, Utrecht, The Netherlands
32 Institute for Advanced Study, University of Amsterdam, The Netherlands
33 Department of Surgery and Cancer, The Imperial Centre for Translational and Experimental Medicine, Imperial College London, UK
34 Centre for Molecular and Biomolecular Informatics, Radboud University Medical Center, Nijmegen, The Netherlands
35 European Research Institute for the Biology of Ageing, University Medical Center Groningen, University of Groningen, The Netherlands
36 European Molecular Biology Laboratory, European Bioinformatics Institute, Heidelberg, Germany
37 Institute of Pathology, University Hospital Essen, University of Duisburg-Essen, Germany
38 Computation molecular design, Zuse Institute Berlin, Germany
39 Mathematics department, Mount Saint Vincent, New York, USA
40 Helmholtz Institute for RNA-based Infection Research, Helmholtz-Center for Infection Research, Würzburg, Germany
41 Institute of Computational Biology, Helmholtz Zentrum München – German Research Center for Environmental Health, Neuberger, Germany
42 Division of Drug Discovery and Safety, Leiden Academic Center for Drug Research – LACDR — Leiden University, The Netherlands
43 Division of Computational Genomics and Systems Genetics, German Cancer Research Center – DKFZ, Heidelberg, Germany
44 Institute of Computational Biology, Helmholtz Centre for Infection Research, Helmholtz-Zentrum Berlin, Germany
45 Joint last authors, workshop organizers.
46 Corresponding author: Alexander Schönthuth, as@cw1.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.

3 Challenges in single-cell transcriptomics

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

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

3.3 Challenge III: Mapping single cells to a reference atlas

3.4 Challenge IV: Generalizing trajectory inference

3.5 Challenge V: Finding patterns in spatially resolved measurements

4 Challenges in single-cell genomics

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

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

5 Challenges in single-cell phylogenomics

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

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

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

Contents

1 Introduction 3

2 Single-Cell Data Science: Themes and Categories 4

2.1 Varying levels of resolution 5

2.2 Quantifying uncertainty of measurements and analysis results 5

2.3 Scaling to higher dimensionalities: more cells, more features, broader coverage 6

2.4 Challenge categories 7

4 Challenges in single-cell genomics 19

4.1 Challenge VI: Improving single-cell DNA sequencing data quality and scaling to more cells 20

4.2 Challenge VII: Errors and missing data in the identification of features / variation from single-cell DNA sequencing data 23

5 Challenges in single-cell phylogenomics 25

5.1 Challenge VIII: Scaling phylogenetic models to many cells and many sites 26

5.2 Challenge IX: Integrating multiple types of features / variation into phylogenetic models 27

5.3 Challenge X: Inferring population genetic parameters of tumor heterogeneity by model integration 29
6 Overarching challenges  
6.1 Challenge XI: Integration of single-cell data: across samples, experiments and types of measurement  
6.2 Challenge XII: Validating and benchmarking analysis tools for single-cell measurements  
7 Acknowledgements

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-
1 Cell Data Science (SCDS).

2 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, over-
arching question: (iii) Which exact level of
resolution is appropriate relative to the par-
ticular research question one has in mind (sec-
tion 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 characteriza-
tions supported by bulk sequencing experi-
ments. 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 cer-
tain freedom in terms of resolution, and de-
pending on the limits imposed by the particu-
lar experimental setup, achieving the targeted
level of resolution or granularity for the in-
tended 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 devel-
opment and tissue organization; (ii) account
for continuous cell states in addition to dis-
crete cell types (i.e. reflecting cell state tra-
jectories within cell types and smooth trans-
itions between cell types, as observed in tis-
sue 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 or-
ganisms 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 lev-
els of resolution are the structure-rich topolo-
gies generated by PAGA based on scRNA-seq [Wolf et al., 2019], see Figure 1 for an
illustration1. At the highest levels of resolu-
tion, these topologies also reflect intermedia-
te cell states and the developmental trajec-
tories passing through them. A similar ap-
proach that also allows for consistently zoom-
ing 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 [Ünen 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 pro-
vide sound theories about reasonable, contin-
uous 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 exper-
iments, because the latter are based on ex-
amining 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 reduc-
tion of the stability of the supporting signals.
The reduction in signal stability, in turn, im-
plies that data becomes substantially more

1Figure 1 was adapted from Wolf et al. [2019],
Fig. 3, provided under Creative Commons
Attribution 4.0 International License (http://
creativecommons.org/licenses/by/4.0/).
Figure 1: Different levels of resolution are of interest, depending on the research question and the data available. Thus, analysis tools and reference systems (such as cell atlases) will have to accommodate for multiple levels of resolution from whole organs and tissues over discrete cell types to continuously mappable intermediate cell states, indistinguishable even at the microscopic level. A graph abstraction that enables such multiple levels of focus is provided by PAGA [Wolf et al., 2019], a structure that allows for discretely grouping cells, as well as inferring trajectories as paths through a graph.

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

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

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
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.

