Eleven grand challenges in single-cell data science

Lähnemann, David; Köster, Johannes; Robinson, Mark D.; Vallejos, Catalina A.; Campbell, Kieran R.; Beerenwinkel, Niko; Pinello, Luca; Lelieveldt, Boudewijn P.F.; Reinders, Marcel; More Authors

DOI
10.1186/s13059-020-1926-6

Publication date
2020

Document Version
Final published version

Published in
Genome biology

Citation (APA)
Lähnemann, D., Köster, J., Robinson, M. D., Vallejos, C. A., Campbell, K. R., Beerenwinkel, N., Pinello, L., Lelieveldt, B. P. F., Reinders, M., & More Authors (2020). Eleven grand challenges in single-cell data science. *Genome biology*, *21*(1), Article 31. https://doi.org/10.1186/s13059-020-1926-6

Lähnemann *et al. Genome Biology* (2020) 21:31
https://doi.org/10.1186/s13059-020-1926-6

Genome Biology

REVIEW Open Access

Eleven grand challenges in single-cell data science

David Lähnemann[1,2,3], Johannes Köster[1,4], Ewa Szczurek[5], Davis J. McCarthy[6,7], Stephanie C. Hicks[8], Mark D. Robinson[9], Catalina A. Vallejos[10,11], Kieran R. Campbell[12,13,14], Niko Beerenwinkel[15,16], Ahmed Mahfouz[17,18], Luca Pinello[19,20,21], Pavel Skums[22], Alexandros Stamatakis[23,24], Camille Stephan-Otto Attolini[25], Samuel Aparicio[13,26], Jasmijn Baaijens[27], Marleen Balvert[27,28], Buys de Barbanson[29,30,31], Antonio Cappuccio[32], Giacomo Corleone[33], Bas E. Dutilh[28,34], Maria Florescu[29,30,31], Victor Guryev[35], Rens Holmer[36], Katharina Jahn[15,16], Thamar Jessurun Lobo[35], Emma M. Keizer[37], Indu Khatri[38], Szymon M. Kielbasa[39], Jan O. Korbel[40], Alexey M. Kozlov[23], Tzu-Hao Kuo[3], Boudewijn P.F. Lelieveldt[41,42], Ion I. Mandoiu[43], John C. Marioni[44,45,46], Tobias Marschall[47,48], Felix Mölder[1,49], Amir Niknejad[50,51], Lukasz Raczkowski[5], Marcel Reinders[17,18], Jeroen de Ridder[29,30], Antoine-Emmanuel Saliba[52], Antonios Somarakis[42], Oliver Stegle[40,46,53], Fabian J. Theis[54], Huan Yang[55], Alex Zelikovsky[56,57], Alice C. McHardy[3], Benjamin J. Raphael[58], Sohrab P. Shah[59] and Alexander Schönhuth[27,28*]

Abstract

The recent boom in microfluidics and combinatorial indexing strategies, combined with low sequencing costs, has empowered single-cell sequencing technology. Thousands—or even millions—of cells analyzed in a single experiment amount to a data revolution in single-cell biology and pose unique data science problems. Here, we outline eleven challenges that will be central to bringing this emerging field of single-cell data science forward. For each challenge, we highlight motivating research questions, review prior work, and formulate open problems. This compendium is for established researchers, newcomers, and students alike, highlighting interesting and rewarding problems for the coming years.

Introduction

Since being highlighted as "Method of the Year" in 2013 [1], 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) enables transcriptome-wide gene expression measurement at single-cell resolution, allowing for cell type clusters to be distinguished (for an early example, see [2]), the arrangement of populations of cells according to novel hierarchies, and the identification of cells transitioning between states. This can lead to a much clearer view of the dynamics of tissue and organism development, and on structures within cell populations that had so far been perceived as homogeneous. In a similar vein, analyses based on single-cell DNA sequencing (scDNA-seq) can highlight somatic clonal structures (e.g., in cancer, see [3, 4]), thus helping to track the formation of cell lineages and provide insight into evolutionary processes acting on somatic mutations.

*Correspondence: as@cwi.nl
Johannes Köster, Alice C. McHardy, Benjamin J. Raphael, Sohrab P. Shah, and Alexander Schönhuth are joint last authors and workshop organizers.
David Lähnemann, Johannes Köster, Ewa Szczurek, Davis J. McCarthy, Stephanie C. Hicks, Mark D. Robinson, Catalina A. Vallejos, Kieran R. Campbell, Niko Beerenwinkel, Ahmed Mahfouz, Luca Pinello, Pavel Skums, Alexandros Stamatakis, Camille Stephan-Otto Attolini, and Alexander Schönhuth are joint first authors and have major contributions to the manuscript.
[27]Life Sciences and Health, Centrum Wiskunde & Informatica, Amsterdam, The Netherlands
[28]Theoretical Biology and Bioinformatics, Science for Life, Utrecht University, Utrecht, The Netherlands
Full list of author information is available at the end of the article

The opportunities arising from single-cell sequencing (sc-seq) are enormous: only now is it possible to re-evaluate hypotheses about differences between predefined sample groups at the single-cell level—no matter if such sample groups are disease subtypes, treatment groups, or simply morphologically distinct cell types. It is therefore no surprise that enthusiasm about the possibility to screen the genetic material of the basic units of life has continued to grow. A prominent example is the Human Cell Atlas [5], an initiative aiming to map the numerous cell types and states comprising a human being.

Encouraged by the great potential of investigating DNA and RNA at the single-cell level, the development of the corresponding experimental technologies has experienced considerable growth. In particular, the emergence of microfluidics techniques and combinatorial indexing strategies [6–10] has led to hundreds of thousands of cells routinely being sequenced in one experiment. This development has even enabled a recent publication analyzing millions of cells at once [11]. Sc-seq datasets comprising very large cell numbers are becoming available worldwide, constituting a data revolution for the field of single-cell analysis.

These vast quantities of data and the research hypotheses that motivate them need to be handled in a computationally efficient and statistically sound manner [12]. As these aspects clearly match a recent definition of "Data Science" [13], we posit that we have entered the era of single-cell data science (SCDS).

SCDS exacerbates many of the data science issues arising in bulk sequencing, but it also constitutes a set of new, unique challenges for the SCDS community to tackle. Limited amounts of material available per cell lead to high levels of uncertainty about observations. When amplification is used to generate more material, technical noise is added to the resulting data. Further, any increase in resolution results in another—rapidly growing—dimension in data matrices, calling for scalable data analysis models and methods. Finally, no matter how varied the challenges are—by research goal, tissue analyzed, experimental setup, or just by whether DNA or RNA is sequenced—they are all rooted in data science, i.e., are computational or statistical in nature. Here, we propose the data science challenges that we believe to be among the most relevant for bringing SCDS forward.

This catalog of SCDS challenges aims at focusing the development of data analysis methods and the directions of research in this rapidly evolving field. It shall serve as a compendium for researchers of various communities, looking for rewarding problems that match their personal expertise and interests. To make it accessible to these different communities, we categorize challenges into the following: transcriptomics (see "Challenges in single-cell transcriptomics"), genomics (see the "Challenges in single-cell genomics"), and phylogenomics (see "Challenges in single-cell phylogenomics"). For each challenge, we provide a thorough review of the status relative to existing approaches and point to possible directions of research to solve it.

Several themes and aspects recur across the boundaries of research communities and methodological approaches. We represent these overlaps in three different ways. First, we decided to discuss some problems in multiple contexts, highlighting the relevant aspects for the respective research communities (e.g., data sparsity in transcriptomics and genomics). Second, we separately introduce recurring themes (see "Single-cell data science: recurring themes"), thereby keeping respective discussions in each challenge succinct. Third, if challenges were identified as independent of the chosen categorization, they are discussed as recapitulatory challenges at the end (see "Overarching challenges").

Single-cell data science: recurring themes

A number of challenging themes are common to many or all single-cell analyses, regardless of the particular assay or data modality generated. We will start our review by introducing them. Later, when discussing the specific challenges, we will refer to these broader themes wherever appropriate and outline what they 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.

The themes may reflect issues one also experiences when analyzing bulk sequencing data. However, even if not unique to single-cell experiments, these issues may dominate the analysis of sc-seq data and therefore require particular attention. The two most urgent elementary themes, not necessarily unique to sc-seq, are the need to quantify measurement uncertainty (see "Quantifying uncertainty of measurements and analysis results") and the need to benchmark methods systematically, in a way that highlights the metrics that are particularly critical in sc-seq. Since the latter is of central importance and an aspect that has gained visibility only recently, we not only mention its importance in relevant challenges, but also consider it a challenge in its own right (see "Challenge XI: Validating and benchmarking analysis tools for single-cell measurements").

We identify three sweeping themes that are more specific to sc-seq, exacerbated by the rapid advances in experimental technologies. First, there is a need to scale to higher dimensional data, be it more cells measured or more data measured per cell (see "Scaling to higher dimensionalities: more cells, more features, and broader coverage"). This need often arises in combination with a second one: the need to integrate data across different

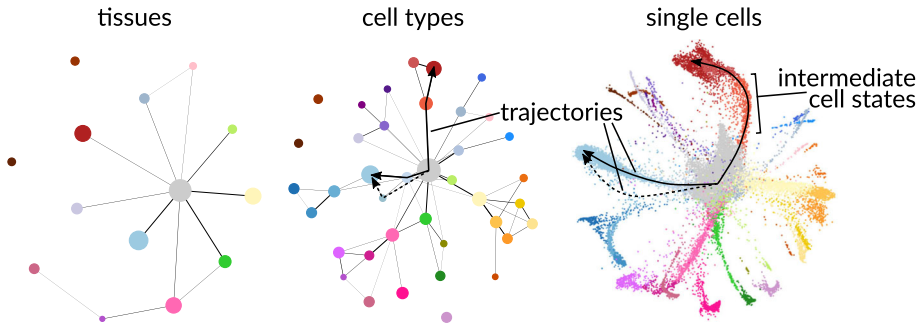

Fig. 1 Different levels of resolution are of interest, depending on the research question and the data available. Thus, analysis tools and reference systems (such as cell atlases) will have to accommodate multiple levels of resolution from whole organs and tissues over discrete cell types to continuously mappable intermediate cell states, which are indistinguishable even at the microscopic level. A graph abstraction that enables such multiple levels of focus is provided by PAGA [14], a structure that allows for discretely grouping cells, as well as inferring trajectories as paths through a graph

types of single-cell measurements (e.g., RNA, DNA, proteins, and methylation) and across samples, be it from different time points, treatment groups, or even organisms. This integration theme runs throughout multiple challenges and is so central that we consider it a challenge worth highlighting (see "Challenge X: Integration of single-cell data across samples, experiments, and types of measurement"). Third, the possibility to operate on the finest levels of resolution casts an important, overarching question: what level of resolution is appropriate relative to the particular research question one has in mind (see "Varying levels of resolution")? We will start by qualifying this last one.

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 those supported by bulk sequencing experiments. However, even though sc-seq operates at the most basic level, mapping cell types and states at a particular level of resolution of interest may be challenging: Achieving the targeted level of resolution or granularity for the intended map of cells may require substantial methodological efforts and will depend on whether the research question allows for a certain freedom in terms of resolution and on the limits imposed by the particular experimental setup.

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

An exemplary illustration of how maps of cell types and states can support different levels of resolution is the structure-rich topologies generated by PAGA based on scRNA-seq [14], see Fig. 1[1]. At the highest levels of resolution, these topologies also reflect intermediate cell states and the developmental trajectories passing through them. A similar approach that also allows for consistently zooming into more detailed levels of resolution is provided by hierarchical stochastic neighbor embedding (HSNE, Pezzotti et al. [15]), a method pioneered on mass cytometry datasets [16, 17]. In addition, manifold learning [18, 19] and metric learning [20, 21] may provide further theoretical support for even more accurate maps, because they provide sound theories about reasonable, continuous distance metrics, instead of just distinct, discrete clusters.

Quantifying uncertainty of measurements and analysis results

The amount of material sampled from single cells is considerably less than that used in bulk experiments. Signals become more stable when individual signals are summarized (such as in a bulk experiment); thus, the increase in resolution due to sc-seq also means a reduction of the stability of the supporting signals. The reduction in signal stability, in turn, implies that data becomes substantially more uncertain and tasks so far considered routine, such as single nucleotide variation (SNV) calling in bulk sequencing, require considerable methodological care with sc-seq data.

[1] Figure 1 was adapted from [14], Fig. 3, provided under Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/).

These issues with data quality and in particular missing data pose challenges that are unique to sc-seq, and are thus at the core of several challenges: regarding scDNA-seq data quality (see "Challenges in single-cell genomics") and especially regarding missing data in scDNA-seq ("Challenge VI: Dealing with errors and missing data in the identification of variation from single-cell DNA sequencing data") and scRNA-seq ("Challenge I: Handling sparsity in single-cell RNA sequencing"). In contrast, the non-negligible batch effects that scRNA-seq can suffer from reflect a common problem in high-throughput data analysis [22], 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, see, for example, BEARscc [23]).

Optimally, sc-seq analysis tools would accurately quantify all uncertainties arising from experimental errors and biases. Such 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.

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

The current blossoming of experimental methods poses considerable statistical challenges, and would do so 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. Some scRNA-seq SCDS methodology has started to address scalability [12, 24–27], but the respective issues have not been fully resolved and experimental methodology will scale further. For scDNA-seq, experimental methodology has just been scaling up to more cells recently (see Table 1 and "Challenge VII: Scaling phylogenetic models to many cells and many sites"), 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 [37, 38], DNA methylation [39], protein levels [40], protein binding, and also for performing multiple simultaneous measurements [41, 42] 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 "Challenge VIII: Integrating multiple types of variation into phylogenetic models"). Also, with spatial and temporal sampling becoming available (see "Challenge V: Finding patterns in spatially resolved measurements" and "Challenge IX: Inferring population genetic parameters of tumor heterogeneity by model integration"), data integration methods need to scale to more and new types of context information for individual cells (see "Challenge X: Integration of single-cell data across samples, experiments, and types of measurement" for a comprehensive discussion of data integration approaches).

Challenges in single-cell transcriptomics

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

Table 1 Whole genome amplification: recent improvements

Recent improvements of whole genome amplification (WGA) methods promise to reduce amplification biases and errors, while scaling throughput to larger cell numbers:

1. Improved coverage uniformity for multiple displacement amplification (MDA) has been achieved using droplet microfluidics-based methods (eWGA [28]; sd-MDA [29]; ddMDA [30]). A second approach has been to couple the Φ29 DNA polymerase to a primase to reduce priming bias [31].
2. One way to reduce the amplification error rate of the polymerase chain reaction (PCR)-based methods (including multiple annealing and looping-based amplification cycles (MALBAC)) would be to employ a thermostable polymerase (necessary for use in PCR) with proof-reading activity similar to Φ29 DNA polymerase, but we are not aware of any PCR DNA polymerases with a fidelity in the range of Φ29 DNA polymerase [32].
3. Three newer methods use an entirely different approach: they randomly insert transposons into the whole genome and then leverage these as priming sites for amplification and library preparation. Transposon Barcoded (TnBC) library preparation (with a PCR amplification, [33]) and direct library preparation (DLP) (with a shallow library without any amplification, [34]) allow only for copy number variation (CNV) calling, but DLP scales up to 80,000 single cells [35]. Linear—as opposed to exponential—Amplification via Transposon Insertion (LIANTI, [36]) also addresses amplification errors: all copies are generated based on the original genomic DNA through in vitro transcription. With errors unique to individual barcoded copies, the authors report a false positive rate that is even lower than for MDA [36].

has no unique molecular identifiers or reads mapping to it. The term "dropout" is often used to denote observed zero values in scRNA-seq data. But this term usually conflates two distinct types of zero values: those attributable to methodological noise, where a gene is expressed but not detected by the sequencing technology, and those attributable to biologically-true absence of expression. Thus, we recommend against the term "dropout" as a catch-all term for observed zeros. Beyond biological variation in the number of unexpressed genes, the proportion of observed zeros, or degree of sparsity, is attributed to technical limitations [43, 44]. Those can result in artificial zeros that are either 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, due to sampling variation—will sometimes be detected and sometimes not). Accordingly, the degree of sparsity depends on the scRNA-seq platform used, the sequencing depth, and the underlying expression level of the gene.

Sparsity in scRNA-seq data can hinder downstream analyses and is still challenging to model or handle appropriately, calling for further method development. Sparsity pervades all aspects of scRNA-seq data analysis, but in this challenge, we focus on the linked problems of learning latent spaces and "imputing" expression values from scRNA-seq data (Fig. 2). Imputation approaches are closely linked to the challenges of normalization. But whereas normalization generally aims to make expression values between cells and experiments more comparable to each other, imputation approaches aim to achieve adjusted data values that 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.

Status

The imputation of missing values has been very successful for genotype data [45]. Crucially, when imputing genotypes, we typically know which data are missing (e.g., when no genotype call is possible due to no coverage of a locus; although see the "Challenge VI: Dealing with errors and missing data in the identification of variation from single-cell DNA sequencing data" for the challenges with scDNA-seq data). In addition, 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 [46].

The situation is somewhat different for scRNA-seq data, as we do not routinely have external reference information to apply (see "Challenge III: Mapping single cells to a reference atlas"). In addition, we can never be sure which of the observed zeros represent "missing data" and which accurately represent a true absence of gene expression in the cell [43].

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 (i.e., quantifying uncertainty, see the "Quantifying uncertainty of measurements and analysis results"), 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 (Fig. 2). We prefer to use the first option where possible, and for many single-cell data analysis problems, there already are statistical models appropriate for sparse count data that should be used or extended (e.g., for differential expression analysis, see the "Challenge II: Defining flexible statistical frameworks for discovering complex differential patterns in gene expression"). 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, depending on the amount of sparsity, imputation could potentially improve results of dimension reduction, visualization, and clustering applications.

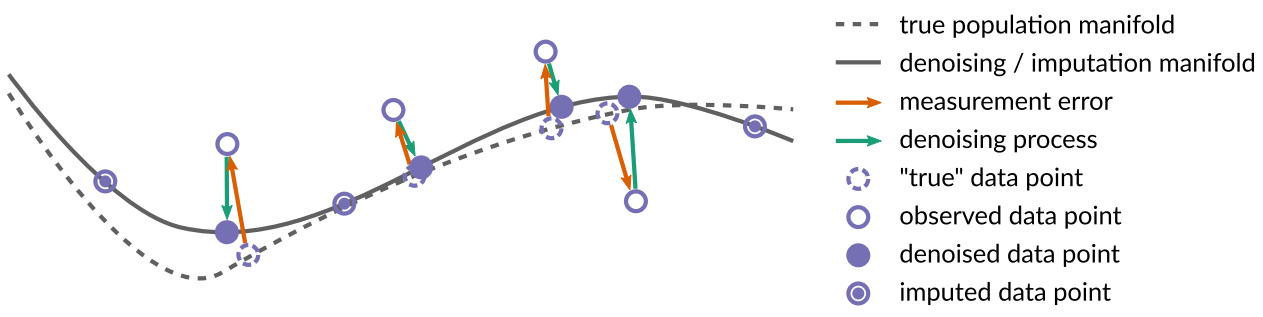

Fig. 2 Measurement error requires denoising methods or approaches that quantify uncertainty and propagate it down analysis pipelines. Where methods cannot deal with abundant missing values, imputation approaches may be useful. While the true population manifold that generated data is never known, one can usually obtain some estimation of it that can be used for both denoising and imputation

We define three broad (and often overlapping) categories of methods that can be used to "impute" scRNA-seq data in the absence of an external reference (Table 2): (A) *Model-based imputation methods* of technical zeros use probabilistic models to identify which observed zeros represent technical rather than biological zeros. They aim to impute expression levels only for the technical zeros, leaving other observed expression levels untouched. (B) *Data-smoothing methods* define a "similarity" between 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 usually adjust all expression values, including technical zeros, biological zeros, and observed non-zero values. (C) *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). 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 (or estimated latent spaces if using, for example, variational autoencoders) can be used for downstream applications such as clustering or trajectory inference (see "Challenge IV: Generalizing trajectory inference"). A fourth—distinct—category is (T) imputation with an external dataset or reference, using it for transfer learning.

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 probabilistically determine 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 (Table 2 (A)).

Data-smoothing methods adjust all gene expression levels based on expression levels in "similar" cells, aiming to "denoise" the values (Fig. 2). Several such methods have been proposed to handle imputation problems (Table 2 (B)). To take a simplified example (Fig. 2), we might imagine that single cells originally refer to points along a curve across a two-dimensional space. 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).

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 a third route for imputation: they can *reconstruct* the observed data matrix from simplified representations of it.

Principal component analysis (PCA) is one 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 (Table 2 (C)), with some specifically intended for imputation (ALRA, ENHANCE, scRMD).

Additionally, machine learning methods have been proposed for scRNA-seq data analysis that can, but need not, use probabilistic data generative processes to capture low-dimensional or latent space representations of a dataset (Table 2 (C)). Some of them are expressly aimed at imputation (e.g., AutoImpute, DeepImpute, EnImpute, DCA, and scVI). But even if imputation is not the 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).

Finally, a small number of scRNA-seq imputation methods extend approaches from any (combination) of the three categories above by incorporating information external to the current dataset (Table 2 (T)). Approaches using cell atlas-type reference resources are further discussed in the "Challenge III: Mapping single cells to a reference atlas" section and classified as approach `+X+S` in the "Challenge X: Integration of single-cell data across samples, experiments, and types of measurement" (see Fig. 6 and Table 4).

