nature protocols

REVIEW ARTICLE

https://doi.org/10.1038/s41596-018-0073-y

Tutorial: guidelines for the experimental design of single-cell RNA sequencing studies

Atefeh Lafzi[1,5], Catia Moutinho[1,5], Simone Picelli[2,4], Holger Heyn[1,3*]

Single-cell RNA sequencing is at the forefront of high-resolution phenotyping experiments for complex samples. Although this methodology requires specialized equipment and expertise, it is now widely applied in research. However, it is challenging to create broadly applicable experimental designs because each experiment requires the user to make informed decisions about sample preparation, RNA sequencing and data analysis. To facilitate this decision-making process, in this tutorial we summarize current methodological and analytical options, and discuss their suitability for a range of research scenarios. Specifically, we provide information about best practices for the separation of individual cells and provide an overview of current single-cell capture methods at different cellular resolutions and scales. Methods for the preparation of RNA sequencing libraries vary profoundly across applications, and we discuss features important for an informed selection process. An erroneous or biased analysis can lead to misinterpretations or obscure biologically important information. We provide a guide to the major data processing steps and options for meaningful data interpretation. These guidelines will serve as a reference to support users in building a single-cell experimental framework —from sample preparation to data interpretation—that is tailored to the underlying research context.

Single-cell transcriptomics studies have markedly improved our understanding of the complexity of tissues, organs and organisms[1]. Gene-expression profiling in individual cells has revealed an unprecedented variety of cell types and subpopulations that were invisible with traditional experimental techniques. As well as providing profound insights into cell composition, single-cell studies have changed established paradigms regarding cell plasticity in dynamic processes such as development[2] and differentiation[3]. Cell states are now known to be more flexible than previously thought, and present multipotent characteristics before reaching fate-decision endpoints. Although various approaches are available for phenotyping of individual cells (e.g., transcriptomics[4], proteomics[5] and epigenomics[6]), single-cell RNA sequencing (scRNA-seq) is currently at the forefront, facilitating ever-larger-scale experiments. The scalability of scRNA-seq experiments has advanced rapidly through the use of automation and sophisticated microfluidics systems, producing datasets from more than 1 million cells[7]. As a result, experimental designs have shifted from a focus on specific cell types to unbiased analysis of entire organs[8–10] and organisms[11,12], thereby enabling a hypothesis-free approach to exploration of the cellular composition of a sample.

Most scRNA-seq methods are now broadly applied in both basic research and clinically translational contexts, even though they require specialized equipment and expertise in sample handling, sequencing-library preparation and data analysis. As a result, single-cell research has become one of the fastest-growing fields in life science, producing fascinating new insights into tissue composition and dynamic biological processes. Large-scale scRNA-seq experiments have yielded cellular maps of *Caenorhabditis elegans*[12], the planarian *Schmidtea mediterranea*[13], *Drosophila*[11,14] and different mouse organs[8,15] to be defined. In humans, single-cell analysis has improved understanding of developmental processes[16], aging[17] and different diseases such as cancer[18–21]. However, it is challenging to create generalizable designs for single-cell transcriptomic experiments because each one requires the user to make informed decisions in order to obtain interpretable results. These include the selection of sample types, cell numbers and preparation methods; the choice of scRNA-seq techniques and sequencing parameters; and the design of computational analysis strategies to generate insights from single-cell datasets. Ultimately, successful single-cell transcriptomic studies with interpretable datasets and meaningful scientific output can be achieved only through the use of tailored experimental designs. To inform this decision-making process, in this tutorial we provide a comprehensive description of the phases of single-cell transcriptomic studies, including (1) sample preparation, (2) scRNA-seq, (3) data processing and (4) data analysis (as discussed further below; see Fig. 1). We summarize the methodological and analytical options and highlight their suitability for distinct research scenarios to support users in designing an end-to-end experimental framework tailored to the underlying

[1]CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain. [2]Research Institute for Neurodegenerative Diseases (DZNE), Bonn, Germany. [3]Universitat Pompeu Fabra (UPF), Barcelona, Spain. [4]Present address: Institute of Molecular and Clinical Ophthalmology Basel (IOB) Basel, Switzerland . [5]These authors contributed equally: Atefeh Lafzi, Catia Moutinho.
*e-mail: holger.heyn@cnag.crg.eu