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

- **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].

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]
population differences in

<table>
<thead>
<tr>
<th>mean</th>
<th>variability</th>
<th>pseudotime</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q</td>
<td>R</td>
<td>S</td>
</tr>
<tr>
<td>i</td>
<td>i</td>
<td>i</td>
</tr>
</tbody>
</table>

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

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

While some methods exist to identify more general patterns of gene expression changes (e.g. variability, distributions), these methods could be further improved by integrating with existing approaches that account for confounding effects such as cell cycle [Stegle et al., 2015] and complex batch effects [Butler et al., 2018a, Haghverdi et al., 2018]. Moreover, our capability to discover interesting gene expression patterns will be vastly expanded by connecting with other aspects of single-cell expression dynamics, such as cell type composition, RNA velocity [Manno et al., 2018], splicing and allele-specificity. This will allow us to fully exploit the granularity contained in single-cell level expression measurements.

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

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

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

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

3.3.1 Status

See Table 1 for a list of cell atlas type references that have recently been published. For human, similar 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 Maelcker, 2015, Weber and Robinson, 2016, Saeyes 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, Farrel 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-
<table>
<thead>
<tr>
<th>organism</th>
<th>scale of cell atlas</th>
<th>citation</th>
</tr>
</thead>
<tbody>
<tr>
<td>nematode</td>
<td>whole organism at larval stage L2</td>
<td>[Cao et al., 2017]</td>
</tr>
<tr>
<td><em>Caenorhabditis elegans</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>planaria</td>
<td>whole organism of the adult animal</td>
<td>[Fincher et al., 2018, Plass et al., 2018]</td>
</tr>
<tr>
<td><em>Schmidtea mediterranea</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fruit fly</td>
<td>whole organism at embryonic stage</td>
<td>[Karaiskos et al., 2017b]</td>
</tr>
<tr>
<td><em>Drosophila melanogaster</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zebrafish</td>
<td>whole organism at embryonic stage</td>
<td>[Farrell et al., 2018, Wagner et al., 2018a]</td>
</tr>
<tr>
<td>frog</td>
<td>whole organism at embryonic stage</td>
<td>[Briggs et al., 2018]</td>
</tr>
<tr>
<td><em>Xenopus tropicalis</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mouse</td>
<td>whole adult brain</td>
<td>[Rosenberg et al., 2018, Saunders et al., 2018, Zeisel et al., 2018]</td>
</tr>
<tr>
<td>Mouse</td>
<td>whole adult organism</td>
<td>[Tabula Muris Consortium, 2018, Han et al., 2018]</td>
</tr>
</tbody>
</table>

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

3.4 Challenge IV: Generalizing trajectory inference

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

### 3.4.1 Status

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

### 3.4.2 Open problems

Potentially, many of the above-mentioned methods for trajectory inference can be extended to data obtained with non-transcriptomic assays. Thereby, the following aspects are crucial. First, it is necessary to define the features to use; while for transcriptomic data the features are well 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 chro-
matin 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 cor-
relation methods have been used to detect aggregation of proteins [Shivanandan et al., 2016]. Shah et al. [2016] use seqFISH to measure 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 het-
erogeneity and understanding its dynamics are crucial for improving diagnosis and therapeutic choices (Figure 4).

Classic bulk sequencing data of tumor samples taken during surgery are always a mixture of tumor and normal cells (including e.g. invading immune cells). This means that disentangling mutational profiles of 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 pre-amplification PCR (PEP-PCR/-I-PEP-PCR) [Zhang et al., 1992, Arneson et al., 2008] and others. They require thermostable polymerases that withstand all temperatures during the cycling. More recent MDA-based technologies use the strand-displacing, high-fidelity 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 (CNVs) 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 (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 MAL-BAC) 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
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 SC-caller [Dong et al., 2017]. SC-caller 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 Genotypyer [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].