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 or cells. In turn, this can introduce false positives in downstream analyses such as differential expression testing and gene network inference [90]. Handling batch effects and potential

Table 2 Short description of methods for the imputation of missing data in scRNA-seq data

A: model-based imputation		
bayNorm	Binomial model, empirical Bayes prior	[47]
BISCUIT	Gaussian model of log counts, cell- and cluster-specific parameters	[48]
CIDR	Decreasing logistic model (DO), non-linear least-squares regression (imp)	[49]
SAVER	NB model, Poisson LASSO regression prior	[50]
ScImpute	Mixture model (DO), non-negative least squares regression (imp)	[51]
scRecover	ZINB model (DO identification only)	[52]
VIPER	Sparse non-negative regression model	[53]
B: data smoothing		
DrImpute	*k*-means clustering of PCs of correlation matrix	[54]
knn-smooth	*k*-nearest neighbor smoothing	[55]
LSImpute	Locality sensitive imputation	[56]
MAGIC	Diffusion across nearest neighbor graph	[57]
netSmooth	Diffusion across PPI network	[58]
C: data reconstruction, matrix factorization		
ALRA	SVD with adaptive thresholding	[59]
ENHANCE	Denoising PCA with aggregation step	[60]
scRMD	Robust matrix decomposition	[61]
consensus NMF	Meta-analysis approach to NMF	[62]
f-scLVM	Sparse Bayesian latent variable model	[63]
GPLVM	Gaussian process latent variable model	[64]
pCMF	Probab. count matrix factorization with Poisson model	[65]
scCoGAPS	Extension of NMF	[66]
SDA	Sparse decomposition of arrays (Bayesian)	[67]
ZIFA	ZI factor analysis	[68]
ZINB-WaVE	ZINB factor model	[69]
C: data reconstruction, machine learning		
AutoImpute	AE, no error back-propagation for zero counts	[70]
BERMUDA	AE for cluster batch correction (MMD and MSE loss function)	[71]
DeepImpute	AE, parallelized on gene subsets	[72]
DCA	Deep count AE (ZINB / NB model)	[73]
DUSC / DAWN	Denoising AE (PCA determines hidden layer size)	[74]
EnImpute	Ensemble learning consensus of other tools	[75]
Expression Saliency	AE (Poisson negative log-likelihood loss function)	[76]
LATE	Non-zero value AE (MSE loss function)	[77]
Lin_DAE	Denoising AE (imputation across *k*-nearest neighbor genes)	[78]
SAUCIE	AE (MMD loss function)	[79]
scScope	Iterative AE	[80]
scVAE	Gaussian-mixture VAE (NB / ZINB / ZIP model)	[81]
scVI	VAE (ZINB model)	[82]
scvis	VAE (objective function based on latent variable model and t-SNE)	[83]
VASC	VAE (denoising layer; ZI layer, double-exponential and Gumbel distribution)	[84]
Zhang_VAE	VAE (MMD loss function)	[85]
T: using external information		
ADImpute	Gene regulatory network information	[86]
netSmooth	PPI network information	[58]
SAVER-X	Transfer learning with atlas-type resources	[87]
SCRABBLE	Matched bulk RNA-seq data	[88]
TRANSLATE	Transfer learning with atlas-type resources	[77]
URSM	Matched bulk RNA-seq data	[89]

Imputation methods using only data from within a dataset are roughly categorized approaches A (model-based), B (data smoothing), and C (data reconstruction), with the latter further differentiated into matrix factorization and machine learning approaches. In contrast to these methods, those in category T (for transfer learning) also use information external to the dataset to be analyzed

AE autoencoder, *DO* dropout, *imp* imputation, *MMD* maximum mean discrepancy, *MSE* mean squared error, *NB* negative binomial, *NMF* non-negative matrix factorization, *P* Poisson, *PC* principal component, *PCA* principal component analysis, *PPI* protein-protein interaction, *SVD* singular value decomposition, *VAE* variational autoencoder, *ZI* zero-inflated

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 "Quantifying uncertainty of measurements and analysis results"). Approaches that allow quantification and propagation of the uncertainties associated with expression measurements (see "Quantifying uncertainty of measurements and analysis results") may help to avoid problems associated with "overimputation" and the introduction of spurious signals noted by Andrews and Hemberg [90].

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., [5]) could provide a valuable resource for this purpose. Some methods have been extended to allow the use of such resources (e.g., SAVER-X and TRANSLATE), but this will need to be done for all approaches (see "Challenge III: Mapping single cells to a reference atlas").

A second approach to avoid circularity is the systematic integration of known biological network structures in the imputation process. This can be achieved by encoding network structure knowledge as prior information, as proposed by ADImpute and netSmooth and the tool by Lin et al. [78].

Finally, a third way of avoiding circularity in imputation 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. Of course, such orthogonal information can also be provided by different types of molecular measurements (see "Challenge X: Integration of single-cell data across samples, experiments, and types of measurement"). Methods designed to integrate multi-omics data could then be extended to enable scRNA-seq imputation, for example, through generative models that explicitly link scRNA-seq with other data types (e.g., clonealign [91]) or by inferring a shared low-dimensional latent structure (e.g., MOFA [92]) 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 (see "Challenge XI: Validating and benchmarking analysis tools for single-cell measurements"). Early attempts by Zhang and Zhang [93] and Andrews and Hemberg [90] 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. Some methods, especially those using deep learning, depend strongly on choice of hyperparameters [94]. There, more detailed comparisons that explore parameter spaces would be helpful, extending work like that from Sun et al. [95] comparing dimensionality reduction methods. Such detailed benchmarking would also help to establish when normalization methods derived from explicit count models (e.g., [96, 97]) may be preferable to imputation.

Finally, scalability for large numbers of cells remains an ongoing concern for methods allowing for imputation, as for all high-throughput single-cell methods and software (see "Scaling to higher dimensionalities: more cells, more features, and broader coverage").

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 (Fig. 6, approach `+S`). Further understanding of gene expression changes will enable deeper knowledge across a myriad of applications, such as immune responses [98, 99], development [100], and drug responses [101].

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, current analysis pipelines typically 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 [102, 103], but it appears that single cell-specific methods do not uniformly outperform the state-of-the-art bulk methods [104]. Some attention has been given to more general patterns of differential expression (Fig. 3), such as changes in variability that account for mean expression confounding [105], changes in trajectory along pseudotime [106, 107], or more generally, changes in distributions [108]. Furthermore, methods for cross-sample comparisons of gene expression (e.g., cell type-specific changes in cell state across samples; see the "Challenge X: Integration of single-cell data across samples, experiments, and types of measurement", Fig. 6 and Table 4) are now emerging, such as pseudo-bulk analyses [109–111], where

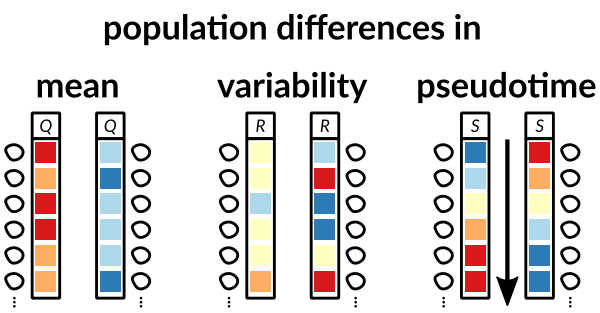

Fig. 3 Differential expression of a gene or transcript between cell populations. The top row labels the specific gene or transcript, as is also done in Fig. 6. A difference in *mean* gene expression manifests in a consistent difference of gene expression across all cells of a population (e.g., high vs. 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, for example, along a developmental trajectory (compare Fig. 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

expression is aggregated over multiple cells within each sample, or mixed models, where both within- and between-sample variation is captured [111, 112]. 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.

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 [113] 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 [114] could provide an alternative solution, with an approach based on correcting the selection bias recently proposed [115].

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 [116] and complex batch effects [117, 118]. 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 [119], 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 [120–124]. 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 or exploring changes in full distributions, while allowing for sample-to-sample variability and subpopulation-specific patterns [111].

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 within a cell type varies across different biological conditions (for differential expression analyses, see the "Challenge II: Defining flexible statistical frameworks for discovering complex differential patterns in gene expression" and data integration approach `+S` in Fig. 6). To put the results of such studies on a map, reliable reference systems with a resolution down to cell states are required—and depending on the research question at hand, even intermediate transition states might be of interest (see "Varying levels of resolution").

The lack of appropriate, available references has so far implied that only reference-free approaches were conceivable. Here, unsupervised clustering approaches were the predominant option (see data integration approach `1S` in Fig. 6). Method development for such unsupervised clustering of cells has already reached a certain level of maturity; for a systematic identification of available techniques, we refer to the respective reviews [125–127].

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 `+X+S` in Fig. 6). They will need to be able to embed new data points into a stable reference framework that allows for different levels of resolution and will have to eventually capture transitional cell states that fall in between clearly annotated cell clusters (see Fig. 1 for an idea of what cell atlas type reference systems could look like).

Table 3 Published cell atlases of whole tissues or whole organisms

Organism	Scale of cell atlas	Citation
Nematode (*Caenorhabditis elegans*)	Whole organism at larval stage L2	[128]
Planaria (*Schmidtea mediterranea*)	Whole organism of the adult animal	[129, 130]
Fruit fly (*Drosophila melanogaster*)	Whole organism at embryonic stage	[131]
Zebrafish	Whole organism at embryonic stage	[132, 133]
Frog (*Xenopus tropicalis*)	Whole organism at embryonic stage	[134]
Mouse	Whole adult brain	[135–137]
Mouse	Whole adult organism	[138, 139]

Status

See Table 3 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 [5]. Towards this end, initial consortia focus on specific organs, for example, the lung [140].

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 [141–147]. Also, the systematic benchmarking of this dynamic field of tools has begun [148]. A field that can serve as a source of further inspiration is flow/mass cytometry, where several methods already address the classification of high-dimensional cell type data [149–152].

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 [5, 140, 153]. Here, we focus on the mapping of cells or rather their molecular profiles onto stable existing reference atlases to further highlight the importance of these fundamental themes. A computationally and statistically sound method for mapping cells onto atlases for a range of conceivable research questions will need to (i) enable operation at various levels of resolution of interest, and also cover continuous, transient cell states (see "Varying levels of resolution"); (ii) quantify the uncertainty of a particular mapping of cells of unknown type/state (see "Quantifying uncertainty of measurements and analysis results"); (iii) scale to ever more cells and broader coverage of types and states (see "Scaling to higher dimensionalities: more cells, more features, and broader coverage"); and (iv) eventually integrate information generated not only through scRNA-seq experiments, but also through other types of measurements, for example, scDNA-seq or protein expression data (see "Challenge X: Integration of single-cell data across samples, experiments, and types of measurement" for a discussion of data integration, especially approaches `+M+C` and `+all` in Fig. 6).

Finally, for further benchmarking of methods that map cells of unknown type or state onto reference atlases (see "Challenge XI: Validating and benchmarking analysis tools for single-cell measurements" for benchmarking in general), atlases of model organisms where full lineages of cells have been determined can form the basis [129, 130, 132, 134, 154]. Importantly, additional information available from lineage tracing of such simpler organisms can provide a cross-check with respect to the transcriptome profile-based classification [134, 155].

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 ([156] and Fig. 1), and the ordering induced by this path is called pseudotime. Several models have been proposed to describe cell state dynamics starting from transcriptomic data [157]. 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 "Challenge X: Integration of single-cell data across samples, experiments, and types of measurement"), is still at its infancy. We believe the study of complex trajectories integrating different data types, especially epigenetics and proteomics information in addition to transcriptomics data, will lead to a more systematic understanding of the processes determining cell fate.

Status

Trajectory 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 [156, 158], principal curve or graph fitting [159–161], or random walks and diffusion operations on graphs ([162, 163] among others) are used to infer pseudotime and/or

branching trajectories. Alternative probabilistic descriptions can be obtained using optimal transport analysis [164] or approximation of the Fokker-Planck equations [165] or by estimating pseudotime through dimensionality reduction with a Gaussian process latent variable model [166–168].

Open problems

Many of the abovementioned methods for trajectory inference can be extended to data obtained with non-transcriptomic assays. For this, the following aspects are crucial. First, it is necessary to define the features to use. For transcriptomic data, the features are well annotated and correspond to expression levels of genes. In contrast, 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 [169–171]. 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 into account. An alternative approach consists of pooling/combining information from several cells or data imputation (see "Challenge I: Handling sparsity in single-cell RNA sequencing"). For example, imputation has been used for single-cell DNA methylation [172], aggregation over chromatin accessibility peaks from bulk or pseudo-bulk sample [173], and k-mer-based approaches have been proposed [160, 174, 175]. However, so far, no systematic evaluation (see "Challenge XI: Validating and benchmarking analysis tools for single-cell measurements") 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 various trajectory inference methods perform on different data types and importantly to define metrics that are suitable. Also, it is necessary to reason on the ground truth or propose reasonable surrogates (e.g., previous knowledge about developmental processes). Some recent papers explore this idea using scATAC-seq data, an assay to measure chromatin accessibility [160, 174, 176].

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 could be extended [118, 177, 178]. A related unsolved problem is that of comparing different trajectories obtained from the same data type but across individuals or conditions, in order to highlight unique and common aspects.

Challenge V: Finding patterns in spatially resolved measurements

Single-cell spatial transcriptomics or proteomics [179–181] 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 `+M1C` in "Challenge X: Integration of single-cell data across samples, experiments, and types of measurement", Fig. 6 and Table 4). 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 [182].

Status

Experimental approaches have been tailored either to 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 [183]. Whole tissues can be decomposed using the Niche-seq approach [184]: 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 beads with known barcodes to dissolve corresponding slide sites and sequence them with the respective barcodes [185]. Ultimately, one would like to sequence inside a tissue without dissociating the cells and without compromising on the unbiased nature of scRNA-seq. First approaches aiming to implement sequencing by synthesis in situ were proposed by Ke et al. [186] and Lee et al. [187], the latter being referred to as FISSEQ (Fluorescent in situ sequencing). Recently, starMAP [188] was presented. Here, RNA within an intact 3D tissue can be amplified and transferred into a hydrogel. Within the hydrogel, amplified DNA barcodes can be sequenced in situ, in order to distinguish RNA species while retaining spatial coordinates. Instead of performing a direct identification of (parts of) the RNA sequence, fluorescent in situ hybridization (FISH)-based methods require to design probes for targeting RNA species of interest. When multiplexing several rounds of FISH in combination with designed barcodes for each RNA species, it becomes possible to measure hundreds to thousands of RNA species simultaneously. Lubeck et al. [189] have shown a first approach of multiplexed, barcoded FISH to measure tens of RNA species simultaneously, called seqFISH. Later, MERFISH was proposed by Chen et al. [190], which enabled the measurement of hundreds to thousands of transcripts in single cells simultaneously while

Table 4 Approaches for data integration, highlighting their promises and challenges

	Integration	Example MT combination	Example AMs	Promises	Challenges
`1S`	None	scDNA-seq	Clustering/unsupervised	Discover new subclones, cell types, or cell states	Technical noise ■; data sparsity ■
`+S`	Within 1 MT, within 1 exp, across > 1 smps	scRNA-seq	Differential analyses, time series, spatial sampling	Identify effects across sample groups, time, and space	Batch effects ■; validate cell type assignments ■
`+X+S`	Within 1 MT, across > 1 exp, across > 1 smps	merFISH	Map cells to stable reference (cell atlas)	Accelerate analyses, increase sample size, generalize observations	Standards across experimental centers
`+M1C`	Across > 1 MTs, within 1 exp, within 1 cell	scM&T-seq (scRNA-seq + methylome)	MOFA, DIABLO, MINT	Holistic view of cell state; quantify dependency of MTs	Scaling cell throughput; MT combinations limited; dependency of MTs ■
`+M+C`	Across > 1 MTs, within 1 exp, across > 1 cells, within 1 cell pop	scDNA-seq + scRNA-seq	Cardelino, Clonealign, MATCHER	Use existing datasets (faster than `+M1C`); flexible experimental design	Validate cell/data matching; test assumptions for integrating data
`+all`	Across > 1 MTs, across > 1 exps, across > 1 smps, within cells	Hypothetical (any combination)	Hypothetical (map cells to multi-omic HCA, single-cell TCGA)	Holistic view of biological systems	All from approaches `+X+S`, `+M1C`, and `+M+C`

The labeling corresponds to Fig. 6. For each approach, one (combination of) measurement type(s) that is available is given, but more exist and several are discussed in the text. As example analysis methods, actual tool names are given where few tools exist to date; otherwise, broader categories or imaginable methodologies are described

Abbreviations: "■" same challenge also applies to all approaches below, *AM* analysis method, *exp(s)* experiment(s), *HCA* human cell atlas, *MT* measurement type, *smps* samples, *TCGA* The Cancer Genome Atlas

retaining spatial coordinates [191]. Subsequently, Shah et al. [192] have scaled seqFISH to hundreds of RNA species as well. This year, Eng et al. [193] presented SeqFISH+, which scales the FISH barcoding strategy to 10,000 RNA species by splitting each of 4 barcode locations to be scanned into 20 separate readings to avoid optical signal crowding. The latter can also be an issue when fewer RNA species are measured, in particular at densely populated regions such as the nucleus [190]. To solve such issues at the expense of measuring fewer RNA species, Codeluppi et al. [194] have proposed osmFISH, which uses a single fluorescent probe per RNA species and leverages FISH iterations to measure different species instead of building up a barcode. This leads to a number of recognizable RNA species that is linear in the number of FISH iterations. In addition to the methods that provide in situ measurements of RNA, mass cytometry [195, 196] and multiplexed immunofluorescence [197–199] can be used to quantify the abundance of proteins while preserving subcellular resolution. Finally, the recently described Digital Spatial Profiling [200, DSP; 201] 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 [147, 177, 202], but none of these currently use spatial information directly. In contrast, spatial correlation methods have been used to detect the aggregation of proteins [203]. Shah et al. [204] use seqFISH to measure transcript abundance of a set of marker genes while retaining the spatial coordinates of the cells. Cells are clustered by gene expression profiles and then assigned to regions in the brain based on their coordinates in the sample. Recently, Esgärd et al. [205] presented a method to detect spatial differential expression patterns per gene based on marked point processes [206], and Svensson et al. [207] 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.

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. 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 [208] and minor subclones [209] to larger subclonal structures or the embedding in surrounding normal tissue and the tumor microenvironment [210]. 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. The expected tissue heterogeneity therefore also impacts the desired properties of the assignment method itself. For example, in order to also recognize groups or types of interest that are expected to occur at multiple locations, applicable methods should not strictly rely on co-localization of transcriptional profiles.

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 [102, 211] and propagate this information to the spatial analysis. Since many spatial measurement approaches entail an optical, microscopy-based component, it would be beneficial to extract additional information from these measurements. For example, cell shape and size, as well as the subcellular spatial distribution of transcripts or proteins, could be used to additionally guide the clustering or classification process. Finally, in light of issues with sparsity in single-cell measurements (see "Challenge I: Handling sparsity in single-cell RNA sequencing"), it appears desirable to integrate spatial information into the quantification itself, and, for example, use neighboring cells within the same tissue for imputation or the inference of a posterior distribution of transcript expression.

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 variations and complex structural variants. In cancer, the entire repertoire of these genetic events can occur during disease progression (Fig. 4). The resulting tumor cell populations are highly heterogeneous. As tumor heterogeneity can predict patient survival and response to therapy [4, 212], including immunotherapy, quantifying this heterogeneity and understanding its dynamics are crucial for improving diagnosis and therapeutic choices (Fig. 4).

Classic bulk sequencing data of tumor samples taken during surgery are always a mixture of tumor and normal cells (including 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 bearing

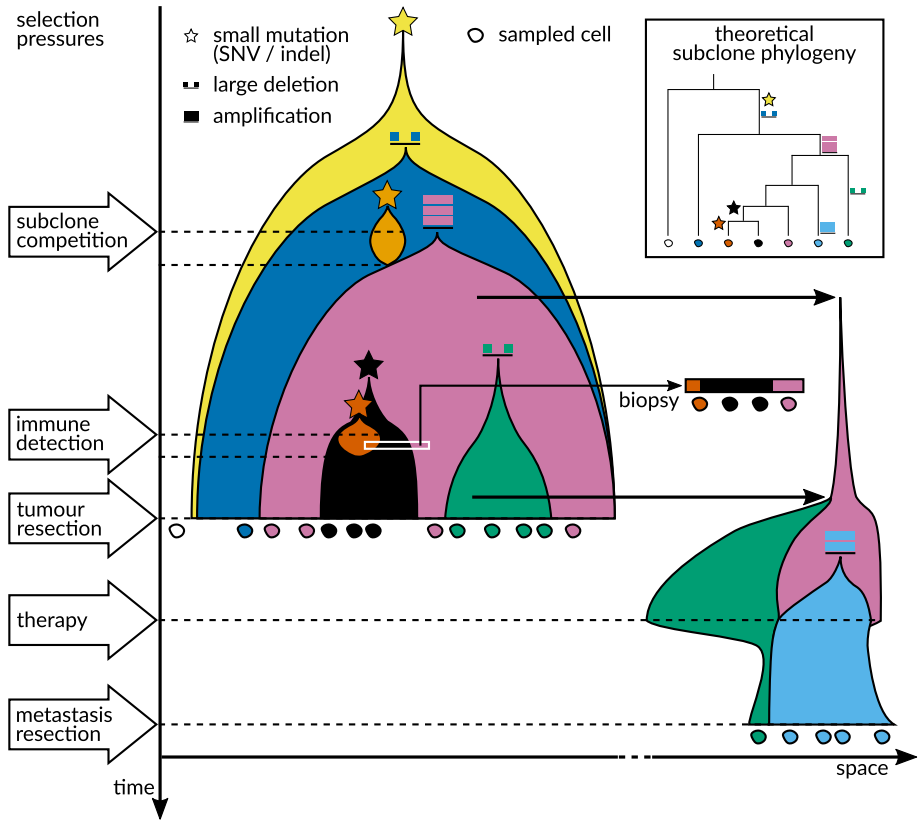

Fig. 4 A tumor evolves somatically—from initiation to detection, to resection, and to possible metastasis. New genomic mutations can confer a selective advantage to the resulting new subclone that allows it to outperform other tumor subclones (subclone competition). At the same time, the acting selection 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

resistance mutation combinations prior to a treatment. Here, the sequencing of single cells holds the exciting promise of directly identifying and characterizing those subclone profiles (Fig. 4).

