Abstract: Single-cell and single-nuclei sequencing experiments reveal previously unseen molecular details. The number of sequencing procedures and computational data analysis approaches have been increasing rapidly in recent years. This chapter provides an overview of the current developments in single-cell analysis. An introduction and practical guidance for choosing the most suitable sequencing procedure to match individual experimental demands in the course of investigating biological hypotheses are presented. Basic data analysis approaches are highlighted, followed by a discussion on advanced downstream approaches to enrich the information obtained from single-cell experiments; for example, trajectory analysis, pseudotime assumptions, and network inference. Currently unsolved challenges are discussed to allow the reader to avoid the most common pitfalls.

Keywords: network integration; pseudotemporal reconstruction; RNA velocity; single-cell RNA-sequencing; scRNA-seq data
INTRODUCTION

A typical single-cell sequencing workflow involves initial tissue and cell preparation, cell capturing and library preparation, sequencing and raw data processing, as well as visualization and downstream analyses. A plethora of protocols are available for the preparation of single-cell suspensions because the optimal procedure differs for each tissue and cell type to be isolated (1–3). Hence, this chapter focuses on the steps of data generation and data analysis, with emphasis on various capture and sequencing techniques, which are foundations for the subsequent computational data analyses.

THE CAPTURE TECHNIQUE SIGNIFICANTLY DETERMINES THE QUANTITY OF MEASURABLE CELLS

Several protocols for single cell RNA-sequencing (scRNA-seq) have been published over the last few years, and it remains a rapidly evolving field (4). The capture technique determines throughput, sorting options, and the type of additional information that can be obtained. The most widely used options are microwell- and droplet-based. These techniques differ in their strategies for tagging transcripts based on cell origin, and in the ways libraries are generated for sequencing. Table 1 summarizes the most common techniques to date and provides an overview of their main characteristics.

<table>
<thead>
<tr>
<th>Technique</th>
<th>Method</th>
<th>Detected cells</th>
<th>Sensitivity</th>
<th>Costs/cell</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>CEL-seq2</td>
<td>Microwell-based</td>
<td>&lt; 400</td>
<td>Very high</td>
<td>~ 3</td>
<td>25h</td>
</tr>
<tr>
<td>Hashimshony et al. (5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Drop-seq</td>
<td>Droplet-based</td>
<td>5,000 – 10,000</td>
<td>Moderate</td>
<td>&lt; 0.1€</td>
<td>10h</td>
</tr>
<tr>
<td>Macosko et al. (6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ICELL8</td>
<td>Microwell-based</td>
<td>1,000 – 1,800</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Goldstein et al. (7)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>InDrop-seq</td>
<td>Droplet-based</td>
<td>5,000 – 10,000</td>
<td>Moderate</td>
<td>&lt; 0.1€</td>
<td>10h</td>
</tr>
<tr>
<td>Klein et al. (8)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MARS-seq</td>
<td>Microwell-based</td>
<td>100 – 1,000</td>
<td>Low</td>
<td>~ 0.50€</td>
<td>NA</td>
</tr>
<tr>
<td>Jaitin et al. (9)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seq-well</td>
<td>Microwell-based</td>
<td>&gt; 10,000</td>
<td>Moderate</td>
<td>&lt; 0.1€</td>
<td>10h</td>
</tr>
<tr>
<td>Gierahn et al. (10)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SmartSeq2</td>
<td>Microwell-based</td>
<td>&lt; 400</td>
<td>Very high</td>
<td>~ 10€</td>
<td>25h</td>
</tr>
<tr>
<td>Picelli et al. (11)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10x Genomics Chromium</td>
<td>Droplet-based</td>
<td>1,000 – 10,000</td>
<td>High</td>
<td>~ 0.25€</td>
<td>9h</td>
</tr>
<tr>
<td>Zheng et al. (12)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The most common single cell sequencing techniques listed together with protocol-dependent key data, such as the range of detectable cell numbers, sensitivity in terms of gene detection rates, and economic factors (average costs, amount of time to complete the procedure). NA, not available.
Microwell-based techniques allow for visual inspection