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, Beerli 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 Sci\( \Phi \) [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 avail-
able. 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 learn-
ing 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 timepoints 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 [Ramala 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.
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 [Szczerk et al., 2013, Jerby-Arnon et al., 2014], and patient survival [Matlak and Szczerk, 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 [Tujrajlic 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.
<table>
<thead>
<tr>
<th>Integration</th>
<th>example MTs</th>
<th>example AMs</th>
<th>Promises</th>
<th>Challenges</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 none</td>
<td>scDNA-seq,</td>
<td>clustering /</td>
<td>identify new cell</td>
<td>technical noise</td>
</tr>
<tr>
<td></td>
<td>scRNA-seq,</td>
<td>unsupervised</td>
<td>types and states</td>
<td></td>
</tr>
<tr>
<td></td>
<td>merFISH</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 within 1 MT,</td>
<td>scDNA-seq,</td>
<td>differential</td>
<td>identify effects</td>
<td>technical noise; batch effects; validate cell type assignments</td>
</tr>
<tr>
<td>within 1 exp,</td>
<td>scRNA-seq,</td>
<td>analyses,</td>
<td>across sample groups, time and space</td>
<td></td>
</tr>
<tr>
<td>across &gt; 1 smps</td>
<td>merFISH</td>
<td>time series,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>spatial sampling</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 within 1 MT,</td>
<td>scRNA-seq,</td>
<td>map cells to</td>
<td>accelerate analyses; increase sample size &amp; generalize observations</td>
<td>technical noise; batch effects; validate cell type assignments; standards across experimental centers</td>
</tr>
<tr>
<td>across &gt; 1 exp,</td>
<td>merFISH</td>
<td>stable reference</td>
<td></td>
<td></td>
</tr>
<tr>
<td>across &gt; 1 smps,</td>
<td></td>
<td>(cell atlas)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 across &gt; 1 MTs,</td>
<td>scG&amp;T-seq,</td>
<td>MOFA, DIABLO, MINT</td>
<td>holistic view of biological processes within cell; quantification of dependency of MTs</td>
<td>scaling cell throughput; MT combinations limited; dependency of MTs; data sparsity</td>
</tr>
<tr>
<td>within 1 exp,</td>
<td>scM&amp;T-seq,</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>within 1 cell</td>
<td>seqFISH</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 across &gt; 1 MTs,</td>
<td>scDNA-seq +</td>
<td>Cardelino, Clonealign, MATCHER</td>
<td>use existing datasets (faster than 3); flexible experimental design</td>
<td>technical noise; validate cell / data grouping; test assumptions for integrating data</td>
</tr>
<tr>
<td>within 1 exp,</td>
<td>scRNA-seq,</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>across &gt; 1 cells,</td>
<td>DNA-seq +</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>within 1 cell pop</td>
<td>scRNA-seq</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 across &gt; 1 MTs,</td>
<td>hypothetical:</td>
<td>hypothetical: multi-omic HCA, single-cell TCGA</td>
<td>comprehensive characterizations of biological systems</td>
<td>all from approaches 2, 3 &amp; 4; standards across experimental centers</td>
</tr>
<tr>
<td>across &gt; 1 exps,</td>
<td>any combina-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>across &gt; 1 smps,</td>
<td>tion</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>within cells</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.

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 concurrent 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.

---

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

### 6.1.1 Status

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

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

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

Evaluation of tool performance requires benchmarking datasets with known ground truth. Such data should include cell populations with known genomic compositions and population structures, i.e. where frequencies of clones and alleles are known. Currently, such datasets are scarce—because generating them in genuine laboratory settings is time-, labor- and cost-intensive. Experimental benchmark datasets for evolutionary analysis of single-cell populations are even harder to obtain, as they require follow-up samples with known information about evolutionary trajectories and developmental times. With lack of time-resolved measurements, only anecdotal evidence exists on, for instance, how the accuracy of phylogenetic inferences is affected by data quality. Availability of such gold-standard datasets would benefit single-cell genomics research enormously.

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

6.2.1 Status

Recent studies [Soneson and Robinson, 2018, Saelens et al., 2019] show that systematic benchmarking of different single-cell analysis methodologies has begun. However, to the best of our knowledge, there is still a shortage of single-cell data simulation tools. Many single-cell data analysis packages include their own ad hoc data simulators [Vallejos et al., 2015, Korthauer et al., 2016a, Lun et al., 2016, Lun and Marioni, 2017, Jahn et al., 2016, Satas and Raphael, 2018, Rizzetto et al., 2017, Köster et al., 2017]. However, these simulators are usually not available as separate tools or even as a source code, tailored to specific
problems studied in corresponding papers and sometimes not comprehensively documented, thus limiting their utility for the broad research community. Furthermore, since such simulators are used only as auxiliary subroutines inside particular projects and are not published as stand-alone tools, they themselves are usually not evaluated, and therefore the accuracy of their reflection of real biological and technological processes remain unclear. There are few exceptions known to us, including the tools Splatter [Zappia et al., 2017], powsimR [Vieth et al., 2017], and SymSim [Zhang et al., 2019d], which provide frameworks for simulation of scRNA-seq data and whose accuracy has been validated by comparison of its results with real data. For single-cell phylogenomics, cancer genome evolution simulators are being designed [Smeraro 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-
This approach has been very successful, e.g. in protein structure prediction and metagenomic analyses. A first step in this direction was the recent single-cell transcriptomics DREAM challenge.

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


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...


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.


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


Niko Beerenwinkel, Tibor Antal, David Dingli, Arne Traulsen, Kenneth W Kinzler, Victor E Velculescu, Bert Vogelstein, and Martin A Nowak. Genetic progression and


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


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 chro-


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

52


Yong Hou, Kui Wu, Xulian Shi, Fuqiang Li, Luting Song, Hanjie Wu, Michael Dean, Guibo Li, Shirley Tsang, Runze Jiang, Xiaulong 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, Xin 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.


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


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,


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


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


Roberto Semeraro, Valerio Orlandini, and Alberto Magi. Xome-Blender: A novel can-


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


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.