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 and genomic rearrangements, together with SNVs and smaller insertion and deletion variants. However, scDNA-seq requires WGA of the DNA extracted from single cells and this amplification introduces errors and biases that present a serious challenge to variant calling [213–216]. It is broadly accepted that different WGA technologies should be used to detect different types of variation. PCR-based approaches [217–220] are best suited for CNV calling, as they achieve a more uniform coverage. But they require thermostable polymerases that withstand the temperature maxima during PCR cycling, and all such polymerases have relatively high error rates. In contrast, MDA-based techniques are the method of choice for SNV calling, as they achieve much lower error rates with the high-fidelity Φ29 DNA polymerase [31, 221–225] (in an isothermal reaction, as it would not be stable at common PCR temperature maxima). But MDA suffers from stronger allelic bias in the amplification, possibly because it is more sensitive to DNA input quality [226] and biased priming [227]. The goal must thus be to (i) improve the coverage uniformity of MDA-based methods, (ii) reduce the error rate of the PCR-based methods, or (iii) create new methods that exhibit both a low error rate and a more uniform amplification of alleles. Recent years witnessed intensive research in these directions (see Table 1), promising scalable methodology for scDNA-seq comparable to that already available for scRNA-seq, while at the same time reducing previously limiting errors and biases. While this is not a SCDS challenge, it remains central 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.

Challenge VI: Dealing with errors and missing data in the identification of variation from single-cell DNA sequencing data

The aim of scDNA-seq 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, Fig. 5). Examples are identifying heterogeneity and tracking evolution in cancer, as the likely most predominant use case (also see below in "Challenges in single-cell phylogenomics"), but also 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 (Fig. 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 above). 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 distorted read counts; (ii) allele dropout, i.e., loci harboring heterozygous mutations where only one of the alleles was amplified and sequenced; and (iii) site dropout, 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.

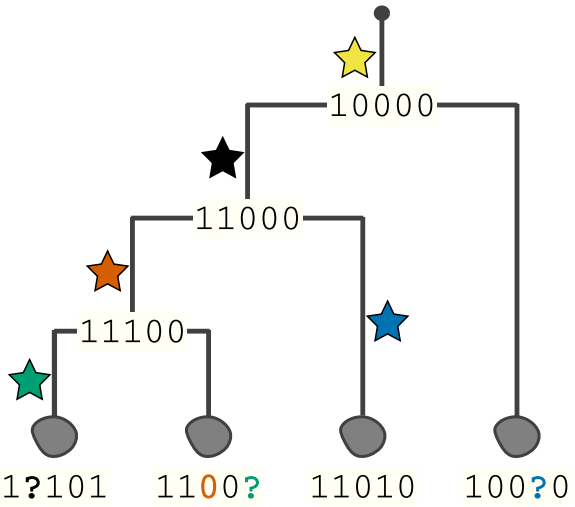

Fig. 5 Mutations (colored stars) accumulate in cells during somatic cell divisions and can be used to reconstruct the developmental lineages of individual cells within an organism (leaf nodes of the tree with mutational presence/absence profiles attached). However, insufficient or unbalanced WGA can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

Status

Current single cell-specific SNV callers include Monovar [228], SCcaller [229], and SCAN-SNV [230]. SCcaller detects somatic variants independently for each cell, but accounts for local allelic amplification biases by integrating across neighboring germline single-nucleotide polymorphisms. It exploits the fact that allele dropout affects contiguous regions of the genome large enough to harbor several, and not only one, heterozygous mutation loci. SCAN-SNV works along similar lines, fitting a region-specific allelic balance model to germline heterozygous variants called in a reference bulk sample. 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 "Challenges in

single-cell phylogenomics"). This tree, however, is in general unknown and can in turn only be reconstructed from single-cell mutation profiles. A possible solution is to infer both mutation calls and a cell lineage tree at the same time, an approach taken by a number of existing tools: single-cell Genotyper [231], SciCloneFit [232], and SciΦ [233]. Finally, SSrGE identifies SNVs correlated with gene expression from scRNA-seq data [234].

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, for example, in Aneufinder [235]. Another tool, Ginkgo, provides interactive CNV detection using circular binary segmentation, but is only available as a web-based tool [236]. ScRNA-seq data, which does not suffer from the errors and biases of WGA, can also be used to call CNVs or loss of heterozygosity events: an approach called HoneyBADGER [237] utilizes a probabilistic hidden Markov model, whereas the R package inferCNV simply averages the expression over adjacent genes [238].

Open problems

SNV callers for scDNA-seq data have already incorporated amplification error rates and allele dropout in their models. Beyond these rates, the challenge remains to further extend this by directly modeling the amplification process using statistics that would inherently account for both errors and biases, and more accurately quantify the resulting uncertainties (see "Quantifying uncertainty of measurements and analysis results"). This could be achieved by expanding models that accurately quantify uncertainties in related settings [239] 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, for example, on a formalization in a recent preprint [240]. 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 [241] 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 [242, 243]. In addition, the integration with deep bulk sequencing data, as well as with scRNA-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 [241].

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 real sample-based benchmarking datasets with a reasonably reliable ground truth (see "Challenge XI: Validating and benchmarking analysis tools for single-cell measurements").

Challenges in single-cell phylogenomics

Single-cell variation profiles from scDNA-seq, as described above ("Challenge VI: Dealing with errors and missing data in the identification of variation from single-cell DNA sequencing data"), can be used in computational models of somatic evolution, including cancer evolution as an important special case (Fig. 4). For cancer, there is an ongoing, lively discussion about the very nature of evolutionary processes at play, with competing theories such as linear, branching, neutral, and punctuated evolution [244].

Models of cancer evolution may range from a simple binary representation of the presence versus the absence of a particular mutational event (Fig. 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 [245]: phylogenetics and population genetics.

Phylogenetics comes with a rich repertoire of computational methods for likelihood-based inference of phylogenetic trees [246]. 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 (Fig. 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 (Fig. 4). To date, population

genetic theory has been used to model the initiation, progression, and spread of tumors from bulk sequencing data [247–249]. The general mathematical framework behind these models are branching processes [250], for example, in models of the accumulation of driver and passenger mutations [251, 252]. Here, the driver mutations carry a fitness advantage, as might epistatic interactions among them [253]. In contrast, 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, for example, 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 will, for example, be essential in the design of adaptive cancer treatment strategies that aim at managing subclonal tumor composition [254, 255].

Challenge VII: 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 and (ii) the increasing numbers of sites that can be queried per genome (see "Scaling to higher dimensionalities: more cells, more features, and broader coverage").

Open problems

(i) While adding data from more single cells will help improve the resolution of tumor phylogenies [256, 257], 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. Phylogenetic inference is NP-hard [258] 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 [259, 260] and other optimization efforts [261] 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. 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 [262].

Challenge VIII: Integrating multiple types of variation into phylogenetic models

Naturally, downstream analyses—like characterizing intratumoral heterogeneity and inferring its evolutionary history—suffer from the unreliable variant detection in single cells. However, the better the quality of the variant calls becomes, the more important it becomes to model all types of available signal in mathematical models of tumor evolution: from SNVs, over smaller insertions and deletions, to large structural variation and CNVs (Fig. 4). In turn, this should increase the resolution and reliability of the resulting trees.

Status

For phylogenetic tree inference from SNVs of single cells, a considerable number of tools exist. The early tools OncoNEM [263] and SCITE [264] 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 [265] 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[2]. 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) [266, 267]. Single-cell genotyper [231], SciCloneFit [232], or SciΦ [233] 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 [268], a standard phylogenetic inference tool RAxML-NG [269] 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 outper-

[2] The infinite sites assumption posits a genome with an infinite number of sites, thus rendering a repeated mutational hit of the same genomic site along a phylogeny impossible.

formed 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 (preliminary analysis by Kozlov).

While there are no tools yet available to identify insertions and deletions from scDNA-seq (see "Challenge VI: Dealing with errors and missing data in the identification of variation from single-cell DNA sequencing data"), it is only a matter of time until such callers will become available. As they can already be identified from bulk sequencing data, some precious efforts to incorporate indels in addition to substitutions into classical phylogenetic models exist: A decade ago, a simple probabilistic model of indel evolution was proposed [270]. But although some progress has been made since then, such models are less tractable than the respective substitution models [271].

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 [272]. CNVs in single cells were extensively studied in the context of tumor evolution and clonal dynamics [273, 274]. Reconstructing a phylogeny with CNVs is not straightforward. The challenges not only are related to experimental limits, such as the complexity of bulk sequencing data [275] and amplification biases [276], but also involve computational constraints. First of all, the causal mechanisms, such as breakage-fusion-bridge cycles [277] and chromosome missegregation [278], can lead to overlapping copy number events [279]. Secondly, inferring a phylogeny with CNV data requires quantifying biologically motivated transition probabilities for changes in copy numbers. Towards that goal, approaches to calculate the distance between whole copy number profiles [280] are a first step. But for them, a number of challenges remain, with several of the underlying problems known to be NP-hard [280].

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 [281]. This undermines the validity of the so-called infinite sites assumption, commonly made by phylogenetic models.

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 Table 1 and "Challenge VI: Dealing with errors and missing data in the identification of variation from single-cell DNA sequencing data"). For indels, variant callers for scDNA-seq data are anticipated but remain to be developed (see "Challenge VI: Dealing with errors and missing data in the identification of variation from single-cell DNA sequencing data"). 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 problem 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 and called (see "Challenge VI: Dealing with errors and missing data in the identification of variation from single-cell DNA sequencing data") 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 IX: Inferring population genetic parameters of tumor heterogeneity by model integration").

Challenge IX: 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 [209, 212]. 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 [282]. 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 ([283, 284] and Fig. 4). However, even the basic questions about the resulting dynamics remain unanswered [285]. For example, it is unclear 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 (seeding multiple locations with a

genotype closer to the late-stage primary tumor). Moreover, it is unknown 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 [273].

Although many mathematical models of tumor evolution have been proposed [245, 247, 251, 252, 286, 287], 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.

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 `+M1C` in "Challenge X: Integration of single-cell data across samples, experiments, and types of measurement", Fig. 6, and Table 4). While in vivo imaging techniques might also become interesting for obtaining time series data in the future [288], the automated analysis of whole slide immunohistochemistry images [289, 290] 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 [291], hundreds of single cells have recently been isolated from tissue sections and analyzed for copy number variation [292]. For cell and tissue characterization in immunohistochemical images, machine learning models are trained to segment the images and recognize structures within tissues and cells [293–295]: They can, for example, determine the densities and quantities of mitotic nuclei, vascular invasion, and 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 of tumor cells with their environment in space [296, 297], that are key to mathematical models of cancer evolution. Development of reliable classifiers for immunohistochemical images, however, is challenging due to scarcity of training data. Solutions such as active learning can speed up the training process and reduce the workload of annotating pathologists [298].

Classically, mathematical models of tumor population genetics have 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 [299] and intratumor heterogeneity [300]. In particular, spatial statistics have been proposed for the quantitative statistical analysis of cancer digital pathology imaging [297], but the idea is applicable to other spatially resolved readouts. Further, a number of methods were proposed to model cell-cell interactions [301, 302] or to predict single-cell expression from microenvironmental features [199, 303].

Regarding temporal resolution, it is already common to sequence tumor material from different time points: biopsies used for diagnosis, resected tumors, lymph nodes and metastases upon surgery, and tumors after relapse. These time points already lend themselves to temporal analyses of clonal dynamics using bulk DNA sequencing data [304], but scDNA-seq is required for a higher resolution of subclonal genotypes. In addition, time resolved measurements and resulting proliferation and death rates promise a higher accuracy in detecting epistatic interactions in cancer genomes than available from previous analyses of bulk sequenced tumor genomes [305–308].

Eventually, population genetic methods and models should be integrated with approaches from phylogenetics, to also leverage the kinship relationships between cells. One prominent example of this recent trend—albeit on bulk data—is the use of the multi-species coalescent model for analyzing MSAs that contain several individuals for several populations [309, 310]. This naturally translates into analyzing tumor subclones as populations of single cells, capturing some of the population structure seen in cancers. Another recent example is a computational model for inference of fitness landscapes of cancer clone populations using scDNA-seq data, SCIFIL [311]. It estimates the maximum likelihood fitness of clone variants by fitting a replicator equation model onto a character-based tumor phylogeny.

For a comprehensive integration, key parameters will need to be quantified with higher resolution. For the detection of positive selection—for example, important in the discussion whether the evolution of tumors is driven by selection or neutral—a number of phylogenetic and population genetic approaches have been proposed in a bulk context. Phylogenetic trees may be used for detecting branches on which positive [312] or diversifying episodic selection [313] is acting.

In this setting, we will have to account for heterotachy (e.g., [314]), 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., [315]) to an empirical mixture model as used for protein evolution [316] could be considered.

Open problems

With an increased resolution of scDNA-seq (see "Challenges in single-cell genomics", Table 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. Studies utilizing multiple region samples from the same tumor and from distant metastases already paved the way in investigating these questions (e.g., [285]). Still, only single-cell spatial resolution will allow identification of specific individual genotypes in specific locations and drawing precise conclusions.

In addition, it will become possible to determine subclone-specific model parameters and their variability in more detail. For example, rates of proliferation, mutation, and death could be obtained 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 the detection of positive and negative selection in cancer, and improve the prediction of subclone resistance (and thus expected treatment success) from subclone fitness estimates. The fitness of individual subclones could be calculated from comparing expanded subclones in drug screens under different treatment regimes.

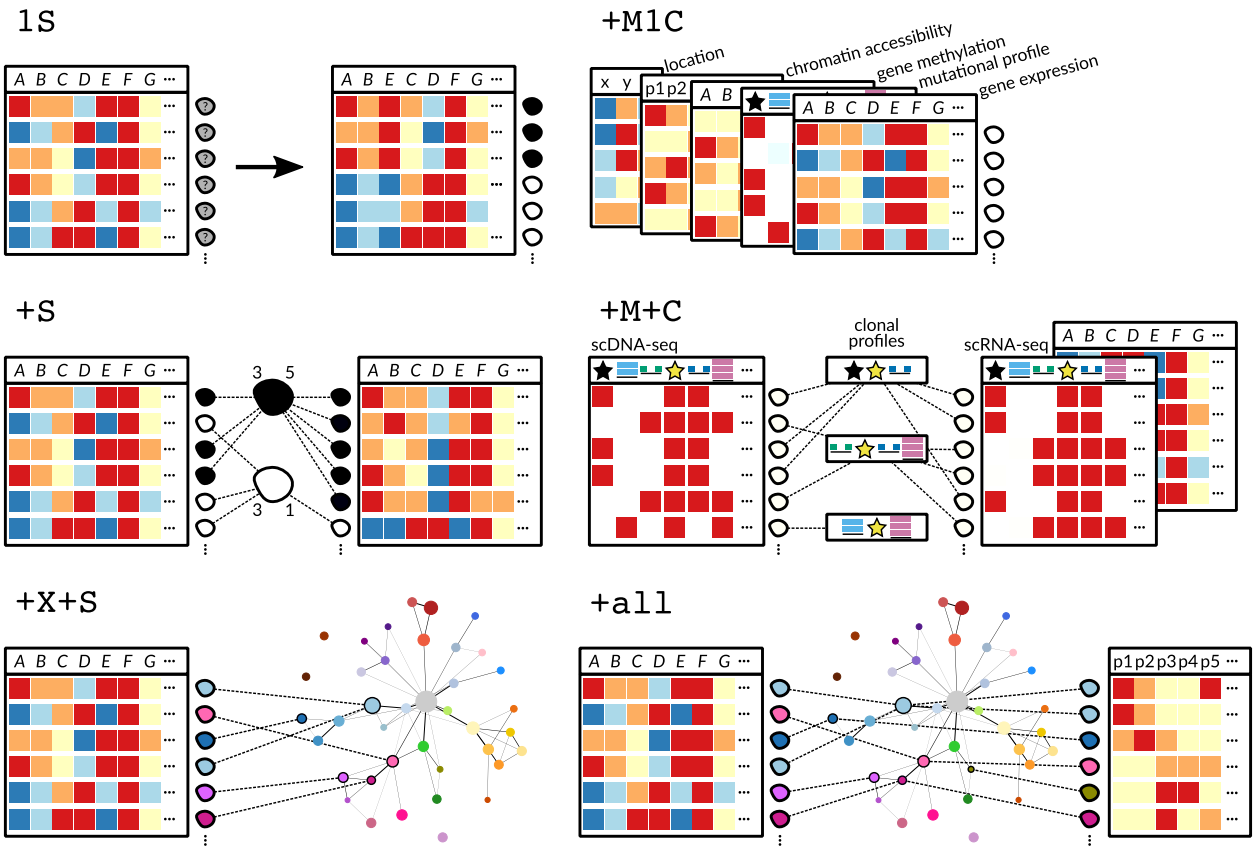

Fig. 6 Approaches for integrating single-cell measurement datasets across measurement types, samples, and experiments, as also described in Table 4. `1S`: clustering of cells from one sample from one experiment requires no data integration. `+S`: integration of one measurement type across samples requires the linking of cell populations/clusters. `+X+S`: integration of one measurement type across experiments conducted in separate laboratories requires stable reference systems like cell atlases (compare Fig. 1). `+M1C`: integration of multiple measurement types obtained from the same cell highlights the problem of data sparsity of all available measurement types and the dependency of measurement types that needs to be accounted for. `+M+C`: integration of different measurement types from different cells of the same cell population requires special care in matching cells through meaningful profiles. `+all`: 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, capturing different cell states, cell populations, and organisms. Effectively, this combines the challenges and promises of the approaches `+X+S`, `+M1C`, and `+M+C`

For some of the rates, for example, subclone-specific rates of mutation, the integration of models from population genetics and phylogenetics holds promise and poses a genuine SCDS challenge. But for all of these rates, having better estimates implies follow-up challenges.

One of these resulting challenges will be to detect positive or diversifying selection with greater resolution, building on approaches from the bulk context. Here, 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 [317]. Other measures have been proposed as better suited for detecting selection within populations based on time series data [318–320] and could potentially be transferred to tumor cell populations.

A particular problem with the detection of positive or diversifying selection is to which extent the above tests will be sensitive to errors in cancer data—the tests are already known to produce high false positive rates in the classic phylogenetic setting when the error rate in the input data is too high [321]. Computationally intense solutions for decreasing the high false positive rate have been proposed [322], but they might not scale to single-cell cancer datasets.

Another resulting problem will be to adapt models for the detection of epistatic interactions to single-cell data. As some of these epistatic interactions can be hard to spot in bulk sequencing data (they may simply disappear because of a low frequency), time-resolved scDNA-seq might be the only way to spot them. If integrated across individuals and cells (see "Challenge X: Integration of single-cell data across samples, experiments, and types of measurement"), it will be possible to identify pairs or even larger combinations of mutations that often occur simultaneously in the same genome, and combinations 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. At the same time, cell death 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, resulting in an even larger data integration challenge (see "Challenge X: Integration of single-cell data across samples, experiments, and types of measurement").

A final step will then be to integrate all these parameters 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.

Overarching challenges

Challenge X: 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 can be different time points, tissues, or organisms. For their integration, we need flexible but rigorous statistical and computational frameworks. Figure 6 and Table 4 provide an overview of the promises and challenges of creating such frameworks that we outline here in terms of six approaches of data integration[3]. All of these approaches 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 (see "Varying levels of resolution"), (ii) the uncertainty of any measurements and how to quantify them for and during the analyses (see "Quantifying uncertainty of measurements and analysis results"), and (iii) the scaling of single-cell methodology to more cells and more features measured at once (see "Scaling to higher dimensionalities: more cells, more features, and broader coverage"). All of these further compound the most important challenge in the integration of single-cell data: to link data from different sources in a way that is biologically meaningful and supports the intended analysis. The maps that describe how data from different sources is linked will increase in complexity on increasing amounts of samples, time points, and types of measurements.

In the simplest setup, we obtain one measurement type from multiple cells of a single sample, to identify subpopulations of cells (e.g., subclones or cell types). As any analysis of sc-seq data, it needs to take into account the data's sparsity (see "Challenge I: Handling sparsity in single-cell RNA sequencing" and "Challenge VI: Dealing with errors and missing data in the identification of variation from single-cell DNA sequencing data"; approach `1S` in Fig. 6 and Table 4).

When aiming at identifying patterns of differential expression or characterizing variability across organisms, individuals, or locations, the same measurement type (for example, only scRNA-seq) is taken from multiple samples from different time points, different locations (e.g., different tissues or sites in a tumor), or different organisms

[3]Graph representation in Fig. 6 approaches `+X+S` and `+all` taken from [14], Fig. 3, provided under Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/)

(approach +S). Any such combination of samples requires accounting for batch effects among those samples and calls for a validation cell type assignments across samples.

Such batch effects are further aggravated when integrating across multiple experiments, possibly run in different experimental centers with similar but distinct setups (approach +X+S). But standardizing experimental procedures and statistically accounting for batch effects will be well worth the effort wherever this enables a significant increase in sample size, so as to generalize (and statistically corroborate) observations. Nevertheless, even if standards have been successfully established and known batches accounted for, additional validation of, for example, assignments of cells to types and states may be required. Eventually, an increase in generality will support the construction of reference systems, such as a cell atlas, the existence of which can support decisive speed-ups when classifying cells or cell states in subsequent experiments (see "Challenge III: Mapping single cells to a reference atlas").

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. Here, an optimal setup is to collect several types of measurements from each cell at once; for example, both scDNA-seq and scRNA-seq captured from the same cell, possibly further augmented by measurements of chromatin accessibility, gene methylation, proteins, or metabolites (approach +M1C). The most prominent challenge for this setup is to model inherent dependencies between measurement types wherever phenomena are concurrent (e.g., measuring CNV through scDNA-seq at the same time as obtaining scRNA-seq, with CNV impacting transcription levels).

However, co-measuring different types of quantities in the same cell can be experimentally challenging or even just impossible at this point in time. An exit strategy to this problem is to analyze a population of cells that is homogeneous in terms of some cell type or state, taking different measurement types in different single cells (approach +M+C). After collecting different measurement types in different single cells, one needs to combine the data in a way that is biologically meaningful. An example is to group cells based on commonalities in their genotype profile (Fig. 6), having become evident only after the application of a scDNA-seq experiment. This will require careful validation of the assumptions made when matching cells via such a grouping, possibly including functional validation of group differences.

Finally, the most comprehensive goal will be a holistic view of the complexity of (intra-)cellular circuits, and charting their variability across time, tissues, populations, and organisms (approach +all). Mapping cellular circuits in this comprehensive manner requires integrating complementary and possibly interdependent measurements in single cells and across multiple single cells from diverse samples.

Status

For *unsupervised clustering* (approach 1S in Fig. 6 and Table 4), method development is a well-established field. Remaining challenges have already been identified systematically (see [125–127]).