For well-based platforms, cells are usually transferred into micro- or nano-well plates using pipette or laser capture methods, such as fluorescent activated cell sorting (FACS) based on surface markers. This option renders well-based platforms particularly useful when isolation of a specific subset of cells is required, for example to explore rare cell types. Another advantage is the ability to visually inspect captured cells, allowing for identification of wells containing damaged or no cells and/or providing additional morphological information. The main drawback of well-based platforms is that they are often low-throughput and require a considerable amount of hands-on work per cell in contrast to other methods. These drawbacks are overcome to some extent by utilization of microfluidic platforms, such as Fluidigm C1 (13), which can be integrated in the workflow of some microwell-based platforms, providing a higher throughput. However, only around 10% of cells are typically captured in a microfluidic platform, rendering it inappropriate for the detection of rare-cell types. The C1 system also allows for visual inspection under the microscope, thereby enabling the user to exclude empty wells and wells containing damaged cells or doublets prior to downstream library preparation. The high cost of the microfluidic cartridges can limit the sample size used in each project, but expenses can be reduced on reagents since reactions can be carried out in a smaller volume.

Droplet-based techniques allow for high throughput

Droplet-based methods use microfluidics to encapsulate each individual cell together with a bead inside a nanoliter droplet that includes specific enzymes required to construct the library. The bead carries primers with a unique barcode, which bind the cell’s mRNA and thus will be attached to all reads originating from that cell. All droplets can be pooled to produce a sequencing library. After sequencing, the reads can be assigned to the cell of origin based on the barcodes. Since the library preparation costs are comparably low, and the downstream processes are less elaborate due to the pooling step, droplet platforms typically have the highest throughput. Usually, the costs for the subsequent sequencing become the limiting factor, so that in typical experiments the coverage is rather low with only a few thousand different transcripts detected per cell. One major drawback is that these protocols offer little control over the cell input and thus are susceptible to selection bias, leading to inaccurate reflection of the biology of the studied system.

THE SEQUENCING TECHNIQUE DETERMINES THE OPTIONS FOR DATA ANALYSES

Once single-cell resolution is achieved via one of the approaches mentioned above, the individual transcriptomes must be sequenced. There are two main forms of sequencing techniques: full-length and tag-based protocols. Full-length based protocols aim to achieve a uniform read coverage of each transcript, whereas tag-based protocols only capture either the 5’- or 3’-end of each RNA molecule.
The choice of capture and quantification method has important implications on the types of analyses the data can be used for.

Tag-based protocols can be combined with unique molecular identifiers (UMIs), which permit multiplexing and improve quantification. However, the restriction to one end of the transcript may hamper the alignment and renders these protocols unsuitable for studies on allele-specific expression or isoform usage (14). To diminish these limitations, paired-end sequencing can be conducted, which involves sequencing both ends of cDNA fragments in a library and aligning the forward and reverse reads as read pairs. This procedure facilitates the detection of genomic rearrangements, such as insertions, deletions, and inversions, allowing for the discovery of gene fusions, novel transcripts, and novel splice isoforms. Moreover, tag-based methods have been established that are able to detect the co-occurrence of a specific transcription start site and a polyadenylation site (15). However, the generation of full-length cDNAs from very long transcripts still poses a technical limitation for any 5′-3′-sequencing method.

In contrast, full-length protocols provide an even coverage of transcripts and are suitable for the discovery of alternative-splicing events and allele-specific expression using single-nucleotide polymorphisms. A disadvantage of these protocols is that it is not possible to incorporate UMIs and barcodes for exact gene level quantification or multiplexing, leading to increased complexity of downstream processing.

For the sequencing step, the Illumina platform is widely used (e.g., HiSeq4000, NextSeq500, and NovaSeq™6000), being responsible for more than 90% of the world’s sequencing data (16). All Illumina platforms use a sequencing by synthesis approach, yielding reliable base calls for highly repetitive sequences.