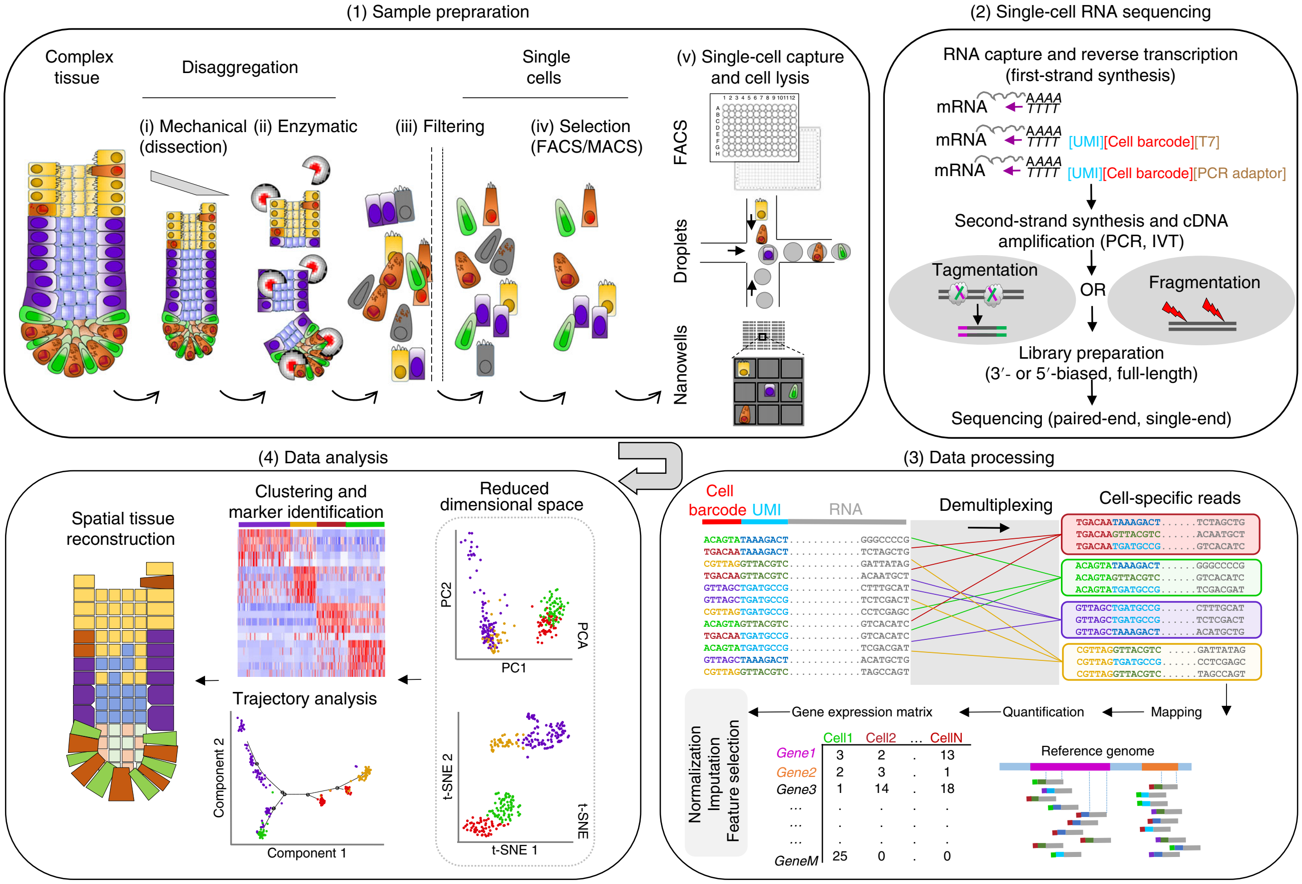
(1) Sample preparation
Complex tissue
Disaggregation
Single cells
(i) Mechanical (dissection)
(ii) Enzymatic
(iii) Filtering
(iv) Selection (FACS/MACS)
(v) Single-cell capture and cell lysis
FACS
Droplets
Nanowells
(2) Single-cell RNA sequencing
RNA capture and reverse transcription (first-strand synthesis)
mRNA AAAA TTTT
mRNA AAAA TTTT [UMI][Cell barcode][T7]
mRNA AAAA TTTT [UMI][Cell barcode][PCR adaptor]
Second-strand synthesis and cDNA amplification (PCR, IVT)
Tagmentation
OR
Fragmentation
Library preparation (3′- or 5′-biased, full-length)
Sequencing (paired-end, single-end)
(3) Data processing
Cell barcode UMI RNA
ACAGTATAAAGACT............GGGCCCCG
TGACAATAAAGACT............TCTAGCTG
CGTTAGGTTACGTC............GATTATAG
TGACAAGTTACGTC............ACAATGCT
GTTAGCTGATGCCG............CTTTGCAT
GTTAGCTGATGCCG............TCTCGACT
CGTTAGTGATGCCG............CCTCGAGC
ACAGTAGTTACGTC............GTCACATC
TGACAATGATGCCG............GTCACATC
ACAGTATGATGCCG............TCGACGAT
GTTAGCTAAAGACT............ACATGCTG
CGTTAGGTTACGTC............TAGCCAGT
Demultiplexing
Cell-specific reads
TGACAATAAAGACT......TCTAGCTG
TGACAAGTTACGTC......ACAATGCT
TGACAATGATGCCG......GTCACATC
ACAGTATAAAGACT......GGGCCCCG
ACAGTAGTTACGTC......GTCACATC
ACAGTATGATGCCG......TCGACGAT
GTTAGCTGATGCCG......CTTTGCAT
GTTAGCTGATGCCG......TCTCGACT
GTTAGCTAAAGACT......ACATGCTG
CGTTAGGTTACGTC .....GATTATAG
CGTTAGTGATGCCG......CCTCGAGC
CGTTAGGTTACGTC......TAGCCAGT
Mapping
Quantification
Gene expression matrix
Reference genome
Cell1 Cell2 ... CellN
Gene1 3 2 . 13
Gene2 2 3 . 1
Gene3 1 14 . 18
GeneM 25 0 . 0
Normalization
Imputation
Feature selection
(4) Data analysis
Spatial tissue reconstruction
Clustering and marker identification
Trajectory analysis
Component 2
Component 1
Reduced dimensional space
PC2
PC1
PCA
t-SNE 2
t-SNE 1
t-SNE