For *integrating datasets across samples in one experiment* (approach +S), a few approaches are available. See for example MNN [118], and the methodologies included in the Seurat package [177, 323, 324]. 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 "Challenge V: Finding patterns in spatially resolved measurements" and "Challenge IX: Inferring population genetic parameters of tumor heterogeneity by model integration".

For *integrating datasets across experiments* (approach +X+S), mapping cells to reference datasets such as the Human Cell Atlas [5] is currently emerging as the most promising strategy. We refer the reader to more particular and detailed discussions in "Challenge III: Mapping single cells to a reference atlas". While applicable reference systems are not (fully) available, assembling cell type clusters from different experiments is a reasonable strategy, as implemented by several recently published tools [202, 325–332].

Integrating across multiple measurement types from the same cell (approach +M1C) has become necessary (and possible) with the advent of experimental protocols that enable the collection of such data [333]. Such protocols combine scDNA-seq and scRNA-seq [333–335]; methylation data and scRNA-seq [336]; all of scRNA-seq, scDNA-seq, methylation, and chromatin accessibility data [41]; or targeted queries on a cell's genotype, expression (scRNA-seq), and methylation status (sc-GEM [337]). For these single cell-specific approaches, bulk approaches that address the integration of data from different types of experiments have the potential to be adapted to single cell-specific noise characteristics (MOFA [92], DIABLO [338], mixOmics [339], and MINT [340]).

For *integrating across multiple measurement types from separate cells* (approach +M+C), all of which stem from a population of cells that is homogeneous with respect to some selection criterion, technologies such as 10X genomics [171] for scRNA-seq and direct library preparation (DLP [341]) for scDNA-seq establish a scalable experimental basis. The greater analytical challenge is to identify subpopulations that had so far remained invisible, and whose identification is crucial so as to not combine different types of data in mistaken ways. An example for

this is the identification of distinct cancer clones from cells sampled from seemingly homogeneous tumor tissue. Here, only performing scDNA-seq experiments can definitively reveal the clonal structure of a tumor. If one wishes to correctly link mutation with transcription profiles, ignoring the clonal structure of a tumor could be misleading. Several analytical methods that address this problem have recently emerged: (i) clonealign [91] assumes a copy number dosage effect on transcription to assign gene expression states to clones, (ii) cardelino [342] aligns clone-specific SNVs in scRNA-seq to those inferred from bulk exome data in order to infer clone-specific expression patterns, and (iii) MATCHER [18] uses manifold alignment to combine scM&T-seq [336] with scGEM [337], leveraging the common set of loci. All of these methods are based on biologically meaningful assumptions on how to summarize data measurements across different measurement types and samples, despite their different physical origin.

Open problems

Experimental technologies that enable taking multiple measurement types in the same cell (approach `+M1C` in Fig. 6 and Table 4) are on the rise and will allow to assay more cells at higher fidelity and reduced cost. While this type of data naturally links measurement types within single cells, the SCDS challenge is to account for dependencies among those measurement types for any obtainable combinations of them. As a prominent example, consider how gene expression increases with higher genomic copy number, a phenomenon known as measurement linkage [343], which has not been addressed for different measurement types taken in the same cell. Statistical models for leveraging those measurement type combinations thus pose formidable SCDS challenges.

While progress on the approach `+M1C` may gradually render approach `+M+C` obsolete, `+M+C` will remain the easier—or the only feasible—approach for many measurement type combinations for a while. At the same time, any advances in characterizing dependencies between different measurement types acquired from separate cells (`+M+C`) provide further ground work for linking them when acquired from the same cell (`+M1C`). Take the example from above, where copy number profiles will impact gene expression measurements. Here, an approach that accounts for this in `+M+C` exists (clonealign [91]) and could be extended to `+M1C` datasets. For approach `+M+C`, the possibility to integrate data from single cells with data from bulk sequencing of the same cell population also holds promise, for example, by using bulk genotypes for imputation of sites with no sequencing coverage in single cells. Finally, knowing how to link (different) measurement types acquired from different cells is essential for building reference systems across experiments, such as cell atlases (see also approaches `+X+S` and `+all`, and "Challenge III: Mapping single cells to a reference atlas"). Thus, exploring further combinations of measurement types and their measurement linkage in `+M+C` datasets remains as a central SCDS challenge.

No matter which combinations of measurement types become available—the amounts of material underlying most measurements will remain tiny, limited by the amounts within a single cell as well as by a limited number of cells available from a particular cell population. This means that one overarching theme will persist: analyses like training models or mapping quantities on one another will suffer from missing entire views—samples, time points, or measurement types. Thus, integrating data across experiments and different measurement types will further compound the challenge of missing data that we already discussed for non-integrative approaches (see "Challenge I: Handling sparsity in single-cell RNA sequencing" and "Challenge VI: Dealing with errors and missing data in the identification of variation from single-cell DNA sequencing data").

Challenge XI: 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 increasingly 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 addition, benchmarking should be conducted in a systematic way, following established recommendations [344, 345].

Evaluation of tool performance requires benchmarking datasets with known ground truth. Such data should include cell populations with known genomic compositions and population structures, in other words where frequencies of clones and alleles are known. Currently, such datasets are scarce—with some notable exceptions [346, 347]—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 requires design and implementation of models that 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.

Status

Recent studies [104, 111, 148, 157, 348] 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, for all the possible use cases. Many single-cell data analysis packages include their own ad hoc data simulators [111, 211, 241, 264, 349–353]. 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 guaranteed to be evaluated, and therefore, the accuracy of their reflection of real biological and technological processes can remain unclear. There are few exceptions known to us, including the tools Splatter [354], powsimR [355], and SymSim [356], 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 [357–359].

Open problems

Current 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 "Challenge IX: Inferring population genetic parameters of tumor heterogeneity by model integration"). To this end, realistic and comprehensive (w.r.t. the evolutionary phenomena) simulation tools are required.

Another interesting computational problem is the 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 [360], 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 possess multiple biological characteristics to be inferred and analyzed. Development of a single measure that 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 that capture several such characteristics (e.g., based on utilization of earth mover's distance [361]) 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 challenge. This approach has been very successful, for example, in protein structure prediction[4] and metagenomic analyses[5]. A first step in this direction was the recent single-cell transcriptomics DREAM challenge[6].

Finally, drawing on all the exemplary benchmarking studies mentioned above, it would be immensely beneficial to bring all the required efforts together in a community-supported benchmarking platform: (i) simulating datasets and validating that they capture important characteristics of real data, (ii) curating ground truths for

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

real datasets, and (iii) agreeing on comprehensive evaluation metrics. Ideally, such a benchmarking framework would remain dynamic beyond an initial publication—to allow ongoing comparison of methods as new approaches are proposed and to easily extend it to entirely new fields of method development.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13059-020-1926-6.

Additional file 1: Review history.

Abbreviations

CNV: Copy number variation; FISH: Fluorescent in situ hybridization; ICA: Independent component analysis; MALBAC: Multiple annealing and looping-based amplification cycles; MDA: Multiple displacement amplification; MSA: Multiple sequence alignment; NMF: Non-negative matrix factorization; PCA: Principal component analysis; PCR: Polymerase chain reaction; sc-seq: Single-cell sequencing; scDNA-seq: Single-cell DNA sequencing; SCDS: Single-cell data science; scRNA-seq: Single-cell RNA sequencing; SNV: Single nucleotide variation; WGA: Whole genome amplification

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, and 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.

Review history

The review history is available as Additional file 1.

Authors' contributions

DL, JK, ES, KRC, DJM, SCH, MDR, CAV, NB, AM, LP, PS, ASt, CSOA, AMK, THK, IIM, ACM, and ASch authored or reviewed substantial parts of the paper. DL, JK, ES, KRC, DJM, SCH, MDR, CAV, NB, LP, PS, CSOA, TJL, FM, and ASch prepared the figures and/or tables. JK, ACM, BJR, SPS, and ASch organized and coordinated the workshop. All authors actively participated in the discussions underlying this review, which took place in working groups at the Lorentz workshop "Single Cell Data Science: Making Sense of Data from Billions of Single Cells." All authors approved the final manuscript.

Authors' information

Twitter handles: @DLaehnemann (David Lähnemann), @johanneskoester (Johannes Köster), @ewa_szczurek (Ewa Szczurek), @davisjmcc (Davis J. McCarthy), @stephaniehicks (Stephanie C. Hicks), @markrobinsonca (Mark D. Robinson), @CataVallejosM (Catalina A. Vallejos), @kieranrcampbell (Kieran R. Campbell), @cbg_ethz (Niko Beerenwinkel), @ahmedElkoussy (Ahmed Mahfouz), @lucapinello (Luca Pinello), @AlexisCompBio (Alexandros Stamatakis), @yocamilleyo (Camille Stephan-Otto Attolini), @sajraparicio (Samuel Aparicio), @BEDutilh (Bas E. Dutilh), @Vityay (Victor Guryev), (Rens Holmer) @holmrenser, @Lumc_InduK (Indu Khatri), @JanKorbel5 (Jan O. Korbel), @tobiasmarschal (Tobias Marschall), @jeroen_deridder (Jeroen de Ridder), @AE_Saliba (Antoine-Emmanuel Saliba), @OliverStegle (Oliver Stegle), @fabian_theis (Fabian J. Theis), @Alexz81744967 (Alex Zelikovsky), @alicecarolyn (Alice C. McHardy), @benjraphael (Benjamin J. Raphael), @SohrabShah (Sohrab P. Shah).

Funding

AC was supported by an IAS Fellowship for external researchers at the University of Amsterdam. ACM was supported by the Helmholtz Incubator (Sparse2Big ZT-I-0007). AMK and ASt were supported by the Klaus Tschira Foundation. AZ was supported by the National Science Foundation (NSF: DBI-1564899, CCF-1619110) and the National Institutes of Health (NIH: 1R01EB025022-01). BdB was supported by the Oncode Institute (220-H72009 – KWF/2016-1/10158). BED was supported by the Netherlands Organisation for Scientific Research (NWO: Vidi grant 864.14.004). BJR was supported by the NSF (CCF-1053753), the NIH (R01HG007069), and the Chan Zuckerberg Initiative (CZI) Donor-Advised Fund (DAF) (2018-182608), an advised fund of the Silicon Valley Community Foundation. CAV was supported by the University of Edinburgh (Chancellor's Fellowship) and by The Alan Turing Institute (EPSRC grant EP/N510129/1). DJM was supported by the National Health and Medical Research Council of Australia (GNT1112681 and GNT1162829). DL was supported by the Katharina Hardt Stiftung and the Deutsche Krebshilfe (national Network Genomic Medicine Lung Cancer, nNGM). FJT was supported by the German Research Foundation (DFG: Collaborative Research Centre 1243, Subproject A17), the Helmholtz Incubator (Sparse2big ZT-I-0007), the BMBF (01IS18036A, 01IS18053A, and 01ZX1711A), and the CZI DAF (182835). GC was supported by Marie Sklodowska-Curie grant (agreement no. 642691, Epipredict). IIM was supported by the NSF (award 1564936). JCM was supported by core funding from the European Molecular Biology Laboratory (EMBL) and core support from Cancer Research UK (CRUK: C9545/A29580). JdR was supported by the NWO (Vidi grant 639.072.715). JK was supported by the NWO (Veni grant 016.173.076). JOK was supported by core funding from the EMBL. KJ was supported by SystemsX.ch (RTD Grant 2013/150). KRC was funded by postdoctoral fellowships from the Canadian Institutes of Health Research (CIHR), the Canadian Statistical Sciences Institute (CANSSI), and the UBC Data Science Institute. LP was supported by the NIH – National Human Genome Research Institute (NHGRI) Career Development Award (R00HG008399), the Genomic Innovator Award (R35HG010717), and the CZI DAF (2018-182734). MB was supported by the NWO (Vidi grant 639.072.309 and Vidi grant 864.14.004). MDR was supported by the Swiss National Science Foundation (310030 175841, CRSII5 177208) and the CZI DAF (2018-182828) and acknowledges support from the University Research Priority Program Evolution in Action at the University of Zurich. NB was supported by the ERC (Synergy Grant 609883). OS was supported by core funding from the EMBL. PS was supported by the NIH (1R01EB025022). SA was supported by BC Cancer Foundation, CIHR, Canadian Cancer Society (CCS) Research Institute (CCSRI), Terry Fox Research Institute (TFRI), and CRUK. SCH was supported by the NIH – NHGRI (R00HG009007) and by the CZI DAF (182891, 193161, 356-01). SPS was supported by a Susan G. Komen scholar award, the Nicholls Biondi Endowed Chair in Computational Oncology, Cycle for Survival Benefitting Memorial Sloan Kettering Cancer Center, the TFRI, and the CCS. THK was supported by the Award of a Research Fellowships for the Promotion of Scientific Cooperation at the Helmholtz-Centre for Infection Research.

Competing interests

BJR is a co-founder and consultant at Medley Genomics. FJT reports receiving consulting fees from Roche Diagnostics GmbH and Cellarity Inc., and ownership interest in Cellarity, Inc. and Dermagnostix GmbH. IIM is a co-founder and holds an interest in SmplBio LLC, a company developing cloud-based scRNA-Seq analysis software. No products, services, or technologies of SmplBio have been evaluated or tested in this work. JdR is co-founder of Cyclomics BV. SA is a scientific advisor to Sangamo and Repare Therapeutics. SA and SPS are both founders and shareholders of Contextual Genomics Inc. All other authors declare that they have no competing interests.

Author details

[1]Algorithms for Reproducible Bioinformatics, Genome Informatics, Institute of Human Genetics, University Hospital Essen, University of Duisburg-Essen, 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, Warszawa, 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, 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 Zürich, Zürich, Switzerland. [10]MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Edinburgh, UK. [11]The Alan Turing Institute, British Library, London, UK. [12]Department of Statistics, University of British Columbia, Vancouver, Canada. [13]Department of Molecular Oncology, BC Cancer Agency, Vancouver, Canada. [14]Data Science Institute, University of British Columbia, Vancouver, Canada. [15]Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland. [16]SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland. [17]Leiden Computational Biology Center, Leiden University Medical Center, Leiden, The Netherlands. [18]Delft Bioinformatics Lab, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Delft, 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, Heidelberg, Germany. [24]Institute for Theoretical Informatics, Karlsruhe Institute of Technology, Karlsruhe, Germany. [25]Institute for Research in Biomedicine, The Barcelona Institute of Science and Technology, Barcelona, Spain. [26]Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, Canada. [27]Life Sciences and Health, Centrum Wiskunde & Informatica, Amsterdam, The Netherlands. [28]Theoretical Biology and Bioinformatics, Science for Life, Utrecht University, Utrecht, The Netherlands. [29]Center for Molecular Medicine, University Medical Center Utrecht, Utrecht, The Netherlands. [30]Oncode Institute, Utrecht, The Netherlands. [31]Quantitative biology, Hubrecht Institute, Utrecht, The Netherlands. [32]Institute for Advanced Study, University of Amsterdam, Amsterdam, The Netherlands. [33]Department of Surgery and Cancer, The Imperial Centre for Translational and Experimental Medicine, Imperial College London, 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, Groningen, The Netherlands. [36]Bioinformatics Group, Wageningen University, Wageningen, The Netherlands. [37]Biometris, Wageningen University & Research, Wageningen, The Netherlands. [38]Department of Immunohematology and Blood Transfusion, Leiden University Medical Center, Leiden, The Netherlands. [39]Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands. [40]Genome Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany. [41]PRB lab, Delft University of Technology, Delft, The Netherlands. [42]Division of Image Processing, Department of Radiology, Leiden University Medical Center, Leiden, The Netherlands. [43]Computer Science & Engineering Department, University of Connecticut, Storrs, USA. [44]Cancer Research UK Cambridge Institute, Li Ka Shing Centre, University of Cambridge, Cambridge, UK. [45]Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, UK. [46]European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, UK. [47]Center for Bioinformatics, Saarland University, Saarbrücken, Germany. [48]Max Planck Institute for Informatics, Saarbrücken, Germany. [49]Institute of Pathology, University Hospital Essen, University of Duisburg-Essen, Essen, Germany. [50]Computation molecular design, Zuse Institute Berlin, Berlin, Germany. [51]Mathematics Department, Mount Saint Vincent, New York, USA. [52]Helmholtz Institute for RNA-based Infection Research, Helmholtz-Center for Infection Research, Würzburg, Germany. [53]Division of Computational Genomics and Systems Genetics, German Cancer Research Center–DKFZ, Heidelberg, Germany. [54]Institute of Computational Biology, Helmholtz Zentrum München–German Research Center for Environmental Health, Neuherberg, Germany. [55]Division of Drug Discovery and Safety, Leiden Academic Center for Drug Research–LACDR–Leiden University, Leiden, The Netherlands. [56]Department of Computer Science, Georgia State University, Atlanta, USA. [57]The Laboratory of Bioinformatics, I.M. Sechenov First Moscow State Medical University, Moscow, Russia. [58]Department of Computer Science, Princeton University, Princeton, USA. [59]Computational Oncology, Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, USA.

Received: 2 August 2019 **Accepted:** 2 January 2020
Published online: 07 February 2020

References

1. Nature Methods. Method of the year 2013. Nat Methods. 2014;11(1):1–1. https://doi.org/10.1038/nmeth.2801. Accessed 15 Oct 2019.
2. Anchang B, Hart TDP, Bendall SC, Qiu P, Bjornson Z, Linderman M, Nolan GP, Plevritis SK. Visualization and cellular hierarchy inference of single-cell data using SPADE. Nat Protocol. 2016;11(7):1264–79. https://doi.org/10.1038/nprot.2016.066. Accessed 21 June 2016.
3. Francis JM, Zhang C-Z, Maire CL, Jung J, Manzo VE, Adalsteinsson VA, Homer H, Haidar S, Blumenstiel B, Pedamallu CS, Ligon AH, Love JC, Meyerson M, Ligon KL. EGFR variant heterogeneity in glioblastoma resolved through single-nucleus sequencing. Cancer Discov. 2014;4(8):956–71. https://doi.org/10.1158/2159-8290.CD-13-0879. Accessed 01 Aug 2019.
4. Lawson DA, Kessenbrock K, Davis RT, Pervolarakis N, Werb Z. Tumour heterogeneity and metastasis at single-cell resolution. Nat Cell Biol. 2018;20(12):1349. https://doi.org/10.1038/s41556-018-0236-7. Accessed 01 Aug 2019.
5. Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, Bodenmiller B, Campbell P, Carninci P, Clatworthy M, Clevers H, Deplancke B, Dunham I, Eberwine J, Eils R, Enard W, Farmer A, Fugger L, Göttgens B, Hacohen N, Haniffa M, Hemberg M, Kim S, Klenerman P, Kriegstein A, Lein E, Linnarsson S, Lundeberg J, Majumder P, Marioni JC, Merad M, Mhlanga M, Nawijn M, Netea M, Nolan G, Pe'er D, Phillipakis A, Ponting CP, Quake S, Reik W, Rozenblatt-Rosen O, Sanes J, Satija R, Schumacher TN, Shalek A, Shapiro E, Sharma P, Shin JW, Stegle O, Stratton M, Stubbington MJT, Oudenaarden AV, Wagner A, Watt F, Weissman J, Wold B, Xavier R, Yosef N, et al. The Human Cell Atlas. 2017. https://doi.org/10.1101/121202. Accessed 27 Mar 2019.
6. Zilionis R, Nainys J, Veres A, Savova V, Zemmour D, Klein AM, Mazutis L. Single-cell barcoding and sequencing using droplet microfluidics. Nat Protoc. 2017;12(1):44–73.
7. Vitak SA, Torkenczy KA, Rosenkrantz JL, Fields AJ, Christiansen L, Wong MH, Carbone L, Steemers FJ, Adey A. Sequencing thousands of single-cell genomes with combinatorial indexing. Nat Methods. 2017;14(3):302–8. https://doi.org/10.1038/nmeth.4154. Accessed 28 June 2019.
8. Svensson V, Vento-Tormo R, Teichmann SA. Exponential scaling of single-cell RNA-seq in the past decade. Nat Protocols. 2018;13(4):599–604. https://doi.org/10.1038/nprot.2017.149. Accessed 28 June 2019.
9. Luo T, Fan L, Zhu R, Sun D. Microfluidic single-cell manipulation and analysis: methods and applications. Micromachines (Basel). 2019;10(2):104. https://doi.org/10.3390/mi10020104.
10. Gao D, Jin F, Zhou M, Jiang Y. Recent advances in single cell manipulation and biochemical analysis on microfluidics. Analyst. 2019;144(3):766–81.
11. Cao J, Spielmann M, Qiu X, Huang X, Ibrahim DM, Hill AJ, Zhang F, Mundlos S, Christiansen L, Steemers FJ, Trapnell C, Shendure J. The single-cell transcriptional landscape of mammalian organogenesis. Nature. 2019;566(7745):496. https://doi.org/10.1038/s41586-019-0969-x. Accessed 28 June 2019.
12. Amezquita RA, Carey VJ, Carpp LN, Geistlinger L, Lun ATL, Marini F, Rue-Albrecht K, Risso D, Soneson C, Waldron L, Pagès H, Smith M, Huber W, Morgan M, Gottardo R, Hicks SC. Orchestrating single-cell analysis with bioconductor. bioRxiv. 2019590562. https://doi.org/10.1101/590562. Accessed 28 Oct 2019.
13. Hicks SC, Peng RD. Elements and principles of data analysis. arXiv:1903.07639 [stat]. 2019. http://arxiv.org/abs/1903.07639. Accessed 02 Apr 2019.
14. Wolf FA, Hamey FK, Plass M, Solana J, Dahlin JS, Göttgens B, Rajewsky N, Simon L, Theis FJ. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. Genome Biol. 2019;20(1):59. https://doi.org/10.1186/s13059-019-1663-x. Accessed 01 Apr 2019.
15. Pezzotti N, Höllt T, Lelieveldt B, Eisemann E, Vilanova A. Hierarchical stochastic neighbor embedding. Comput Graphics Forum. 2016;35(3):21–30. https://doi.org/10.1111/cgf.12878. Accessed 28 June 2019.
16. Unen VV, Höllt T, Pezzotti N, Li N, Reinders MJT, Eisemann E, Koning F, Vilanova A, Lelieveldt BPF. Visual analysis of mass cytometry data by hierarchical stochastic neighbour embedding reveals rare cell types. Nat Commun. 2017;8(1):1740. https://doi.org/10.1038/s41467-017-01689-9. Accessed 28 June 2019.