**HOW TO CHOOSE THE RIGHT APPROACH?**

The choice of method depends primarily on the individual scientific question and is further influenced by the compromise between cell numbers, information depth, and overall cost. For example, a droplet-based method will be most suitable for the characterization of the composition of a tissue, since it allows for large numbers of cells to be captured. In contrast, for in-depth analysis of rare cell types, it is probably best to enrich them using FACS, if there is a known surface marker, and then sequencing a smaller number of cells. Full-length transcript quantification will be more appropriate for studying different isoforms, since tagged protocols are much more limited. By contrast, UMIs can facilitate gene-level quantification, but they can only be used with tagged protocols. The low cost and high throughput of tag-based approaches has led to their widespread application in studies of gene expression levels, cell-type discovery, and tissue composition. However, it is recommended to consider the different techniques critically before starting the experiment. Svensson *et al.* compared the accuracy and sensitivity of different protocols and reported substantial differences between them (17). Figure 1 is based on current benchmarking studies (18-20) and should serve as a rough orientation for decision-making between the different experimental techniques based on individual experimental demands.
In the following, state-of-the-art computational components of scRNA-seq data analysis are presented, and underlying methods are discussed to contribute an update to previous single-cell analysis reviews (21, 22). We highlight well-executed benchmarking studies for additional in-depth reading and seek to guide new users through the landscape of scRNA-seq analysis tools with regards to data processing, downstream, and network analyses.

An introduction to the broad variety of data analysis platforms

A large number of data analysis platforms (web-based, stand-alone workflows, or integrated into computational frameworks like R or Python) are available to analyze scRNA-seq data. A comprehensive list is continuously updated in the scRNA-Tools database (https://www.scrna-tools.org/ [accessed on 13 January 2021]). There are commercially available analysis software packages, some of which are developed by single-cell sequencing companies and service providers, such as Cell Ranger and Loupe Cell Browser (10X Genomics) (23), as well as SeqGeq (BD Biosystems) (24). Others are designed by companies specializing in software solutions, such as Partek Flow (Partek) (25). While commercially available packages are user-friendly, open-source analysis packages are usually more powerful, transparent, reproducible, and flexible. Bioconductor, an open-development software
project for the analysis of high-throughput genomics data, provides powerful analysis tools, such as Scanpy (26) or Scater (27). Currently, the R package “Seurat” (28) is one of the most popular toolboxes for general single cell sequencing analysis, consistently performing well in benchmarking studies. However, the need for computational basic skills in either R or Python poses a hurdle for many scientists. Galaxy (https://usegalaxy.eu [accessed on 13 January 2021]) serves the same purpose as Bioconductor, but the developers aimed for an enhanced accessibility of the tools and easy usability for scientists with little or no bioinformatics background. An extensive Galaxy online training module (https://galaxyproject.github.io/training-material/ [accessed on 13 January 2021]) is offered, and it is possible to use the most common workflows without having to master any command line-based tools (29).

Foundations of processing raw scRNA-seq data

Standard data processing can be classified into six stages: (i) raw data alignment; (ii) quality control and data normalization; (iii) data integration and correction; (iv) expression recovery; (v) feature-selection of data; and (vi) dimensionality reduction and visualization. The subsequently performed downstream analyses may use different levels of processed data as input (Figure 2). Depending on the experimental setup or data used, it is also possible to skip certain levels or to have slight alterations in their order, for example, data integration and correction might not be needed for single-batch datasets.

![Flow chart summarizing current scRNA-seq data analysis principles.](image-url)
Alignment of scRNA-Seq raw data

Alignment is the first and one of the most critical steps of the scRNA-seq analyses. In general, the aim of the alignment step is to find the original transcriptomic location of the experimentally obtained sequencing reads. Thus, the choice of the alignment tool and its parametrization directly affects the count matrix, all subsequent downstream analysis steps, and, finally, the biological findings. The two most popular tools for alignment are the splice-aware aligner STAR (30) and the pseudoalignment approach Kallisto (31). Recent benchmark studies evaluated the performance of these two methods using real datasets obtained from different platforms (DropSeq, Fluidigm, and 10xChromium) (32, 33). They conclude that Kallisto’s use of computing resources is much less demanding than STAR when only cDNA sequences are used as the reference; however, such efficiency gain is at the cost of loss of information.

Quality control to determine cell viability and sequencing outcome

An important aspect of scRNA-seq protocols is that captured cells, irrespective of the method used, are often stressed, damaged, or broken. In addition, some capture sites can be empty, and some may contain multiple cells. All these events refer to “low quality” cells, which may lead to misinterpretation of the data and, therefore, need to be corrected (34). In general, cell quality control (QC) is commonly performed based on three QC covariates: (i) the number of counts per barcode (count depth); (ii) the number of genes per barcode; and (iii), the fraction of counts from mitochondrial genes per barcode (35). Cells that show an aberrant behavior for these characteristics are typically removed from further analysis, although care must be taken when studying a heterogeneous population of cells as total mRNA content and other features can vary substantially. On the one hand, barcodes with only low count depths, few detected genes, and a high fraction of mitochondrial counts may indicate cells whose cytoplasmic mRNA leaked out through a broken membrane, and thus, only mRNA located in the mitochondria is still conserved (22). On the other hand, cells with very high counts and a large amount of transcripts may represent doublets, which is why these have to be filtered with specific tools like Scrublet (36) or Solo (37).

Generation and normalization of the count matrix to ensure comparability

Counts in a count matrix represent the successful quantification of a sequencing read to a specific genomic location. There are multiple tools that generate a count matrix, e.g., Cell Ranger (38), indrops (39), SEQC (40), or bustools (41). Due to the technical variability inherent in the count matrix generation, the count depths for the same cell can differ. Thus, differences in gene expression between cells based on count data may have been introduced during sampling (42). Normalization addresses these differences to obtain correct relative gene expression abundances between cells, for example, via scaling of count data. The most common normalization approach is count depth scaling, referred to as “counts per million” (CPM). This approach was adapted from bulk RNA-seq analysis to
normalize count data towards a size factor proportional to the count depth per cell. Weinreb et al. introduced an extension of CPM that excludes genes accounting for less than 5% of the total counts in any cell, which allows for molecular count variability in only a few highly expressed genes (43). More cellular heterogeneity is taken into consideration by a pooling-based size factor estimation method that can be applied for more heterogeneous samples in order to increase the validity of the biological conclusions (44). To acquire a well-suited normalization, one may use Scone, which is a tool that provides graphical summaries and quantitative reports, as well as trade-offs and ranks of normalization methods by panel performance (45).

**Batch-effect correction and external data integration may enhance the biological outcome**

Single-cell data is often acquired based on multiple experiments with varying capturing times, consumables, and technology platforms. These differences can lead to large variations or so-called “batch-effects” in the data and may confound biological variations of interest. Tran et al. compared 14 methods in terms of computational runtime, the ability to handle large datasets, and batch-effect correction efficacy, while preserving cell type purity (46). Based on their results, Harmony (47), Liger (48), and Seurat 3 are the recommended methods for batch integration. Due to its significantly shorter runtime, Harmony is recommended as the first method to try, with the other methods as viable alternatives. Luecken et al. found in their data representing >1.2 million cells that highly variable gene selection improves the performance of data integration methods, whereas scaling pushes methods to prioritize batch removal over conservation of biological variation (49).

**Data correction accounting for unusual droplets**

For droplet-based methods only a fraction of droplets will contain an intact cell. Since biological experiments are never flawless, some RNA will leak out of dead or damaged cells, thereby producing ambient RNA. Droplets without an intact cell can still contain this ambient RNA, which in turn will contribute to the sequencing library and final reads. Variations in droplet size, amplification efficiency, and sequencing cause a wide range of library sizes for both “RNA background” and real cells. Most methods try to distinguish between them by utilizing the total molecules/reads per barcode to find an “inflection point” between larger libraries representing cells and smaller libraries assumed to be only background. Using knee plots one can visualize this inflection point where the total number of molecules per barcode suddenly drops. The R package DropletsUtils uses the complete count matrix of all droplets to assess the profile of ambient RNA from those droplets with extremely low counts. Gene-expression profiles deviating from this background are considered as originating from intact cells. Since background RNA often looks similar to the expression profile of the largest cell population, this is combined with an inflection point method. By this means, EmptyDrops can verify barcodes for very small cells in highly diverse samples.
Expression recovery corrects for zero or low read counts