17. Höllt T, Pezzotti N, Unen VV, Koning F, Lelieveldt BPF, Vilanova A. CyteGuide: visual guidance for hierarchical single-cell analysis. IEEE Trans Vis Comput Graph. 2018;24(1):739–48. https://doi.org/10.1109/TVCG.2017.2744318.
18. Welch JD, Hartemink AJ, Prins JF. MATCHER: manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics. Genome Biol. 2017;18(1):138.
19. Moon KR, Stanley JS, Burkhardt D, van Dijk D, Wolf G, Krishnaswamy S. Manifold learning-based methods for analyzing single-cell RNA-sequencing data. Curr Opin Syst Biol. 2018;7:36–46.
20. Hoffer E, Ailon N. Deep metric learning Using triplet network. In: Feragen A, Pelillo M, Loog M, editors. Similarity-Based Pattern Recognition. Lecture Notes in Computer Science. Heidelberg: Springer; 2015. p. 84–92.
21. Bromley J, Bentz JW, Bottou L, Guyon I, Lecun Y, Moore C, Säckinger E, Shah R. Signature verification using a "Siamese" time delay neural network. International Journal of Pattern Recognition and Artificial Intelligence. 1993;07(04):669–88. https://doi.org/10.1142/S0218001493000339. Accessed 28 Mar 2019.
22. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet. 2010;11(10):733–739. https://doi.org/10.1038/nrg2825. Accessed 27 Mar 2019.
23. Severson DT, Owen RP, White MJ, Lu X, Schuster-Böckler B. BEARscc determines robustness of single-cell clusters using simulated technical replicates. Nat Commun. 2018;9(1):1187.
24. Sengupta D, Rayan NA, Lim M, Lim B, Prabhakar S. Fast, scalable and accurate differential expression analysis for single cells. bioRxiv. 2016049734. https://doi.org/10.1101/049734. Accessed 27 Mar 2019.
25. Sinha D, Kumar A, Kumar H, Bandyopadhyay S, Sengupta D. dropclust: efficient clustering of ultra-large scRNA-seq data. Nucleic Acids Res. 2018;46(6):36.
26. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 2018;19(1):15.
27. Iacono G, Mereu E, Guillaumet-Adkins A, Corominas R, Cuscó I, Rodríguez-Esteban G, Gut M, Pérez-Jurado LA, Gut I, Heyn H. bigSCale: an analytical framework for big-scale single-cell data. Genome Res. 2018;28(6):878–90.
28. Fu Y, Li C, Lu S, Zhou W, Tang F, Xie XS, Huang Y. Uniform and accurate single-cell sequencing based on emulsion whole-genome amplification. Proc Natl Acad Sci U S A. 2015;112(38):11923–8.
29. Hosokawa M, Nishikawa Y, Kogawa M, Takeyama H. Massively parallel whole genome amplification for single-cell sequencing using droplet microfluidics. Sci Rep. 2017;7(1):5199.
30. Sidore AM, Lan F, Lim SW, Abate AR. Enhanced sequencing coverage with digital droplet multiple displacement amplification. Nucleic Acids Res. 2016;44(7):66.
31. Picher ÁJ, Budeus B, Wafzig O, Krüger C, García-Gómez S, Martínez-Jiménez MI, Díaz-Talavera A, Weber D, Blanco L, Schneider A. TruePrime is a novel method for whole-genome amplification from single cells based on TthPrimPol. Nat Commun. 2016;7:13296.
32. Potapov V, Ong JL. Examining sources of error in PCR by single-molecule sequencing. PLoS ONE. 2017;12(1):0169774.
33. Xi L, Belyaev A, Spurgeon S, Wang X, Gong H, Aboukhalil R, Fekete R. New library construction method for single-cell genomes. PLoS ONE. 2017;12(7):0181163.
34. Zahn H, Steif A, Laks E, Eirew P, VanInsberghe M, Shah SP, Aparicio S, Hansen CL. Scalable whole-genome single-cell library preparation without preamplification. Nat. Methods. 2017;14(2):167–73.
35. Laks E, Zahn H, Lai D, McPherson A, Steif A, Brimhall J, Biele J, Wang B, Masud T, Grewal D, Nielsen C, Leung S, Bojilova V, Smith M, Golovko O, Poon S, Eirew P, Kabeer F, Algara TRD, Lee SR, Taghiyar MJ, Huebner C, Ngo J, Chan T, Vatrt-Watts S, Walters P, Abrar N, Chan S, Wiens M, Martin L, Scott RW, Underhill MT, Chavez E, Steidl C, Costa DD, Ma Y, Coope RJN, Corbett R, Pleasance S, Moore R, Mungall AJ, Consortium CI, Marra MA, Hansen C, Shah S, Aparicio S. Resource: scalable whole genome sequencing of 40,000 single cells identifies stochastic aneuploidies, genome replication states and clonal repertoires. bioRxiv. 2018411058. https://doi.org/10.1101/411058. Accessed 16 Oct 2018.
36. Chen C, Xing D, Tan L, Li H, Zhou G, Huang L, Xie XS. Single-cell whole-genome analyses by linear amplification via transposon insertion (LIANTI). Science. 2017;356(6334):189–94.
37. Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, Chang HY, Greenleaf WJ. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature. 2015;523(7561):486–90. https://doi.org/10.1038/nature14590. Accessed 30 Apr 2019.
38. Cusanovich DA, Daza R, Adey A, Pliner HA, Christiansen L, Gunderson KL, Steemers FJ, Trapnell C, Shendure J. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. Science (New York). 2015;348(6237):910–4. https://doi.org/10.1126/science.aab1601.
39. Karemaker ID, Vermeulen M. Single-cell DNA methylation profiling: technologies and biological applications. Trends Biotechnol. 2018;36(9):952–65. https://doi.org/10.1016/j.tibtech.2018.04.002. Accessed 30 Apr 2019.
40. Virant-Klun I, Leicht S, Hughes C, Krijgsveld J. Identification of maturation-specific proteins by single-cell proteomics of human oocytes. Mol Cell Proteomics MCP. 2016;15(8):2616–27. https://doi.org/10.1074/mcp.M115.056887.
41. Clark SJ, Argelaguet R, Kapourani C-A, Stubbs TM, Lee HJ, Alda-Catalinas C, Krueger F, Sanguinetti G, Kelsey G, Marioni JC, Stegle O, Reik W. scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. Nat Commun. 2018;9(1):781. https://doi.org/10.1038/s41467-018-03149-4. Accessed 27 Mar 2019.
42. Cao J, Cusanovich DA, Ramani V, Aghamirzaie D, Pliner HA, Hill AJ, Daza RM, McFaline-Figueroa JL, Packer JS, Christiansen L, Steemers FJ, Adey AC, Trapnell C, Shendure J. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. Science. 2018;361(6409):1380–5. https://doi.org/10.1126/science.aau0730. Accessed 30 Apr 2019.
43. Hicks SC, Townes FW, Teng M, Irizarry RA. Missing data and technical variability in single-cell RNA-sequencing experiments. Biostatistics. 2018;19(4):562–78. https://doi.org/10.1093/biostatistics/kxx053. Accessed 27 Mar 2019.
44. Bacher R, Kendziorski C. Design and computational analysis of single-cell RNA-sequencing experiments. Genome Biol. 2016;17(1):63. https://doi.org/10.1186/s13059-016-0927-y. Accessed 27 Mar 2019.
45. Das S, Abecasis GR, Browning BL. Genotype imputation from large reference panels. Annu Rev Genomics Hum Genet. 2018;19(1):73–96. https://doi.org/10.1146/annurev-genom-083117-021602. Accessed 28 Oct 2019.
46. Das S, Abecasis GR, Browning BL. Genotype imputation from large reference panels. Annu Rev Genomics Hum Genet. 2018;19:73–96. https://doi.org/10.1146/annurev-genom-083117-021602.
47. Tang W, Bertaux F, Thomas P, Stefanelli C, Saint M, et al. bayNorm: Bayesian gene expression recovery, imputation and normalisation for single cell RNA-sequencing data. bioRxiv. 2018. https://www.biorxiv.org/content/10.1101/384586v2.abstract.
48. Azizi E, Prabhakaran S, Carr A, Pe'er D. Bayesian inference for single-cell clustering and imputing. Genomics Comput Biol. 2017;3(1):46. https://doi.org/10.18547/gcb.2017.vol3.iss1.e46. Accessed 27 Mar 2019.
49. Lin P, Troup M, Ho JWK. CIDR: ultrafast and accurate clustering through imputation for single-cell RNA-seq data. Genome Biol. 2017;18(1):59. https://doi.org/10.1186/s13059-017-1188-0. Accessed 27 Mar 2019.
50. Huang M, Wang J, Torre E, Dueck H, Shaffer S, Bonasio R, Murray JI, Raj A, Li M, Zhang NR. SAVER: gene expression recovery for single-cell RNA sequencing. Nat Methods. 2018;15(7):539. https://doi.org/10.1038/s41592-018-0033-z. Accessed 27 Mar 2019.
51. Li WV, Li JJ. An accurate and robust imputation method scimpute for single-cell RNA-seq data. Nat Commun. 2018;9(1):997.
52. Miao Z, Li J, Zhang X. scRecover: discriminating true and false zeros in single-cell RNA-seq data for imputation. bioRxiv. 2019665323. https://doi.org/10.1101/665323. Accessed 15 Oct 2019.
53. Chen M, Zhou X. VIPER: variability-preserving imputation for accurate gene expression recovery in single-cell RNA sequencing studies. Genome Biol. 2018;19(1):196.
54. Gong W, Kwak I-Y, Pota P, Koyano-Nakagawa N, Garry DJ. DrImpute: imputing dropout events in single cell RNA sequencing data. BMC Bioinformatics. 2018;19(1):220. https://doi.org/10.1186/s12859-018-2226-y. Accessed 27 Mar 2019.

55. Wagner F, Yan Y, Yanai I. K-nearest neighbor smoothing for high-throughput single-cell RNA-Seq data. bioRxiv. 2018217737. https://doi.org/10.1101/217737. Accessed 15 Oct 2019.
56. Moussa M, Măndoiu II. Locality sensitive imputation for single cell RNA-Seq data. J Comput Biol. 2019. https://doi.org/10.1089/cmb.2018.0236. Accessed 27 July 2019.
57. Dijk DV, Sharma R, Nainys J, Yim K, Kathail P, Carr AJ, Burdziak C, Moon KR, Chaffer CL, Pattabiraman D, Bierie B, Mazutis L, Wolf G, Krishnaswamy S, Pe'er D. Recovering gene interactions from single-cell data using data diffusion. Cell. 2018;174(3):716–72927. https://doi.org/10.1016/j.cell.2018.05.061. Accessed 27 Mar 2019.
58. Jonathan Ronen AA. netsmooth: network-smoothing based imputation for single cell RNA-seq. F1000Res. 2018;7:. https://github.com/BIMSBbioinfo/netSmooth.
59. Linderman GC, Zhao J, Kluger Y. Zero-preserving imputation of scRNA-seq data using low-rank approximation. bioRxiv. 2018. https://www.biorxiv.org/content/10.1101/397588v1.abstract.
60. Wagner F, Barkley D, Yanai I. Accurate denoising of single-cell RNA-Seq data using unbiased principal component analysis. bioRxiv. 2019655365. URL https://doi.org/10.1101/655365. Accessed 15 Nov 2019.
61. Chen C, Wu C, Wu L, Wang Y, Deng M, Xi R. scRMD: imputation for single cell RNA-seq data via robust matrix decomposition. bioRxiv. 2018459404. https://doi.org/10.1101/459404. Accessed 15 Oct 2019.
62. Kotliar D, Veres A, Nagy MA, Tabrizi S, Hodis E, Melton DA, Sabeti PC. Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. Elife. 2019;8:43803.
63. Buettner F, Pratanwanich N, McCarthy DJ, Marioni JC, Stegle O. f-scLVM: scalable and versatile factor analysis for single-cell RNA-seq. Genome Biol. 2017;18(1):212. https://doi.org/10.1186/s13059-017-1334-8.
64. Verma A, Engelhardt BE. A robust nonlinear low-dimensional manifold for single cell RNA-seq data. bioRxiv. 2018443044. https://doi.org/10.1101/443044. Accessed 15 Nov 2019.
65. Durif G, Modolo L, Mold JE, Lambert-Lacroix S, Picard F. Probabilistic count matrix factorization for single cell expression data analysis. Bioinformatics. 2019. https://doi.org/10.1093/bioinformatics/btz177.
66. Stein-O'Brien GL, Clark BS, Sherman T, Zibetti C, Hu Q, Sealfon R, Liu S, Qian J, Colantuoni C, Blackshaw S, Goff LA, Fertig EJ. Decomposing cell identity for transfer learning across cellular measurements, platforms, tissues, and species. Cell Syst. 2019;8(5):395–4118. https://doi.org/10.1016/j.cels.2019.04.004.
67. Jung M, Wells D, Rusch J, Ahmad S, Marchini J, Myers SR, Conrad DF. Unified single-cell analysis of testis gene regulation and pathology in five mouse strains. eLife. 2019;8:. URL https://doi.org/10.7554/eLife.43966.
68. Pierson E, Yau C. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biol. 2015;16(1):241. https://doi.org/10.1186/s13059-015-0805-z. Accessed 27 Mar 2019.
69. Risso D, Perraudeau F, Gribkova S, Dudoit S, Vert J-P. A general and flexible method for signal extraction from single-cell RNA-seq data. Nature Commun. 2018;9(1):284. https://doi.org/10.1038/s41467-017-02554-5.
70. Talwar D, Mongia A, Sengupta D, Majumdar A. AutoImpute: autoencoder based imputation of single-cell RNA-seq data. Sci Rep. 2018;8(1):16329. https://doi.org/10.1038/s41598-018-34688-x.
71. Wang T, Johnson TS, Shao W, Lu Z, Helm BR, Zhang J, Huang K. BERMUDA: a novel deep transfer learning method for single-cell RNA sequencing batch correction reveals hidden high-resolution cellular subtypes. Genome Biol. 2019;20(1):165. https://doi.org/10.1186/s13059-019-1764-6. Accessed 15 Nov 2019.
72. Arisdakessian C, Poirion O, Yunits B, Zhu X, Garmire L. DeepImpute: an accurate, fast and scalable deep neural network method to impute single-cell RNA-Seq data. bioRxiv. 2018. https://www.biorxiv.org/content/10.1101/353607v1.abstract.
73. Eraslan G, Simon LM, Mircea M, Mueller NS, Theis FJ. Single-cell RNA-seq denoising using a deep count autoencoder. Nat Commun. 2019;10(1):390. https://doi.org/10.1038/s41467-018-07931-2. Accessed 27 Mar 2019.
74. Srinivasan S, Johnson NT, Korkin D. A hybrid deep clustering approach for robust cell type profiling using single-cell RNA-seq data. bioRxiv. 2019. https://www.biorxiv.org/content/10.1101/511626v1.abstract.
75. Zhang X-F, Ou-Yang L, Yang S, Zhao X-M, Hu X, Yan H. EnImpute: imputing dropout events in single cell RNA sequencing data via ensemble learning. Bioinformatics. 2019. https://doi.org/10.1093/bioinformatics/btz435.
76. Kinalis S, Nielsen FC, Winther O, Bagger FO. Deconvolution of autoencoders to learn biological regulatory modules from single cell mRNA sequencing data. BMC Bioinformatics. 2019;20(1):379. https://doi.org/10.1186/s12859-019-2952-9.
77. Badsha MB, Li R, Liu B, Li YI, Xian M, Banovich NE, Fu AQ. Imputation of single-cell gene expression with an autoencoder neural network. bioRxiv. 2018504977. https://doi.org/10.1101/504977. Accessed 15 Oct 2019.
78. Lin C, Jain S, Kim H, Bar-Joseph Z. Using neural networks for reducing the dimensions of single-cell RNA-Seq data. Nucleic Acids Res. 2017;45(17):156.
79. Amodio M, Dijk DV, Srinivasan K, Chen WS, Mohsen H, Moon KR, Campbell A, Zhao Y, Wang X, Venkataswamy M, Desai A, Ravi V, Kumar P, Montgomery R, Wolf G, Krishnaswamy S. Exploring single-cell data with deep multitasking neural networks. bioRxiv. 2019237065. https://doi.org/10.1101/237065. Accessed 15 Oct 2019.
80. Deng Y, Bao F, Dai Q, Wu LF, Altschuler SJ. Scalable analysis of cell-type composition from single-cell transcriptomics using deep recurrent learning. Nat Methods. 2019. https://doi.org/10.1038/s41592-019-0353-7.
81. Grønbech CH, Vording MF, Timshel P, Sønderby CK, Pers TH, Winther O. scVAE: Variational auto-encoders for single-cell gene expression data. bioRxiv. 2019318295. https://doi.org/10.1101/318295. Accessed 15 Oct 2019.
82. Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep generative modeling for single-cell transcriptomics. Nat Methods. 2018;15(12):1053–8. https://doi.org/10.1038/s41592-018-0229-2.
83. Ding J, Condon A, Shah SP. Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. Nat Commun. 2018;9(1):2002.
84. Wang D, Gu J. VASC: Dimension reduction and visualization of single-cell RNA-seq data by deep variational autoencoder. Genomics Proteomics Bioinforma. 2018;16(5):320–31.
85. Zhang C. Single-cell data analysis using mmd variational autoencoder for a more informative latent representation. bioRxiv. 2019613414. https://doi.org/10.1101/613414. Accessed 15 Oct 2019.
86. Leote AC, Wu X, Beyer A. Network-based imputation of dropouts in single-cell RNA sequencing data. bioRxiv. 2019611517. URL https://doi.org/10.1101/611517. Accessed 23 Apr 2019.
87. Wang J, Agarwal D, Huang M, Hu G, Zhou Z, Ye C, Zhang NR. Data denoising with transfer learning in single-cell transcriptomics. Nat Methods. 2019;16(9):875–8. https://doi.org/10.1038/s41592-019-0537-1. Accessed 15 Oct 2019.
88. Peng T, Zhu Q, Yin P, Tan K. SCRABBLE: single-cell RNA-seq imputation constrained by bulk RNA-seq data. Genome Biol. 2019;20(1):88. https://doi.org/10.1186/s13059-019-1681-8.
89. Zhu L, Lei J, Devlin B, Roeder K. A unified statistical framework for single cell and bulk RNA sequencing data. Ann Appl Stat. 2018;12(1):609–32. https://doi.org/10.1214/17-AOAS1110. Accessed 15 Nov 2019.
90. Andrews TS, Hemberg M. False signals induced by single-cell imputation. F1000Research. 2019;7:1740. https://doi.org/10.12688/f1000research.16613.2. Accessed 28 June 2019.
91. Campbell KR, Steif A, Laks E, Zahn H, Lai D, McPherson A, Farahani H, Kabeer F, O'Flanagan C, Biele J, Brimhall J, Wang B, Walters P, Consortium I, Bouchard-Côté A, Aparicio S, Shah SP. clonealign: statistical integration of independent single-cell RNA and DNA sequencing data from human cancers. Genome Biol. 2019;20(1):54. https://doi.org/10.1186/s13059-019-1645-z. Accessed 27 Mar 2019.
92. Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, Buettner F, Huber W, Stegle O. Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. Mol Syst Biol. 2018;14(6):8124. https://doi.org/10.15252/msb.20178124. Accessed 27 Mar 2019.
93. Zhang L, Zhang S. Comparison of computational methods for imputing single-cell RNA-sequencing data. IEEE/ACM Trans Comput Biol Bioinforma. 20181. https://doi.org/10.1109/TCBB.2018.2848633.
94. Hu Q, Greene CS. Parameter tuning is a key part of dimensionality reduction via deep variational autoencoders for single cell RNA transcriptomics. Pac Symp Biocomput. 2019;24:362–73.
95. Sun S, Zhu J, Ma Y, Zhou X. Accuracy, robustness and scalability of dimensionality reduction methods for single cell RNAseq analysis. bioRxiv. 2019. https://www.biorxiv.org/content/10.1101/641142v1.abstract.

96. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. bioRxiv. 2019576827. https://doi.org/10.1101/576827. Accessed 15 Oct 2019.
97. Townes FW, Hicks SC, Aryee MJ, Irizarry RA. Feature selection and dimension reduction for single cell RNA-Seq based on a multinomial model. bioRxiv. 2019574574. https://doi.org/10.1101/574574. Accessed 15 Oct 2019.
98. Kang HM, Subramaniam M, Targ S, Nguyen M, Maliskova L, McCarthy E, Wan E, Wong S, Byrnes L, Lanata CM, Gate RE, Mostafavi S, Marson A, Zaitlen N, Criswell LA, Ye CJ. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. Nat Biotechnol. 2018;36(1):89–94.
99. Stubbington MJT, Rozenblatt-Rosen O, Regev A, Teichmann SA. Single-cell transcriptomics to explore the immune system in health and disease. Science. 2017;358(6359):58–63.
100. Karaiskos N, Wahle P, Alles J, Boltengagen A, Ayoub S, Kipar C, Kocks C, Rajewsky N, Zinzen RP. The drosophila embryo at single-cell transcriptome resolution. Science. 2017;358(6360):194–9.
101. Kim K-T, Lee HW, Lee H-O, Kim SC, Seo YJ, Chung W, Eum HH, Nam D-H, Kim J, Joo KM, Park W-Y. Single-cell mRNA sequencing identifies subclonal heterogeneity in anti-cancer drug responses of lung adenocarcinoma cells. Genome Biol. 2015;16:127.
102. Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nat Methods. 2014;11(7):740–2. https://doi.org/10.1038/nmeth.2967.
103. Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, Slichter CK, Miller HW, McElrath MJ, Prlic M, Linsley PS, Gottardo R. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol. 2015;16:. https://doi.org/10.1186/s13059-015-0844-5. Accessed 27 Mar 2019.
104. Soneson C, Robinson MD. Bias, robustness and scalability in single-cell differential expression analysis. Nat Methods. 2018;15(4):255–61. https://doi.org/10.1038/nmeth.4612.
105. Eling N, Richard AC, Richardson S, Marioni JC, Vallejos CA. Correcting the mean-variance dependency for differential variability testing using single-cell RNA sequencing data. Cell Syst. 2018;7(3):284–29412. https://doi.org/10.1016/j.cels.2018.06.011. Accessed 27 Mar 2019.
106. Campbell KR, Yau C. Uncovering pseudotemporal trajectories with covariates from single cell and bulk expression data. Nat Commun. 2018;9(1):2442. https://doi.org/10.1038/s41467-018-04696-6. Accessed 27 Mar 2019.
107. van den Berge K, Bezieux HRD, Street K, Saelens W, Cannoodt R, Saeys Y, Dudoit S, Clement L. Trajectory-based differential expression analysis for single-cell sequencing data. bioRxiv. 2019623397. https://doi.org/10.1101/623397. Accessed 03 May 2019.
108. Korthauer KD, Chu L-F, Newton MA, Li Y, Thomson J, Stewart R, Kendziorski C. A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. Genome Biol. 2016;17(1):222. https://doi.org/10.1186/s13059-016-1077-y.
109. L. Lun A. T, Bach K, Marioni JC. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biol. 2016;17(1):75. https://doi.org/10.1186/s13059-016-0947-7. Accessed 23 Oct 2019.
110. Kang HM, Subramaniam M, Targ S, Nguyen M, Maliskova L, McCarthy E, Wan E, Wong S, Byrnes L, Lanata C, Gate R, Mostafavi S, Marson A, Zaitlen N, Criswell LA, Ye CJ. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. Nat Biotechnol. 2018;36(1):89–94. https://doi.org/10.1038/nbt.4042. Accessed 27 Mar 2019.
111. Crowell HL, Soneson C, Germain P-L, Calini D, Collin L, Raposo C, Malhotra D, Robinson MD. On the discovery of population-specific state transitions from multi-sample multi-condition single-cell RNA sequencing data. bioRxiv. 2019713412. https://doi.org/10.1101/713412. Accessed 23 Oct 2019.
112. Tung P-Y, Blischak JD, Hsiao CJ, Knowles DA, Burnett JE, Pritchard JK, Gilad Y. Batch effects and the effective design of single-cell gene expression studies. Sci Rep. 2017;7:39921. https://doi.org/10.1038/srep39921. Accessed 23 Oct 2019.
113. Vavoulis DV, Francescatto M, Heutink P, Gough J. DGEclust: differential expression analysis of clustered count data. Genome Biol. 2015;16:39.
114. Reid S, Taylor J, Tibshirani R. A general framework for estimation and inference from clusters of features. J Am Stat Assoc. 2018;113(521):280–93.
115. Zhang JM, Kamath GM, Tse DN. Valid post-clustering differential analysis for single-cell RNA-Seq. bioRxiv. 2019463265. https://doi.org/10.1101/463265. Accessed 09 July 2019.
116. Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. Nat Rev Genet. 2015;16(3):133.
117. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018;36(5):411–20.
118. Haghverdi L, Lun ATL, Morgan MD, Marioni JC. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat Biotechnol. 2018;36(5):421–7. https://doi.org/10.1038/nbt.4091.
119. Manno GL, Soldatov R, Zeisel A, Braun E, Hochgerner H, Petukhov V, Lidschreiber K, Kastriti ME, Lönnerberg P, Furlan A, Fan J, Borm LE, Liu Z, Bruggen DV, Guo J, He X, Barker R, Sundström E, Castelo-Branco G, Cramer P, Adameyko I, Linnarsson S, Kharchenko PV. RNA velocity of single cells. Nature. 2018;560(7719):494. https://doi.org/10.1038/s41586-018-0414-6. Accessed 28 Mar 2019.
120. Lun ATL, Richard AC, Marioni JC. Testing for differential abundance in mass cytometry data. Nat. Methods. 2017;14(7):707–9.
121. Bruggner RV, Bodenmiller B, Dill DL, Tibshirani RJ, Nolan GP. Automated identification of stratifying signatures in cellular subpopulations. Proc Natl Acad Sci U S A. 2014;111(26):2770–7.
122. Weber LM, Nowicka M, Soneson C, Robinson MD. diffcyt: differential discovery in high-dimensional cytometry via high-resolution clustering. bioRxiv. 2018349738. https://doi.org/10.1101/349738. Accessed 28 Mar 2019.
123. Nowicka M, Krieg C, Weber LM, Hartmann FJ, Guglietta S, Becher B, Levesque MP, Robinson MD. CyTOF workflow: differential discovery in high-throughput high-dimensional cytometry datasets. F1000Res. 2017;6:748.
124. Arvaniti E, Claassen M. Sensitive detection of rare disease-associated cell subsets via representation learning. Nat Commun. 2017;8(1):1–10. https://doi.org/10.1038/ncomms14825. Accessed 23 Oct 2019.
125. Duò A, Robinson MD, Soneson C. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. F1000Res. 2018;7:1141. https://doi.org/10.12688/f1000research.15666.2.
126. Freytag S, Tian L, Lönnstedt I, Ng M, Bahlo M. Comparison of clustering tools in R for medium-sized 10x Genomics single-cell RNA-sequencing data. F1000Research. 2018;7:. https://doi.org/10.12688/f1000research.15809.2. Accessed 07 Feb 2019.
127. Kiselev VY, Andrews TS, Hemberg M. Challenges in unsupervised clustering of single-cell RNA-seq data. Nat Rev Genet. 20191. https://doi.org/10.1038/s41576-018-0088-9. Accessed 03 Apr 2019.
128. Cao J, Packer JS, Ramani V, Cusanovich DA, Huynh C, Daza R, Qiu X, Lee C, Furlan SN, Steemers FJ, Adey A, Waterston RH, Trapnell C, Shendure J. Comprehensive single-cell transcriptional profiling of a multicellular organism. Science. 2017;357(6352):661–7. https://doi.org/10.1126/science.aam8940. Accessed 03 Apr 2019.
129. Fincher CT, Wurtzel O, Hoog TD, Kravarik KM, Reddien PW. Cell type transcriptome atlas for the planarian Schmidtea mediterranea. Science. 2018;360(6391):1736. https://doi.org/10.1126/science.aaq1736. Accessed 03 Apr 2019.
130. Plass M, Solana J, Wolf FA, Ayoub S, Misios A, Glažar P, Obermayer B, Theis FJ, Kocks C, Rajewsky N. Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. Science. 2018;360(6391):1723. https://doi.org/10.1126/science.aaq1723. Accessed 03 Apr 2019.
131. Karaiskos N, Wahle P, Alles J, Boltengagen A, Ayoub S, Kipar C, Kocks C, Rajewsky N, Zinzen RP. The Drosophila embryo at single-cell transcriptome resolution. Science. 2017;358(6360):194–9. https://doi.org/10.1126/science.aan3235. Accessed 03 Apr 2019.
132. Farrell JA, Wang Y, Riesenfeld SJ, Shekhar K, Regev A, Schier AF. Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. Science. 2018;360(6392):3131. https://doi.org/10.1126/science.aar3131. Accessed 03 Apr 2019.
133. Wagner DE, Weinreb C, Collins ZM, Briggs JA, Megason SG, Klein AM. Single-cell mapping of gene expression landscapes and lineage in the

zebrafish embryo. Science. 2018;360(6392):981–7. https://doi.org/10.1126/science.aar4362. Accessed 03 Apr 2019.
134. Briggs JA, Weinreb C, Wagner DE, Megason S, Peshkin L, Kirschner MW, Klein AM. The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. Science. 2018;360(6392):5780. https://doi.org/10.1126/science.aar5780. Accessed 03 Apr 2019.
135. Rosenberg AB, Roco CM, Muscat RA, Kuchina A, Sample P, Yao Z, Graybuck LT, Peeler DJ, Mukherjee S, Chen W, Pun SH, Sellers DL, Tasic B, Seelig G. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. Science. 2018;360(6385):176–82. URL https://doi.org/10.1126/science.aam8999. Accessed 03 Apr 2019.
136. Saunders A, Macosko EZ, Wysoker A, Goldman M, Krienen FM, de Rivera H, Bien E, Baum M, Bortolin L, Wang S, Goeva A, Nemesh J, Kamitaki N, Brumbaugh S, Kulp D, McCarroll SA. Molecular diversity and specializations among the cells of the adult mouse brain. Cell. 2018;174(4):1015–103016. https://doi.org/10.1016/j.cell.2018.07.028. Accessed 03 Apr 2019.
137. Zeisel A, Hochgerner H, Lönnerberg P, Johnsson A, Memic F, van der Zwan J, Häring M, Braun E, Borm LE, La Manno G, Codeluppi S, Furlan A, Lee K, Skene N, Harris KD, Hjerling-Leffler J, Arenas E, Ernfors P, Marklund U, Linnarsson S. Molecular architecture of the mouse nervous system. Cell. 2018;174(4):999–101422. https://doi.org/10.1016/j.cell.2018.06.021. Accessed 03 Apr 2019.
138. Tabula Muris Consortium T. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. Nature. 2018;562(7727):367. https://doi.org/10.1038/s41586-018-0590-4. Accessed 03 Apr 2019.
139. Han X, Wang R, Zhou Y, Fei L, Sun H, Lai S, Saadatpour A, Zhou Z, Chen H, Ye F, Huang D, Xu Y, Huang W, Jiang M, Jiang X, Mao J, Chen Y, Lu C, Xie J, Fang Q, Wang Y, Yue R, Li T, Huang H, Orkin SH, Yuan G-C, Chen M, Guo G. Mapping the mouse cell atlas by Microwell-Seq. Cell. 2018;172(5):1091–110717.
140. Schiller HB, Montoro DT, Simon LM, Rawlins EL, Meyer KB, Strunz M, Vieira Braga F, Timens W, Koppelman GH, Budinger GRS, Burgess JK, Waghray A, van den Berge M, Theis FJ, Regev A, Kaminski N, Rajagopal J, Teichmann SA, Misharin AV, Nawijn MC. The Human Lung Cell Atlas - a high-resolution reference map of the human lung in health and disease. Am J Respir Cell Mol Biol. 2019. https://doi.org/10.1165/rcmb.2018-0416TR. Accessed 29 Apr 2019.
141. Lieberman Y, Rokach L, Shay T. CaSTLe – classification of single cells by transfer learning: harnessing the power of publicly available single cell RNA sequencing experiments to annotate new experiments. PLoS ONE. 2018;13(10):0205499. https://doi.org/10.1371/journal.pone.0205499. Accessed 03 Apr 2019.
142. Srivastava D, Iyer A, Kumar V, Sengupta D. CellAtlasSearch: a scalable search engine for single cells. Nucleic Acids Res. 2018;46(W1):141–7. https://doi.org/10.1093/nar/gky421. Accessed 03 Apr 2019.
143. Cao Z-J, Wei L, Lu S, Yang D-C, Gao G. Cell BLAST: searching large-scale scRNA-seq database via unbiased cell embedding. bioRxiv. 2019587360. https://doi.org/10.1101/587360. Accessed 03 Apr 2019.
144. DePasquale EA, Ferchen K, Hay S, Grimes HL, Salomonis N. cellHarmony: cell-level matching and comparison of single-cell transcriptomes. bioRxiv. 2019412080. https://doi.org/10.1101/412080. Accessed 04 Apr 2019.
145. Kanter J. K. d., Lijnzaad P, Candelli T, Margaritis T, Holstege F. CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing. bioRxiv. 2019558908. https://doi.org/10.1101/558908. Accessed 01 Apr 2019.
146. Sato K, Tsuyuzaki K, Shimizu K, Nikaido I. CellFishing.jl: an ultrafast and scalable cell search method for single-cell RNA sequencing. Genome Biol. 2019;20(1):31. https://doi.org/10.1186/s13059-019-1639-x. Accessed 03 Apr 2019.
147. Zhang AW, O'Flanagan C, Chavez E, Lim JL, McPherson A, Wiens M, Walters P, Chan T, Hewitson B, Lai D, Mottok A, Sarkozy C, Chong L, Aoki T, Wang X, Weng AP, McAlpine JN, Aparicio S, Steidl C, Campbell KR, Shah SP. Probabilistic cell type assignment of single-cell transcriptomic data reveals spatiotemporal microenvironment dynamics in human cancers. bioRxiv. 2019521914. https://doi.org/10.1101/521914. Accessed 12 Mar 2019.
148. Abdelaal T, Michielsen L, Cats D, Hoogduin D, Mei H, Reinders MJT, Mahfouz A. A comparison of automatic cell identification methods for single-cell RNA sequencing data. Genome Biol. 2019;20(1):194. https://doi.org/10.1186/s13059-019-1795-z. Accessed 23 Oct 2019.
149. Chester C, Maecker HT. Algorithmic tools for mining High-Dimensional cytometry data. J Immunol. 2015;195(3):773–9.
150. Weber LM, Robinson MD. Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data. Cytom A. 2016;89(12):1084–96. https://doi.org/10.1002/cyto.a.23030. Accessed 30 Apr 2019.
151. Saeys Y, Van Gassen S, Lambrecht BN. Computational flow cytometry: helping to make sense of high-dimensional immunology data. Nat Rev Immunol. 2016;16(7):449–62. https://doi.org/10.1038/nri.2016.56. Accessed 30 Apr 2019.
152. Guilliams M, Dutertre C-A, Scott C, McGovern N, Sichien D, Chakarov S, Van Gassen S, Chen J, Poidinger M, De Prijck S, Tavernier S, Low I, Irac S, Mattar C, Sumatoh H, Low G, Chung T, Chan D, Tan K, Hon T, Fossum E, Bogen B, Choolani M, Chan J, Larbi A, Luche H, Henri S, Saeys Y, Newell E, Lambrecht B, Malissen B, Ginhoux F. Unsupervised high-dimensional analysis aligns dendritic cells across tissues and species. Immunity. 2016;45(3):669–84. https://doi.org/10.1016/j.immuni.2016.08.015. Accessed 30 Apr 2019.
153. Hon C-C, Shin JW, Carninci P, Stubbington MJT. The Human Cell Atlas: technical approaches and challenges. Brief Funct Genom. 2018;17(4):283–94. https://doi.org/10.1093/bfgp/elx029. Accessed 05 June 2019.
154. Spanjaard B, Hu B, Mitic N, Olivares-Chauvet P, Janjuha S, Ninov N, Junker JP. Simultaneous lineage tracing and cell-type identification using CRISPR-Cas9-induced genetic scars. Nat Biotechnol. 2018;36(5):469–73.
155. Kester L, van Oudenaarden A. Single-cell transcriptomics meets lineage tracing. Cell Stem Cell. 2018;23(2):166–79. https://doi.org/10.1016/j.stem.2018.04.014. Accessed 20 Nov 2019.
156. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, Lennon NJ, Livak KJ, Mikkelsen TS, Rinn JL. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol. 2014;32(4):381–6. https://doi.org/10.1038/nbt.2859.
157. Saelens W, Cannoodt R, Todorov H, Saeys Y. A comparison of single-cell trajectory inference methods. Nat Biotechnol. 20191. https://doi.org/10.1038/s41587-019-0071-9. Accessed 30 Apr 2019.
158. Ji Z, Ji H. TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. Nucleic Acids Res. 2016;44(13):117. https://doi.org/10.1093/nar/gkw430.
159. Qiu X, Mao Q, Tang Y, Wang L, Chawla R, Pliner HA, Trapnell C. Reversed graph embedding resolves complex single-cell trajectories. Nat Methods. 2017;14(10):979–82. https://doi.org/10.1038/nmeth.4402. Accessed 30 Apr 2019.
160. Chen H, Albergante L, Hsu JY, Lareau CA, Bosco GL, Guan J, Zhou S, Gorban AN, Bauer DE, Aryee MJ, Langenau DM, Zinovyev A, Buenrostro JD, Yuan G-C, Pinello L. Single-cell trajectories reconstruction, exploration and mapping of omics data with STREAM. Nat Commun. 2019;10(1):1903. https://doi.org/10.1038/s41467-019-09670-4. Accessed 30 Apr 2019.
161. Rizvi AH, Camara PG, Kandror EK, Roberts TJ, Schieren I, Maniatis T, Rabadan R. Single-cell topological RNA-seq analysis reveals insights into cellular differentiation and development. Nat Biotechnol. 2017;35(6):551–60. https://doi.org/10.1038/nbt.3854.
162. Haghverdi L, Büttner M, Wolf FA, Buettner F, Theis FJ. Diffusion pseudotime robustly reconstructs lineage branching. Nat Methods. 2016;13(10):845–8. https://doi.org/10.1038/nmeth.3971. Accessed 30 Apr 2019.
163. Setty M, Tadmor MD, Reich-Zeliger S, Angel O, Salame TM, Kathail P, Choi K, Bendall S, Friedman N, Pe'er D. Wishbone identifies bifurcating developmental trajectories from single-cell data. Nat Biotechnol. 2016;34(6):637–45. https://doi.org/10.1038/nbt.3569. Accessed 30 Apr 2019.
164. Schiebinger G, Shu J, Tabaka M, Cleary B, Subramanian V, Solomon A, Liu S, Lin S, Berube P, Lee L, Chen J, Brumbaugh J, Rigollet P, Hochedlinger K, Jaenisch R, Regev A, Lander ES. Reconstruction of developmental landscapes by optimal-transport analysis of single-cell gene expression sheds light on cellular reprogramming. bioRxiv. 2017191056. https://doi.org/10.1101/191056. Accessed 30 Apr 2019.
165. Weinreb C, Wolock S, Tusi BK, Socolovsky M, Klein AM. Fundamental limits on dynamic inference from single-cell snapshots. Proc Natl Acad

Sci. 2018;115(10):2467–76. https://doi.org/10.1073/pnas.1714723115. Accessed 30 Apr 2019.
166. Campbell KR, Yau C. Order under uncertainty: robust differential expression analysis using probabilistic models for pseudotime inference. PLoS Comput Biol. 2016;12(11):1005212. https://doi.org/10.1371/journal.pcbi.1005212. Accessed 09 July 2019.
167. Reid JE, Wernisch L. Pseudotime estimation: deconfounding single cell time series. Bioinformatics. 2016;32(19):2973–80. https://doi.org/10.1093/bioinformatics/btw372. Accessed 09 July 2019.
168. Ahmed S, Rattray M, Boukouvalas A. GrandPrix: scaling up the Bayesian GPLVM for single-cell data. Bioinformatics. 2019;35(1):47–54. https://doi.org/10.1093/bioinformatics/bty533. Accessed 09 July 2019.
169. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, Trombetta JJ, Weitz DA, Sanes JR, Shalek AK, Regev A, McCarroll SA. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell. 2015;161(5):1202–14. https://doi.org/10.1016/j.cell.2015.05.002.
170. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA, Kirschner MW. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell. 2015;161(5):1187–201. https://doi.org/10.1016/j.cell.2015.04.044.
171. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J, Gregory MT, Shuga J, Montesclaros L, Underwood JG, Masquelier DA, Nishimura SY, Schnall-Levin M, Wyatt PW, Hindson CM, Bharadwaj R, Wong A, Ness KD, Beppu LW, Deeg HJ, McFarland C, Loeb KR, Valente WJ, Ericson NG, Stevens EA, Radich JP, Mikkelsen TS, Hindson BJ, Bielas JH. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017;8:14049. https://doi.org/10.1038/ncomms14049. Accessed 30 Apr 2019.
172. Angermueller C, Lee HJ, Reik W, Stegle O. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. Genome Biol. 2017;18(1):67. https://doi.org/10.1186/s13059-017-1189-z. Accessed 30 Apr 2019.
173. Cusanovich DA, Reddington JP, Garfield DA, Daza RM, Aghamirzaie D, Marco-Ferreres R, Pliner HA, Christiansen L, Qiu X, Steemers FJ, Trapnell C, Shendure J, Furlong EEM. The cis-regulatory dynamics of embryonic development at single-cell resolution. Nature. 2018;555(7697):538–42. https://doi.org/10.1038/nature25981. Accessed 30 Apr 2019.
174. Buenrostro JD, Corces MR, Lareau CA, Wu B, Schep AN, Aryee MJ, Majeti R, Chang HY, Greenleaf WJ. Integrated single-cell analysis maps the continuous regulatory landscape of human hematopoietic differentiation. Cell. 2018;173(6):1535–154816. https://doi.org/10.1016/j.cell.2018.03.074.
175. de Boer CG, Regev A. BROCKMAN: deciphering variance in epigenomic regulators by k-mer factorization. BMC Bioinformatics. 2018;19(1):253. https://doi.org/10.1186/s12859-018-2255-6. Accessed 30 Apr 2019.
176. Pliner HA, Packer JS, McFaline-Figueroa JL, Cusanovich DA, Daza RM, Aghamirzaie D, Srivatsan S, Qiu X, Jackson D, Minkina A, Adey AC, Steemers FJ, Shendure J, Trapnell C. Cicero predicts cis-regulatory DNA interactions from single-cell chromatin accessibility data. Mol Cell. 2018;71(5):858–8718. https://doi.org/10.1016/j.molcel.2018.06.044.
177. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018;36(5):411–20. https://doi.org/10.1038/nbt.4096. Accessed 30 Apr 2019.
178. Welch J, Kozareva V, Ferreira A, Vanderburg C, Martin C, Macosko E. Integrative inference of brain cell similarities and differences from single-cell genomics. bioRxiv. 2018459891. https://doi.org/10.1101/459891. Accessed 30 Apr 2019.
179. Crosetto N, Bienko M, van Oudenaarden A. Spatially resolved transcriptomics and beyond. Nat Rev Genet. 2015;16(1):57–66.
180. Strell C, Hilscher MM, Laxman N, Svedlund J, Wu C, Yokota C, Nilsson M. Placing RNA in context and space - methods for spatially resolved transcriptomics. FEBS J. 2018;286(8):1468–81. https://doi.org/10.1111/febs.14435.
181. Moffitt JR, Bambah-Mukku D, Eichhorn SW, Vaughn E, Shekhar K, Perez JD, Rubinstein ND, Hao J, Regev A, Dulac C, Zhuang X. Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. Science. 2018;362(6416):5324. https://doi.org/10.1126/science.aau5324. Accessed 27 Mar 2019.
182. Tanay A, Regev A. Scaling single-cell genomics from phenomenology to mechanism. Nature. 2017;541(7637):331–8.
183. Ståhl PL, Salmén F, Vickovic S, Lundmark A, Navarro JF, Magnusson J, Giacomello S, Asp M, Westholm JO, Huss M, Mollbrink A, Linnarsson S, Codeluppi S, Borg k., Pontén F, Costea PI, Sahlén P, Mulder J, Bergmann O, Lundeberg J, Frisén J. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science (New York). 2016;353(6294):78–82. https://doi.org/10.1126/science.aaf2403.
184. Medaglia C, Giladi A, Stoler-Barak L, Giovanni MD, Salame TM, Biram A, David E, Li H, Iannacone M, Shulman Z, Amit I. Spatial reconstruction of immune niches by combining photoactivatable reporters and scRNA-seq. Science. 2017;358(6370):1622–6. https://doi.org/10.1126/science.aao4277. Accessed 27 Mar 2019.
185. Rodriques SG, Stickels RR, Goeva A, Martin CA, Murray E, Vanderburg CR, Welch J, Chen LM, Chen F, Macosko EZ. Slide-seq: a scalable technology for measuring genome-wide expression at high spatial resolution. Science. 2019;363(6434):1463–7. https://doi.org/10.1126/science.aaw1219. Accessed 16 Apr 2019.
186. Ke R, Mignardi M, Pacureanu A, Svedlund J, Botling J, Wählby C, Nilsson M. In situ sequencing for RNA analysis in preserved tissue and cells. Nat Methods. 2013;10(9):857–60. https://doi.org/10.1038/nmeth.2563. Accessed 10 Oct 2019.
187. Lee JH, Daugharthy ER, Scheiman J, Kalhor R, Ferrante TC, Terry R, Turczyk BM, Yang JL, Lee HS, Aach J, Zhang K, Church GM. Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. Nat Protoc. 2015;10(3):442–58.
188. Wang X, Allen WE, Wright MA, Sylwestrak EL, Samusik N, Vesuna S, Evans K, Liu C, Ramakrishnan C, Liu J, Nolan GP, Bava F-A, Deisseroth K. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. Science. 2018;361(6400):5691. https://doi.org/10.1126/science.aat5691. Accessed 14 Oct 2019.
189. Lubeck E, Coskun AF, Zhiyentayev T, Ahmad M, Cai L. Single-cell in situ RNA profiling by sequential hybridization. Nat Methods. 2014;11(4):360–1.
190. Chen KH, Boettiger AN, Moffitt JR, Wang S, Zhuang X. RNA imaging spatially resolved, highly multiplexed RNA profiling in single cells. Science. 2015;348(6233):6090.
191. Moffitt JR, Hao J, Bambah-Mukku D, Lu T, Dulac C, Zhuang X. High-performance multiplexed fluorescence in situ hybridization in culture and tissue with matrix imprinting and clearing. Proc Natl Acad Sci U S A. 2016;113(50):14456–61.
192. Shah S, Lubeck E, Zhou W, Cai L. In situ transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus. Neuron. 2016;92(2):342–57. https://doi.org/10.1016/j.neuron.2016.10.001. Accessed 10 Oct 2019.
193. Eng C-HL, Lawson M, Zhu Q, Dries R, Koulena N, Takei Y, Yun J, Cronin C, Karp C, Yuan G-C, Cai L. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. Nature. 2019;568(7751):235. URL https://doi.org/10.1038/s41586-019-1049-y. Accessed 16 Apr 2019.
194. Codeluppi S, Borm LE, Zeisel A, Manno GL, Lunteren JAV, Svensson CI, Linnarsson S. Spatial organization of the somatosensory cortex revealed by osmFISH. Nat Methods. 2018;15(11):932–5. https://doi.org/10.1038/s41592-018-0175-z. Accessed 14 Oct 2019.
195. Giesen C, Wang HAO, Schapiro D, Zivanovic N, Jacobs A, Hattendorf B, Schüffler P. J, Grolimund D, Buhmann JM, Brandt S, Varga Z, Wild PJ, Günther D, Bodenmiller B. Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. Nat Methods. 2014;11(4):417–22. https://doi.org/10.1038/nmeth.2869. Accessed 27 Mar 2019.
196. Angelo M, Bendall SC, Finck R, Hale MB, Hitzman C, Borowsky AD, Levenson RM, Lowe JB, Liu SD, Zhao S, Natkunam Y, Nolan GP. Multiplexed ion beam imaging of human breast tumors. Nat Med. 2014;20(4):436–42. https://doi.org/10.1038/nm.3488. Accessed 15 Nov 2019.
197. Lin J-R, Izar B, Wang S, Yapp C, Mei S, Shah PM, Santagata S, Sorger PK. Highly multiplexed immunofluorescence imaging of human tissues and tumors using t-CyCIF and conventional optical microscopes. eLife. 2018;7:31657. https://doi.org/10.7554/eLife.31657. Accessed 14 Oct 2019.
198. Saka SK, Wang Y, Kishi JY, Zhu A, Zeng Y, Xie W, Kirli K, Yapp C, Cicconet M, Beliveau BJ, Lapan SW, Yin S, Lin M, Boyden ES, Kaeser PS, Pihan G, Church GM, Yin P. Immuno-SABER enables highly multiplexed