An additional challenge during the analysis of scRNA-seq data derives from the low transcript capture and sequencing efficiency of current methods. This leads to a large proportion of genes (often more than 90%) with zero or low read counts \(^{50}\). Although many of the observed zero counts reflect a true absence of expression, a considerable fraction is due to technical factors that can vary between less than 1% and more than 60% across cells \(^{17}\). Early expression recovery approaches pooled data for each gene across similar cells, but this may lead to over-smoothing and can disturb the natural cell-to-cell stochasticity in gene expression. For this reason, more advanced approaches such as single-cell analysis via expression recovery (Saver) \(^{50}\), were introduced. Saver assumes that the count of each gene in each cell follows a negative binomial model, which is used to estimate the recovered expression module. In contrast to approaches that impute dropout events by borrowing information across only genes or cells, scTSSR simultaneously leverages information from both similar genes and similar cells using a two-side sparse self-representation model and shows superior use in specific cases \(^{51}\). Another approach showing an actual benefit over existing ones is Viper, which is based on nonnegative sparse regression models \(^{52}\). Viper can progressively infer a sparse set of local neighborhood cells instead of only using similarly expressed genes and cells.

Feature selection filters the most important genes

Feature selection is the process of choosing genes that contain useful information about the underlying biology of the sample, while removing genes that contain no useful information or random noise. This reduces the data size to facilitate the computationally time-consuming steps, while aiming to preserve the relevant biological structure between cells. A simple approach for feature selection is to pick the most variable genes based on their expression between the identified or clustered cell populations. This assumes that actual biological differences occur as increased variation of highly regulated genes, in contrast to genes that are not changed at baseline level or are only slightly regulated via technical noise. There are built in standard approaches for feature selection in Seurat 3, but also further advanced ones available based on a multinomial model \(^{53}\) or ensemble feature selection and similarity measurements \(^{54}\). After the technical process of choosing appropriate features per cellular group, manual curation and cross-validation with existing enrichment databases, for example, Enrichr \(^{55}\), or gprofiler \(^{56}\), may help to evaluate the in silico findings. For a first description of novel cell populations or subpopulations, experimental validation of highly expressed genes is still considered the gold standard.

Dimensionality reduction highly facilitates data interpretability

Dimensionality reduction is an essential tool required to tame the highly complex information content in scRNA-Seq data analysis. A proper reduction of the dimensions allows for effective noise removal and is pivotal for many downstream analyses that include cell clustering or lineage reconstruction. Principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), and uniform
manifold approximation and projection (UMAP), as well as many extensions of these three, are commonly used algorithms in scRNA-seq. Sun et al. compared 18 different dimensionality reduction methods on 30 publicly available scRNA-seq datasets (57). They suggest that applying sophisticated gene filtering approaches prior to running dimensionality reduction will help to improve their performance. In addition, they see a benefit in even more stringent gene filtering approaches because these result in a smaller subset of genes and, therefore, facilitate the application of some of the slow dimensionality reduction methods to larger data sets. A major problem during dimensionality reduction is to preserve the global structure of the data because removing dimensions might likewise suppress some information. Some algorithms, such as the scvis algorithm, try to overcome this limitation by computing low-dimensional embeddings of scRNA-seq data while preserving global structure of the high-dimensional measurements (58). Recently, Heiser et al. presented an unbiased framework that defines metrics of global and local structure preservation in dimensionality reduction transformations (59).

**Important visualizations to share results**

Common visualizations in single cell related publications include 2D and 3D clustering, heatmaps of highly expressed genes, violin plots, and dotplots (60, 61). A recent comparison of interactive single cell visualization tools (for example celxgene, Loom-viewer, iSEE, single-cell explorer) was made by Cakir et al. who regard this specific type of visualization as beneficial to the whole research community, facilitating scientific progress (62). One can also consider using R-Shiny approaches to turn one’s own research results into interactively explorable plots (63). To enrich one’s own research results with additional data, one can utilize scAVI (http://amp.pharm.mssm.edu/scavi/ [accessed on 13 January 2021]), which is a web-based platform developed to enable users to analyze and visualize published and unpublished scRNA-seq datasets with state-of-the-art algorithms and visualization methods. The scAVI platform supports the analysis and visualization of 463 publicly available scRNA-seq studies from GEO covering 194,653 single cells. Once there is a very high number of cells within an analysis, algorithms like “TooManyCells”, a suite of graph-based algorithms, can be used to efficiently identify and visualize cell populations (64).

**DOWNSTREAM ANALYSES**

After utilizing the main processing steps and visualizing the scRNA-seq data, more in-depth analyses should be carried out at a cellular level and at the gene level to fully decipher the identified single-cell profiles. Here, we present further advanced approaches that will help to exploit the information content of a sample of interest.

**Cluster analysis and cell cluster annotation**

A major use case of scRNA-seq is to identify, quantify, and characterize cell populations in heterogeneous samples or tissues. From a biological perspective,
such cell populations include different cell types, or may refer to different states of identical cell types, for example, stimulated and unstimulated cells, or cells in a different maturation state. A de novo identification of cells, mathematically speaking, is an unsupervised clustering problem, which has been widely studied with machine learning algorithms, and there are several well-established strategies that have been adapted for scRNASeq data (65). However, the annotation of the newly identified clusters is still a bottleneck in terms of time consumption and expertise needed because one has to manually curate the clusters using so-called cell atlases, such as “Cellatlas” (https://data.humancellatlas.org/ [accessed on 13 January 2021]) or the “Single Cell Expression Atlas” (https://www.ebi.ac.uk/gxa/sc/home [accessed on 13 January 2021]). To overcome these limitations, promising automated annotation tools, such as SCSA (66) or further ML-based oversampling techniques of Bej et al. (67) have already shown significant potential.

Trajectory analysis and inference to investigate cellular origins

The newly acquired resolution of scRNA-seq allows researchers to distinguish between closely related cell populations, potentially revealing functionally distinct groups with complex relationships (68). For many cellular investigations, there are no distinct borders between cellular states, but instead a smooth transition, where individual cells represent points along a continuum or lineage, in which cells change states by undergoing gradual transcriptional changes, representing a temporal variable or pseudotime (69). The inference of lineage structures is considered as pseudotemporal reconstruction of a sample of interest that finally infers changing cell states and cell fate decisions (70). In addition, many cell populations contain several lineages that share a common initial group branching into different further subgroups, which requires additional analyses to distinguish between cells that fall along those different lineages (71). The two most popular approaches for pseudotemporal reconstruction are Monocle (70) and Slingshot (69), which have been recently compared and benchmarked (54, 69). These benchmarks showed that dimensionality reduction results based on Monocle3 are in line with recommendations by the Monocle3 software itself, which uses UMAP as the default dimensionality reduction method (72). Moreover, the set of the best dimensionality reduction methods for Monocle3 are consistent with those for Slingshot, with only one method difference between the two (GLMPCA [generalized principal component analysis] in place of common PCA).

Gene expression dynamics and pseudotime can reveal cell fate decisions

A central challenge in trajectory inference is the destructive nature of scRNA-seq, which reveals only static snapshots of cellular states; additional information is required to constrain possible dynamics that could make a reasonable prediction towards the same trajectory (73). The concept of RNA velocity enabled the investigation of such dynamic information by assuming that newly transcribed (unspliced pre-mRNAs containing introns) and mature (spliced mRNAs) can be distinguished in common scRNA-seq protocols (74). However, errors in velocity estimates may arise if the central assumptions of a common splicing rate and the
observation of the full splicing dynamics with steady-state mRNA levels are violated (75). Thus, Bergen et al. (75) developed “scVelo”, as a further extension of “RNAvelocity” and solving its limitations, by utilizing the full transcriptional dynamics of splicing kinetics using a likelihood-based dynamical model. This model generalizes RNA velocity computations with transient cell states, which are common in the development of and the response to perturbations. Another interesting approach to simulate time-series trajectories is proposed by Yeo et al. (76), who are using a generative model framework that is able to predict trajectories for cells, which are not found in the model’s training set (including cells in which genes or sets of genes have been perturbed).