and amplified protein imaging in tissues. Nat Biotechnol. 2019;37(9): 1080–90. https://doi.org/10.1038/s41587-019-0207-y. Accessed 14 Oct 2019.

199. Goltsev Y, Samusik N, Kennedy-Darling J, Bhate S, Hale M, Vazquez G, Black S, Nolan GP. Deep profiling of mouse splenic architecture with CODEX multiplexed imaging. Cell. 2018;174(4):968–98115. https://doi.org/10.1016/j.cell.2018.07.010. Accessed 14 Oct 2019.
200. Merritt CR, Ong GT, Church S, Barker K, Geiss G, Hoang M, Jung J, Liang Y, McKay-Fleisch J, Nguyen K, Sorg K, Sprague I, Warren C, Warren S, Zhou Z, Zollinger DR, Dunaway DL, Mills GB, Beechem JM. High multiplex, digital spatial profiling of proteins and RNA in fixed tissue using genomic detection methods. bioRxiv. 2019559021. https://doi.org/10.1101/559021. Accessed 01 Aug 2019.
201. Van TM, Blank CU. A user's perspective on GeoMxTM digital spatial profiling. Immuno-Oncol Technol. 2019;1:11–18. https://doi.org/10.1016/j.iotech.2019.05.001. Accessed 01 Aug 2019.
202. Kiselev VY, Yiu A, Hemberg M. scmap: projection of single-cell RNA-seq data across data sets. Nat Methods. 2018;15(5):359–62. https://doi.org/10.1038/nmeth.4644. Accessed 27 Mar 2019.
203. Shivanandan A, Unnikrishnan J, Radenovic A. On characterizing protein spatial clusters with correlation approaches. Sci Rep. 2016;6:31164.
204. Shah S, Lubeck E, Zhou W, Cai L. In situ transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus. Neuron. 2016;92(2):342–57.
205. Edsgärd D, Johnsson P, Sandberg R. Identification of spatial expression trends in single-cell gene expression data. Nat Methods. 2018;15(5): 339–42.
206. Jacobsen M. Point process theory and applications: marked point and piecewise deterministic processes. Basel: Springer Science & Business Media; 2005.
207. Svensson V, Teichmann SA, Stegle O. SpatialDE: identification of spatially variable genes. Nat Methods. 2018;15(5):343–6.
208. Fridman WH, Galon J, Dieu-Nosjean M-C, Cremer I, Fisson S, Damotte D, Pagès F, Tartour E, Sautès-Fridman C. Immune infiltration in human cancer: prognostic significance and disease control. In: Dranoff G, editor. Cancer Immunology and Immunotherapy. Berlin, Heidelberg: Springer; 2011. p. 1–24.
209. Swanton C. Intratumor heterogeneity: evolution through space and time. Cancer Res. 2012;72(19):4875–82.
210. Cretu A, Brooks PC. Impact of the non-cellular tumor microenvironment on metastasis: potential therapeutic and imaging opportunities. J Cell Physiol. 2007;213(2):391–402.
211. Köster J, Brown M, Liu XS. A Bayesian model for single cell transcript expression analysis on MERFISH data. Bioinformatics. 2019;35(6): 995–1001. https://doi.org/10.1093/bioinformatics/bty718. Accessed 15 Nov 2019.
212. McGranahan N, Swanton C. Clonal heterogeneity and tumor evolution: past, present, and the future. Cell. 2017;168(4):613–28.
213. de Bourcy CFA, De Vlaminck I, Kanbar JN, Wang J, Gawad C, Quake SR. A quantitative comparison of single-cell whole genome amplification methods. PLoS ONE. 2014;9(8):105585.
214. Hou Y, Wu K, Shi X, Li F, Song L, Wu H, Dean M, Li G, Tsang S, Jiang R, Zhang X, Li B, Liu G, Bedekar N, Lu N, Xie G, Liang H, Chang L, Wang T, Chen J, Li Y, Zhang X, Yang H, Xu X, Wang L, Wang J. Comparison of variations detection between whole-genome amplification methods used in single-cell resequencing. Gigascience. 2015;4:37.
215. Huang L, Ma F, Chapman A, Lu S, Xie XS. Single-cell whole-genome amplification and sequencing: methodology and applications. Annu Rev Genomics Hum Genet. 2015;16:79–102.
216. Estévez-Gómez N, Prieto T, Guillaumet-Adkins A, Heyn H, Prado-López S, Posada D. Comparison of single-cell whole-genome amplification strategies. bioRxiv. 2018443754. https://doi.org/10.1101/443754. Accessed 27 July 2019.
217. Telenius H, Carter NP, Bebb CE, Nordenskjöld M, Ponder BA, Tunnacliffe A. Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer. Genomics. 1992;13(3):718–25.
218. Zhang L, Cui X, Schmitt K, Hubert R, Navidi W, Arnheim N. Whole genome amplification from a single cell: implications for genetic analysis. Proc Natl Acad Sci U S A. 1992;89(13):5847–51.
219. Klein CA, Schmidt-Kittler O, Schardt JA, Pantel K, Speicher MR, Riethmüller G. Comparative genomic hybridization, loss of heterozygosity, and DNA sequence analysis of single cells. Proc Natl Acad Sci U S A. 1999;96(8):4494–9.
220. Arneson N, Hughes S, Houlston R, Done S. Whole-genome amplification by improved primer extension preamplification PCR (I-PEP-PCR). Cold Spring Harb Protocol. 2008;2008(1):4921. https://doi.org/10.1101/pdb.prot4921. Accessed 15 Nov 2019.
221. Blanco L, Bernad A, Lázaro JM, Martín G, Garmendia C, Salas M. Highly efficient DNA synthesis by the phage phi 29 DNA polymerase: symmetrical mode of DNA replication. J Biol Chem. 1989;264(15): 8935–40.
222. Dean FB, Hosono S, Fang L, Wu X, Faruqi AF, Bray-Ward P, Sun Z, Zong Q, Du Y, Du J, Driscoll M, Song W, Kingsmore SF, Egholm M, Lasken RS. Comprehensive human genome amplification using multiple displacement amplification. Proc Natl Acad Sci U S A. 2002;99(8):5261–6.
223. Spits C, Le Caignec C, De Rycke M, Van Haute L, Van Steirteghem A, Liebaers I, Sermon K. Whole-genome multiple displacement amplification from single cells. Nat Protoc. 2006;1(4):1965–70.
224. Paez JG, Lin M, Beroukhim R, Lee JC, Zhao X, Richter DJ, Gabriel S, Herman P, Sasaki H, Altshuler D, Li C, Meyerson M, Sellers WR. Genome coverage and sequence fidelity of phi29 polymerase-based multiple strand displacement whole genome amplification. Nucleic Acids Res. 2004;32(9):71.
225. Spits C, Le Caignec C, De Rycke M, Van Haute L, Van Steirteghem A, Liebaers I, Sermon K. Optimization and evaluation of single-cell whole-genome multiple displacement amplification. Hum Mutat. 2006;27(5):496–503.
226. Bäumer C, Fisch E, Wedler H, Reinecke F, Korfhage C. Exploring DNA quality of single cells for genome analysis with simultaneous whole-genome amplification. Sci Rep. 2018;8(1):1–10. https://doi.org/10.1038/s41598-018-25895-7. Accessed 24 Oct 2019.
227. Picher ÁJ, Budeus B, Wafzig O, Krüger C, García-Gómez S, Martínez-Jiménez MI, Díaz-Talavera A, Weber D, Blanco L, Schneider A. TruePrime is a novel method for whole-genome amplification from single cells based on *Tth*PrimPol. Nat Commun. 2016;7:13296. https://doi.org/10.1038/ncomms13296. Accessed 07 Mar 2019.
228. Zafar H, Wang Y, Nakhleh L, Navin N, Chen K. Monovar: single-nucleotide variant detection in single cells. Nat Methods. 2016;13(6): 505–7. https://doi.org/10.1038/nmeth.3835. Accessed 28 Mar 2019.
229. Dong X, Zhang L, Milholland B, Lee M, Maslov AY, Wang T, Vijg J. Accurate identification of single-nucleotide variants in whole-genome-amplified single cells. Nat Methods. 2017;14(5):491–3. https://doi.org/10.1038/nmeth.4227. Accessed 28 Mar 2019.
230. Luquette LJ, Bohrson CL, Sherman MA, Park PJ. Identification of somatic mutations in single cell DNA-seq using a spatial model of allelic imbalance. Nat Commun. 2019;10(1):1–14. https://doi.org/10.1038/s41467-019-11857-8. Accessed 02 Sept 2019.
231. Roth A, McPherson A, Laks E, Biele J, Yap D, Wan A, Smith MA, Nielsen CB, McAlpine JN, Aparicio S, Bouchard-Côté A, Shah SP. Clonal genotype and population structure inference from single-cell tumor sequencing. Nat Methods. 2016;13(7):573–6.
232. Zafar H, Navin N, Chen K, Nakhleh L. SiCloneFit: Bayesian inference of population structure, genotype, and phylogeny of tumor clones from single-cell genome sequencing data: Cold Spring Harbor Laboratory; 2018. https://doi.org/10.1101/394262.
233. Singer J, Kuipers J, Jahn K, Beerenwinkel N. Single-cell mutation identification via phylogenetic inference. Nat Commun. 2018;9(1):5144. https://doi.org/10.1038/s41467-018-07627-7. Accessed 28 Mar 2019.
234. Poirion O, Zhu X, Ching T, Garmire LX. Using single nucleotide variations in single-cell RNA-seq to identify subpopulations and genotype-phenotype linkage. Nat Commun. 2018;9(1):4892. https://doi.org/10.1038/s41467-018-07170-5. Accessed 28 Mar 2019.
235. Bakker B, Taudt A, Belderbos ME, Porubsky D, Spierings DCJ, de Jong TV, Halsema N, Kazemier HG, Hoekstra-Wakker K, Bradley A, de Bont ESJM, van den Berg A, Guryev V, Lansdorp PM, Colomé-Tatché M, Foijer F. Single-cell sequencing reveals karyotype heterogeneity in murine and human malignancies. Genome Biol. 2016;17:115. https://doi.org/10.1186/s13059-016-0971-7. Accessed 14 Feb 2017.
236. Garvin T, Aboukhalil R, Kendall J, Baslan T, Atwal GS, Hicks J, Wigler M, Schatz MC. Interactive analysis and assessment of single-cell copy-number variations. Nat Methods. 2015;12(11):1058–60.
237. Fan J, Lee H-O, Lee S, Ryu D-E, Lee S, Xue C, Kim SJ, Kim K, Barkas N, Park PJ, Park W-Y, Kharchenko PV. Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data. Genome Res. 2018;28(8):1217–27.

238. Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, Cahill DP, Nahed BV, Curry WT, Martuza RL, Louis DN, Rozenblatt-Rosen O, Suvà ML, Regev A, Bernstein BE. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science. 2014;344(6190):1396–401.
239. Köster J, Dijkstra L, Marschall T, Schönhuth A. Enhancing sensitivity and controlling false discovery rate in somatic indel discovery. bioRxiv. 2019741256. https://doi.org/10.1101/741256. Accessed 22 Aug 2019.
240. Koptagel H, Jun S-H, Lagergren J. SCuPhr: a probabilistic framework for cell lineage tree reconstruction. bioRxiv. 2018357442. https://doi.org/10.1101/357442. Accessed 09 Aug 2018.
241. Satas G, Raphael BJ. Haplotype phasing in single-cell DNA-sequencing data. Bioinformatics. 2018;34(13):211–7.
242. Bohrson CL, Barton AR, Lodato MA, Rodin RE, Luquette LJ, Viswanadham VV, Gulhan DC, Cortés-Ciriano I, Sherman MA, Kwon M, Coulter ME, Galor A, Walsh CA, Park PJ. Linked-read analysis identifies mutations in single-cell DNA-sequencing data. Nature Genetics. 20191. https://doi.org/10.1038/s41588-019-0366-2. Accessed 28 Mar 2019.
243. Hård J, Al Hakim E, Kindblom M, Björklund SK, Sennblad B, Demirci I, Paterlini M, Reu P, Borgström E, Ståhl PL, Michaelsson J, Mold JE, Frisén J. Conbase: a software for unsupervised discovery of clonal somatic mutations in single cells through read phasing. Genome Biol. 2019;20(1):68. https://doi.org/10.1186/s13059-019-1673-8. Accessed 27 July 2019.
244. Davis A, Gao R, Navin N. Tumor evolution: Linear, branching, neutral or punctuated? Biochim Biophys Acta. 2017;1867(2):151–61.
245. Altrock PM, Liu LL, Michor F. The mathematics of cancer: integrating quantitative models. Nat Rev Cancer. 2015;15(12):730–45. https://doi.org/10.1038/nrc4029. Accessed 07 Mar 2019.
246. Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol. 1981;17(6):368–76.
247. Foo J, Leder K, Michor F. Stochastic dynamics of cancer initiation. Phys Biol. 2011;8(1):015002.
248. Beerenwinkel N, Antal T, Dingli D, Traulsen A, Kinzler KW, Velculescu VE, Vogelstein B, Nowak MA. Genetic progression and the waiting time to cancer. PLoS Comput Biol. 2007;3(11):225.
249. Haeno H, Gonen M, Davis MB, Herman JM, Iacobuzio-Donahue CA, Michor F. Computational modeling of pancreatic cancer reveals kinetics of metastasis suggesting optimum treatment strategies. Cell. 2012;148(1-2):362–75.
250. Kimmel M, Axelrod D. Branching Processes in Biology, 2nd ed. Interdisciplinary Applied Mathematics. New York: Springer; 2015. https://www.springer.com/gp/book/9781493915583. Accessed 28 Mar 2019.
251. Bozic I, Gerold JM, Nowak MA. Quantifying clonal and subclonal passenger mutations in cancer evolution. PLoS Comput Biol. 2016;12(2):1004731.
252. Bozic I, Antal T, Ohtsuki H, Carter H, Kim D, Chen S, Karchin R, Kinzler KW, Vogelstein B, Nowak MA. Accumulation of driver and passenger mutations during tumor progression. Proc Natl Acad Sci U S A. 2010;107(43):18545–50.
253. Bauer B, Siebert R, Traulsen A. Cancer initiation with epistatic interactions between driver and passenger mutations. J Theor Biol. 2014;358:52–60.
254. Acar A, Nichol D, Fernandez-Mateos J, Cresswell GD, Barozzi I, Hong SP, Spiteri I, Stubbs M, Burke R, Stewart A, Vlachogiannis G, Maley CC, Magnani L, Valeri N, Banerji U, Sottoriva A. Exploiting evolutionary herding to control drug resistance in cancer. bioRxiv. 2019566950. https://doi.org/10.1101/566950. Accessed 02 Apr 2019.
255. Zhang J, Cunningham JJ, Brown JS, Gatenby RA. Integrating evolutionary dynamics into treatment of metastatic castrate-resistant prostate cancer. Nat Commun. 2017;8(1):1816. https://doi.org/10.1038/s41467-017-01968-5. Accessed 02 Apr 2019.
256. Graybeal A. Is it better to add taxa or characters to a difficult phylogenetic problem? Syst Biol. 1998;47(1):9–17.
257. Pollock DD, Zwickl DJ, McGuire JA, Hillis DM. Increased taxon sampling is advantageous for phylogenetic inference. Syst Biol. 2002;51(4):664–71.
258. Roch S. A short proof that phylogenetic tree reconstruction by maximum likelihood is hard. IEEE/ACM Trans Comput Biol Bioinform. 2006;3(1):92–4.
259. Aberer AJ, Kobert K, Stamatakis A. ExaBayes: massively parallel Bayesian tree inference for the whole-genome era. Mol Biol Evol. 2014;31(10):2553–6. https://doi.org/10.1093/molbev/msu236. Accessed 08 Mar 2019.
260. Ayres DL. Research and application of parallel computing algorithms for statistical phylogenetic inference. PhD thesis, University of Maryland. 2017. https://doi.org/10.13016/M2FQ9Q584. http://drum.lib.umd.edu/handle/1903/19951. Accessed 08 Mar 2019.
261. Ogilvie HA, Bouckaert RR, Drummond AJ. StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. Mol Biol Evol. 2017;34(8):2101–14.
262. Leaché A. D, Banbury BL, Felsenstein J, de Oca AN-M, Stamatakis A. Short tree, long tree, right tree, wrong tree: new acquisition bias corrections for inferring SNP phylogenies. Syst Biol. 2015;64(6):1032–47.
263. Ross EM, Markowetz F. OncoNEM: inferring tumor evolution from single-cell sequencing data. Genome Biol. 2016;17:69.
264. Jahn K, Kuipers J, Beerenwinkel N. Tree inference for single-cell data. Genome Biol. 2016;17:86.
265. Zafar H, Tzen A, Navin N, Chen K, Nakhleh L. SiFit: inferring tumor trees from single-cell sequencing data under finite-sites models. Genome Biol. 2017;18(1):178.
266. El-Kebir M. SPhyR: tumor phylogeny estimation from single-cell sequencing data under loss and error. Bioinformatics. 2018;34(17):671–9. https://doi.org/10.1093/bioinformatics/bty589. Accessed 27 July 2019.
267. Ciccolella S, Gomez MS, Patterson M, Vedova GD, Hajirasouliha I, Bonizzoni P. Inferring cancer progression from single-cell sequencing while allowing mutation losses. bioRxiv. 2018268243. https://doi.org/10.1101/268243. Accessed 07 Mar 2019.
268. Kozlov O. Models, optimizations, and tools for large-scale phylogenetic inference, handling sequence uncertainty, and taxonomic validation. PhD thesis, Karlsruhe Institute of Technology (KIT). 2018.
269. Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics. 2019. https://doi.org/10.1093/bioinformatics/btz305. Accessed 27 July 2019.
270. Rivas E, Eddy SR. Probabilistic phylogenetic inference with insertions and deletions. PLoS Comput Biol. 2008;4(9):1000172.
271. Holmes IH. Solving the master equation for indels. BMC Bioinformatics. 2017;18(1):255.
272. Kim T-M, Xi R, Luquette LJ, Park RW, Johnson MD, Park PJ. Functional genomic analysis of chromosomal aberrations in a compendium of 8000 cancer genomes. Genome Res. 2013;23(2):217–27. https://doi.org/10.1101/gr.140301.112.
273. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, Cook K, Stepansky A, Levy D, Esposito D, Muthuswamy L, Krasnitz A, McCombie WR, Hicks J, Wigler M. Tumour evolution inferred by single-cell sequencing. Nature. 2011;472(7341):90–4.
274. Eirew P, Steif A, Khattra J, Ha G, Yap D, Farahani H, Gelmon K, Chia S, Mar C, Wan A, Laks E, Biele J, Shumansky K, Rosner J, McPherson A, Nielsen C, Roth AJL, Lefebvre C, Bashashati A, de Souza C, Siu C, Aniba R, Brimhall J, Oloumi A, Osako T, Bruna A, Sandoval JL, Algara T, Greenwood W, Leung K, Cheng H, Xue H, Wang Y, Lin D, Mungall AJ, Moore R, Zhao Y, Lorette J, Nguyen L, Huntsman D, Eaves CJ, Hansen C, Marra MA, Caldas C, Shah SP, Aparicio S. Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. Nature. 2015;518(7539):422–6. https://doi.org/10.1038/nature13952. Accessed 03 July 2015.
275. Zaccaria S, El-Kebir M, Klau GW, Raphael BJ. The copy-number tree mixture deconvolution problem and applications to multi-sample bulk sequencing tumor data. In: Sahinalp SC, editor. Research in Computational Molecular Biology. Lecture Notes in Computer Science. Heidelberg: Springer; 2017. p. 318–35.
276. Gawad C, Koh W, Quake SR. Single-cell genome sequencing: current state of the science. Nat Rev Genet. 2016;17(3):175–88.
277. Bignell GR, Santarius T, Pole JCM, Butler AP, Perry J, Pleasance E, Greenman C, Menzies A, Taylor S, Edkins S, Campbell P, Quail M, Plumb B, Matthews L, McLay K, Edwards PAW, Rogers J, Wooster R, Futreal PA, Stratton MR. Architectures of somatic genomic rearrangement in human cancer amplicons at sequence-level resolution. Genome Res. 2007;17(9):1296–303. https://doi.org/10.1101/gr.6522707.
278. Santaguida S, Richardson A, Iyer DR, M'Saad O, Zasadil L, Knouse KA, Wong YL, Rhind N, Desai A, Amon A. Chromosome mis-segregation generates cell-cycle-arrested cells with complex karyotypes that are eliminated by the immune system. Dev Cell. 2017;41(6):638–6515. https://doi.org/10.1016/j.devcel.2017.05.022. Accessed 27 Mar 2019.