**Differential expression testing as a major hurdle in data analysis**

In general, scRNA-seq and bulk RNA-seq data have different characteristics that require a new differential expression (DE) analysis definition beyond the common nonzero difference in average expression, which is not adequately addressed yet. Due to the small amount and low capture efficiency of RNA molecules in single cells, many transcripts tend to be missed during reverse transcription. As a result, one observes that some transcripts are highly expressed in one cell but are not expressed in another cell of the same population, which is defined as a “drop-out” event (77). In addition, multimodality, heterogeneity, and sparsity (many zero counts) are the major hurdles for an effective DE calling. A comprehensive, comparative study of differential gene expression analysis tools for single-cell RNA sequencing data was recently made by Wang et al. (78). They observed that the agreement of tools calling DE genes is not high (~10%) and concluded that there is a trade-off between true-positive rates and the precision of calling DE genes. Methods with higher true positive rates tend to show low precision due to introducing false positives, whereas methods with high precision show low true positive rates due to identifying few DE genes. One solution could be to use different DE testing models, such as methods that can capture multimodality (for example, scDD (79)) and model-based approaches (for example, Monocle) designed for handling zero counts.

**Gene regulatory networks and disease maps**

Gene expression is highly regulated by transcription factors, co-factors, and signaling molecules that span cross-related networks. An improved understanding of these networks is a major goal in biology and medicine because it determines essential factors that are responsible for healthy and disease related phenotypes. So far, molecular networks have been solely based on microarray and bulk RNA-seq data, but are now being refined with single-cell resolution. A recent systematic evaluation of state-of-the-art algorithms for inferring gene regulatory networks from single-cell transcriptional data contributes to the use of single-cell data to parametrize novel networks and improves already existing ones (80). In addition, further algorithms are used to infer global, large-scale regulatory networks on organ scale and perturbed systems, such as diabetes and Alzheimer’s disease (81), or reconstruct networks by using scRNA-seq data of barcoded genotypes (82). Novel disease oriented network applications, like the inflammation
resolution disease map, are able to integrate various kinds of omics data, such as scRNA-Seq, to simulate all relevant molecular processes and serve as a comprehensive knowledge base on single-cell level (83).

CONCLUSION

Taken together, scRNA-seq already offers huge potential for many biological and biomedical areas and will be further applied to decipher currently undiscovered molecular processes. Nevertheless, one must be aware of current sequencing technologies and must determine if scRNA-seq is necessary or can be circumvented with bulk RNA-seq of purified or individual cell types. From a computational perspective, a lot of data analysis steps have potential for optimization and require an even higher awareness of complexity than bulk RNA-seq analysis. The presented procedures and technologies are a current snapshot in the lively field of scRNA-seq, which may rapidly develop in the near future.

Acknowledgement: This work was supported by the EU Social Fund (ESF/14-BM-A55-0024/18, ESF/14-BM-A55-0027/18, ESF/14-BM-A55-0028/18), the “Deutsche Forschungs- gemeinschaft” - DFG (DA1296/6-1), the German Heart Foundation (F/01/12), the FORUN Program of Rostock Medical University (889001/889003), the Josef and Käthe Klinz Foundation (T319/29737/2017), the Damp Foundation (2016-11), and the Federal Ministry of Education and Research - BMBF (VIP+00240, 031L0106C).

Conflict of interest: The authors declare no potential conflict of interest with respect to research, authorship, and/or publication of this chapter.

Copyright and permission statement: The authors confirm that the materials included in this chapter do not violate copyright laws. Where relevant, appropriate permissions have been obtained from the original copyright holder(s), and all original sources have been appropriately acknowledged or referenced.

REFERENCES