279. Schwarz RF, Trinh A, Sipos B, Brenton JD, Goldman N, Markowetz F. Phylogenetic quantification of intra-tumour heterogeneity. PLoS Comput Biol. 2014;10(4):1003535. https://doi.org/10.1371/journal.pcbi.1003535. Accessed 27 Mar 2019.
280. Zeira R, Shamir R. Genome rearrangement problems with single and multiple gene copies : a review. 2018. https://pdfs.semanticscholar.org/85e6/7eb03d1b3d004c60a12df08c1f937fbaa974.pdf. Not clear where this was initially published and whether it is peer-reviewed.
281. Kuipers J, Jahn K, Raphael BJ, Beerenwinkel N. Single-cell sequencing data reveal widespread recurrence and loss of mutational hits in the life histories of tumors. Genome Res. 2017;27(11):1885–94.
282. Yang L, Lin PC. Mechanisms that drive inflammatory tumor microenvironment, tumor heterogeneity, and metastatic progression. Semin Cancer Biol. 2017;47:185–95.
283. Junttila MR, de Sauvage FJ. Influence of tumour micro-environment heterogeneity on therapeutic response. Nature. 2013;501(7467):346–54.
284. Corredor G, Wang X, Zhou Y, Lu C, Fu P, Syrigos K, Rimm DL, Yang M, Romero E, Schalper KA, Velcheti V, Madabhushi A. Spatial Architecture and Arrangement of Tumor-Infiltrating Lymphocytes for Predicting Likelihood of Recurrence in Early-Stage NonÛSmall Cell Lung Cancer. Clin Cancer Res. 2018;25:1526–1534. https://doi.org/10.1158/1078-0432.CCR-18-2013.
285. Turajlic S, Swanton C. Metastasis as an evolutionary process. Science. 2016;352(6282):169–75.
286. Michor F, Iwasa Y, Nowak MA. Dynamics of cancer progression. Nat Rev Cancer. 2004;4(3):197–205.
287. Williams MJ, Werner B, Barnes CP, Graham TA, Sottoriva A. Identification of neutral tumor evolution across cancer types. Nat Genet. 2016;48(3):238–44. https://doi.org/10.1038/ng.3489.
288. Larue RTHM, Defraene G, De Ruysscher D, Lambin P, van Elmpt W. Quantitative radiomics studies for tissue characterization: a review of technology and methodological procedures. Br J Radiol. 2017;90(1070):20160665.
289. Ghaznavi F, Evans A, Madabhushi A, Feldman M. Digital imaging in pathology: whole-slide imaging and beyond. Annu Rev Pathol. 2013;8:331–59.
290. Saco A, Ramírez J, Rakislova N, Mira A, Ordi J. Validation of whole-slide imaging for histolopathogical diagnosis: Current state. Pathobiology. 2016;83(2-3):89–98.
291. Datta S, Malhotra L, Dickerson R, Chaffee S, Sen CK, Roy S. Laser capture microdissection: big data from small samples. Histol Histopathol. 2015;30(11):1255–69.
292. Casasent AK, Schalck A, Gao R, Sei E, Long A, Pangburn W, Casasent T, Meric-Bernstam F, Edgerton ME, Navin NE. Multiclonal invasion in breast tumors identified by topographic single cell sequencing. Cell. 2018;172(1-2):205–21712.
293. Gurcan MN, Boucheron L, Can A, Madabhushi A, Rajpoot N, Yener B. Histopathological image analysis: a review. IEEE Rev Biomed Eng. 2009;2:147.
294. Irshad H, Veillard A, Roux L, Racoceanu D. Methods for nuclei detection, segmentation, and classification in digital histopathology: a review—current status and future potential. IEEE Rev Biomed Eng. 2014;7:97–114. https://doi.org/10.1109/RBME.2013.2295804.
295. Komura D, Ishikawa S. Machine learning methods for histopathological image analysis. Comput Struct Biotechnol J. 2018;16:34–42.
296. Yuan Y. Spatial heterogeneity in the tumor microenvironment. Cold Spring Harb Perspect Med. 2016;6(8):a026583. https://doi.org/10.1101/cshperspect.a026583.
297. Heindl A, Nawaz S, Yuan Y. Mapping spatial heterogeneity in the tumor microenvironment: a new era for digital pathology. Lab Invest. 2015;95(4):377–84.
298. Rączkowski u, Możejko M, Zambonelli J, Szczurek E. ARA: accurate, reliable and active histopathological image classification framework with Bayesian deep learning. Sci Rep. 2019;9(1):1–12. https://doi.org/10.1038/s41598-019-50587-1. Accessed 13 Nov 2019.
299. Martens EA, Kostadinov R, Maley CC, Hallatschek O. Spatial structure increases the waiting time for cancer. New J Phys. 2011;13(11):115014. https://doi.org/10.1088/1367-2630/13/11/115014.
300. Waclaw B, Bozic I, Pittman ME, Hruban RH, Vogelstein B, Nowak MA. A spatial model predicts that dispersal and cell turnover limit intratumour heterogeneity. Nature. 2015;525(7568):261–4.
301. Schapiro D, Jackson HW, Raghuraman S, Fischer JR, Zanotelli VRT, Schulz D, Giesen C, Catena R, Varga Z, Bodenmiller B. histoCAT: analysis of cell phenotypes and interactions in multiplex image cytometry data. Nat Methods. 2017;14(9):873–6.
302. Arnol D, Schapiro D, Bodenmiller B, Saez-Rodriguez J, Stegle O. Modelling cell-cell interactions from spatial molecular data with spatial variance component analysis. bioRxiv. 2018265256. https://doi.org/10.1101/265256. Accessed 15 Nov 2019.
303. Battich N, Stoeger T, Pelkmans L. Control of transcript variability in single mammalian cells. Cell. 2015;163(7):1596–610.
304. Johnson BE, Mazor T, Hong C, Barnes M, Aihara K, McLean CY, Fouse SD, Yamamoto S, Ueda H, Tatsuno K, Asthana S, Jalbert LE, Nelson SJ, Bollen AW, Gustafson WC, Charron E, Weiss WA, Smirnov IV, Song JS, Olshen AB, Cha S, Zhao Y, Moore RA, Mungall AJ, Jones SJM, Hirst M, Marra MA, Saito N, Aburatani H, Mukasa A, Berger MS, Chang SM, Taylor BS, Costello JF. Mutational analysis reveals the origin and therapy-driven evolution of recurrent glioma. Science. 2014;343(6167):189–93.
305. Szczurek E, Misra N, Vingron M. Synthetic sickness or lethality points at candidate combination therapy targets in glioblastoma. Int J Cancer. 2013;133(9):2123–32.
306. Jerby-Arnon L, Pfetzer N, Waldman YY, McGarry L, James D, Shanks E, Seashore-Ludlow B, Weinstock A, Geiger T, Clemons PA, Gottlieb E, Ruppin E. Predicting cancer-specific vulnerability via data-driven detection of synthetic lethality. Cell. 2014;158(5):1199–209.
307. Matlak D, Szczurek E. Epistasis in genomic and survival data of cancer patients. PLoS Comput Biol. 2017;13(7):1005626.
308. Wilkins JF, Cannataro VL, Shuch B, Townsend JP. Analysis of mutation, selection, and epistasis: an informed approach to cancer clinical trials. Oncotarget. 2018;9(32):22243–53. https://doi.org/10.18632/oncotarget.25155. Accessed 13 Nov 2019.
309. Rannala B, Yang Z. Efficient bayesian species tree inference under the multispecies coalescent. Syst Biol. 2017;66(5):823–42.
310. Liu L, Xi Z, Wu S, Davis CC, Edwards SV. Estimating phylogenetic trees from genome-scale data. Ann N Y Acad Sci. 2015;1360:36–53.
311. Skums P, Tsyvina V, Zelikovsky A. Inference of clonal selection in cancer populations using single-cell sequencing data. bioRxiv. 2019465211. https://doi.org/10.1101/465211. Accessed 28 June 2019.
312. Zhang J, Nielsen R, Yang Z. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol. 2005;22(12):2472–9.
313. Smith MD, Wertheim JO, Weaver S, Murrell B, Scheffler K, Kosakovsky Pond SL. Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. Mol Biol Evol. 2015;32(5):1342–53.
314. Kolaczkowski B, Thornton JW. A mixed branch length model of heterotachy improves phylogenetic accuracy. Mol Biol Evol. 2008;25(6):1054–66.
315. Yang Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol. 1994;39(3):306–14.
316. Le SQ, Dang CC, Gascuel O. Modeling protein evolution with several amino acid replacement matrices depending on site rates. Mol Biol Evol. 2012;29(10):2921–36.
317. Kryazhimskiy S, Plotkin JB. The population genetics of dN/dS. PLoS Genet. 2008;4(12):1000304.
318. Neher RA, Russell CA, Shraiman BI. Predicting evolution from the shape of genealogical trees. Elife. 2014;3:.
319. Gray RR, Pybus OG, Salemi M. Measuring the temporal structure in serially-sampled phylogenies. Methods Ecol Evol. 2011;2(5):437–45.
320. Steinbrück L, McHardy AC. Allele dynamics plots for the study of evolutionary dynamics in viral populations. Nucleic Acids Res. 2011;39(1):4.
321. Fletcher W, Yang Z. The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. Mol Biol Evol. 2010;27(10):2257–67.
322. Redelings B. Erasing errors due to alignment ambiguity when estimating positive selection. Mol Biol Evol. 2014;31(8):1979–93.
323. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol. 2015;33(5):495–502.

324. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, Stoeckius M, Smibert P, Satija R. Comprehensive integration of single cell data. bioRxiv. 2018460147. https://doi.org/10.1101/460147. Accessed 14 June 2019.
325. Zhang H, Lee CAA, Li Z, Garbe JR, Eide CR, Petegrosso R, Kuang R, Tolar J. A multitask clustering approach for single-cell RNA-seq analysis in recessive dystrophic epidermolysis bullosa. PLoS Comput Biol. 2018;14(4):1006053.
326. Barkas N, Petukhov V, Nikolaeva D, Lozinsky Y, Demharter S, Khodosevich K, Kharchenko PV. Wiring together large single-cell RNA-seq sample collections. bioRxiv. 2018460246. https://doi.org/10.1101/460246. Accessed 11 Apr 2019.
327. Gao X, Hu D, Gogol M, Li H. ClusterMap: comparing analyses across multiple single cell RNA-Seq profiles. bioRxiv. 2018331330. https://doi.org/10.1101/331330. Accessed 04 Apr 2019.
328. Park J-E, Polanski K, Meyer K, Teichmann SA. Fast batch alignment of single cell transcriptomes unifies multiple mouse cell atlases into an integrated landscape. bioRxiv. 2018397042. https://doi.org/10.1101/397042. Accessed 04 Apr 2019.
329. Wagner F, Yanai I. Moana: a robust and scalable cell type classification framework for single-cell RNA-Seq data. bioRxiv. 2018456129. https://doi.org/10.1101/456129. Accessed 04 Apr 2019.
330. Boufea K, Seth S, Batada NN. scID: identification of equivalent transcriptional cell populations across single cell RNA-seq data using discriminant analysis. bioRxiv. 2019470203. https://doi.org/10.1101/470203. Accessed 04 Apr 2019.
331. Johansen N, Quon G. scAlign: a tool for alignment, integration and rare cell identification from scRNA-seq data. bioRxiv. 2019504944. https://doi.org/10.1101/504944. Accessed 04 Apr 2019.
332. Johnson TS, Wang T, Huang Z, Yu CY, Wu Y, Han Y, Zhang Y, Huang K, Zhang J. LAmbDA:label ambiguous domain adaptation dataset integration reduces batch effects and improves subtype detection. Bioinformatics. 2019. https://doi.org/10.1093/bioinformatics/btz295. Accessed 03 May 2019.
333. Macaulay IC, Ponting CP, Voet T. Single-cell multiomics: multiple measurements from single cells. Trends Genet. 2017;33(2):155–68. https://doi.org/10.1016/j.tig.2016.12.003. Accessed 27 Mar 2019.
334. Dey SS, Kester L, Spanjaard B, Bienko M, van Oudenaarden A. Integrated genome and transcriptome sequencing of the same cell. Nat Biotechnol. 2015;33(3):285–9. https://doi.org/10.1038/nbt.3129. Accessed 27 Mar 2019.
335. Macaulay IC, Teng MJ, Haerty W, Kumar P, Ponting CP, Voet T. Separation and parallel sequencing of the genomes and transcriptomes of single cells using G&T-seq. Nat Protoc. 2016;11(11):2081–103. https://doi.org/10.1038/nprot.2016.138. Accessed 27 Mar 2019.
336. Angermueller C, Clark SJ, Lee HJ, Macaulay IC, Teng MJ, Hu TX, Krueger F, Smallwood S, Ponting CP, Voet T, Kelsey G, Stegle O, Reik W. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. Nat Methods. 2016;13(3):229–32. https://doi.org/10.1038/nmeth.3728.
337. Cheow LF, Courtois ET, Tan Y, Viswanathan R, Xing Q, Tan RZ, Tan DSW, Robson P, Loh Y-H, Quake SR, Burkholder WF. Single-cell multimodal profiling reveals cellular epigenetic heterogeneity. Nat Methods. 2016;13(10):833–6. https://doi.org/10.1038/nmeth.3961. Accessed 10 Apr 2019.
338. Singh A, Gautier B, Shannon CP, Rohart F, Vacher M, Tebutt SJ, Cao K-AL. DIABLO: from multi-omics assays to biomarker discovery, an integrative approach. bioRxiv. 2018067611. https://doi.org/10.1101/067611. Accessed 10 Apr 2019.
339. Rohart F, Gautier B, Singh A, Cao K-AL. mixOmics: An R package for 'omics feature selection and multiple data integration. PLoS Comput Biol. 2017;13(11):1005752. https://doi.org/10.1371/journal.pcbi.1005752. Accessed 27 Mar 2019.
340. Rohart F, Eslami A, Matigian N, Bougeard S, Lê Cao K-A. MINT: a multivariate integrative method to identify reproducible molecular signatures across independent experiments and platforms. BMC Bioinformatics. 2017;18(1):128. https://doi.org/10.1186/s12859-017-1553-8. Accessed 10 Apr 2019.
341. Zahn H, Steif A, Laks E, Eirew P, VanInsberghe M, Shah SP, Aparicio S, Hansen CL. Scalable whole-genome single-cell library preparation without preamplification. Nat Methods. 2017;14(2):167–73. https://doi.org/10.1038/nmeth.4140. Accessed 27 Mar 2019.
342. McCarthy DJ, Rostom R, Huang Y, Kunz DJ, Danecek P, Bonder MJ, Hagai T, Consortium H, Wang W, Gaffney DJ, Simons BD, Stegle O, Teichmann SA. Cardelino: integrating whole exomes and single-cell transcriptomes to reveal phenotypic impact of somatic variants. bioRxiv. 2018413047. https://doi.org/10.1101/413047. Accessed 27 Mar 2019.
343. Loper J, Bakken T, Sumbul U, Murphy G, Zeng H, Blei D, Paninski L. The Markov link method: a nonparametric approach to combine observations from multiple experiments. bioRxiv. 2019457283. https://doi.org/10.1101/457283. Accessed 27 Mar 2019.
344. Mangul S, Martin LS, Hill BL, Lam AK-M, Distler MG, Zelikovsky A, Eskin E, Flint J. Systematic benchmarking of omics computational tools. Nat Commun. 2019;10(1):1393. https://doi.org/10.1038/s41467-019-09406-4. Accessed 02 Apr 2019.
345. Weber LM, Saelens W, Cannoodt R, Soneson C, Hapfelmeier A, Gardner PP, Boulesteix A-L, Saeys Y, Robinson MD. Essential guidelines for computational method benchmarking. Genome Biol. 2019;20(1):125. https://doi.org/10.1186/s13059-019-1738-8. Accessed 28 June 2019.
346. Grün D, Kester L, van Oudenaarden A. Validation of noise models for single-cell transcriptomics. Nat Methods. 2014;11(6):637–40. https://doi.org/10.1038/nmeth.2930. Accessed 09 July 2019.
347. Tian L, Dong X, Freytag S, Cao K-AL, Su S, JalalAbadi A, Amann-Zalcenstein D, Weber TS, Seidi A, Jabbari JS, Naik SH, Ritchie ME. Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. Nat Methods. 2019;16(6):479. https://doi.org/10.1038/s41592-019-0425-8. Accessed 09 July 2019.
348. Vieth B, Parekh S, Ziegenhain C, Enard W, Hellmann I. A systematic evaluation of single cell RNA-seq analysis pipelines. Nat Commun. 2019;10(1):1–11. https://doi.org/10.1038/s41467-019-12266-7. Accessed 23 Oct 2019.
349. Vallejos CA, Marioni JC, Richardson S. BASiCS: Bayesian analysis of single-cell sequencing data. PLoS Comput Biol. 2015;11(6):1004333.
350. Korthauer KD, Chu L-F, Newton MA, Li Y, Thomson J, Stewart R, Kendziorski C. A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. Genome Biol. 2016;17(1):222.
351. Lun ATL, Bach K, Marioni JC. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biol. 2016;17:75.
352. Lun ATL, Marioni JC. Overcoming confounding plate effects in differential expression analyses of single-cell RNA-seq data. Biostatistics. 2017;18(3):451–64.
353. Rizzetto S, Eltahla AA, Lin P, Bull R, Lloyd AR, Ho JWK, Venturi V, Luciani F. Impact of sequencing depth and read length on single cell RNA sequencing data of T cells. Sci Rep. 2017;7(1):12781.
354. Zappia L, Phipson B, Oshlack A. Splatter: simulation of single-cell RNA sequencing data. Genome Biol. 2017;18(1):174.
355. Vieth B, Ziegenhain C, Parekh S, Enard W, Hellmann I. powsimr: power analysis for bulk and single cell RNA-seq experiments. Bioinformatics. 2017;33(21):3486–8.
356. Zhang X, Xu C, Yosef N. SymSim: simulating multi-faceted variability in single cell RNA sequencing. bioRxiv. 2019378646. https://doi.org/10.1101/378646. Accessed 28 June 2019.
357. Semeraro R, Orlandini V, Magi A. Xome-Blender: a novel cancer genome simulator. PLoS ONE. 2018;13(4):0194472.
358. Xia LC, Ai D, Lee H, Andor N, Li C, Zhang NR, Ji HP. SVEngine: an efficient and versatile simulator of genome structural variations with features of cancer clonal evolution. Gigascience. 2018;7(7):. https://doi.org/10.1093/gigascience/giy081.
359. Meng J, Chen Y-PP. A database of simulated tumor genomes towards accurate detection of somatic small variants in cancer. PLoS ONE. 2018;13(8):0202982.
360. Soneson C, Robinson MD. Towards unified quality verification of synthetic count data with countsimQC. Bioinformatics. 2017;34(4):691–2.
361. Knyazev S, Tsyvina V, Melnyk A, Artyomenko A, Malygina T, Porozov YB, Campbell E, Switzer WM, Skums P, Zelikovsky A. CliqueSNV: scalable reconstruction of intra-host viral populations from NGS reads. 2018. https://doi.org/10.1101/264242.

Publisher's Note