Fig. 1 | The single-cell RNA sequencing process. The successful design of single-cell transcriptomics experiments includes four major phases: (1) During sample preparation, cells are physically separated into a single-cell solution from which specific cell types can be enriched or excluded (optional). After they have been captured in wells or droplets, single cells are lysed, and the RNA is released for subsequent processing. (2) To convert RNA into sequencing-ready libraries, poly(A)-tailed RNA molecules are captured on poly(T) oligonucleotides that can contain unique molecular identifier (UMI) sequences and single-cell-specific barcodes (5′- and 3′-biased methods). To allow for subsequent amplification of the RNA by PCR or IVT, adaptors or T7 polymerase promoter sequences, respectively, are included in the oligonucleotides. After RT into cDNA and second-strand synthesis (optional), the transcriptome is amplified (PCR or IVT). For conversion into sequencing libraries, the amplicons are fragmented by enzymatic (e.g., tagmentation) or mechanical (e.g., ultrasound) forces. Sequencing adaptors are attached during a final amplification step. Full-length sequencing can be carried out, or 5′ or 3′ transcript ends can be selected for sequencing using specific amplification primers (optional). For most applications, paired-end sequencing is required. (3) The sequencing reads are demultiplexed on the basis of cell-specific barcodes and mapped to the respective reference genome. UMI sequences are used for the digital counting of RNA molecules and for correction of amplification biases. The resulting gene-expression quantification matrix can subsequently be normalized, and missing values imputed, before informative genes are extracted for the analysis. (4) Dimensional-reduction representations guide the estimation of sample heterogeneity and the data interpretation. Data analysis can then be tailored to the underlying dataset, which allows cells to be clustered into potential cell types and states, or ordered along a predicted trajectory in pseudotime. Eventually, the spatial cellular organization can be reconstructed through the interrogation of marker genes (experimentally) or through marker-guided computational reconstruction (inference). PC, principal component.

research context (a glossary of relevant terms is provided in Table 1).

Sample preparation

Preparation of high-quality single-cell suspensions is key to successful single-cell studies. Irrespective of the starting material, the condition of the cells is critical for efficient cell capture and optimal performance of the scRNA-seq protocols. Although most methods use fresh viable single cells, alternatives include preserved samples[22–24] and nuclear RNA from frozen tissue[25–29]. Here we provide common general guidelines applicable to all tissues, and optimized parameters tailored to the major tissues of interest. In principle, scRNA-seq applications are not restricted to specific species as long as poly(A)-tailed RNA is present. However, some organisms might require additional processing steps to efficiently release molecules into the reactions (e.g., cell wall removal for plant material).

Good practices for sterile sample handling are recommended, including the use of nuclease-free reagents and consumables. To minimize cell damage, pipetting and centrifugation should be kept to a minimum. Cell concentration and size both influence pelleting efficiency at a given centrifugation speed, time and temperature, and a tightly packed cell pellet may require extra pipetting, which can damage cells

Table 1 | Glossary

Term	Definition
Algorithm	A process or set of rules to be followed in computational calculations or other problem-solving operations.
Barcode	A stretch of sequence used to uniquely label DNA/RNA molecules, cells or sequencing libraries (to allow multiplexing).
Batch effect	A technical source of variation added during sample handling.
Benchmark	Systematic comparison of different techniques (experimental or computational) for their performance in a given scenario.
Binary classifier	A classification function that predicts the assignment of an element to a set of groups.
Bulk RNA sequencing	The sequencing of RNA isolated from pools of cells.
Cell barcode	A cell-specific unique sequence tag that is added to RNA transcripts during library preparation.
Cell capture	Positioning of single cells in reaction volumes (e.g., droplets or wells) for downstream processing.
Cluster annotation	Assigning a function or identity to a group of cells on the basis of the expression of marker genes.
Clustering	The task of grouping cells in such a way that cells in the same group (cluster) are more similar to each other than to cells of another group.
Combinatorial barcoding	The use of combinations of cell barcodes with repeated assignment of barcodes to cells during multiple indexing rounds.
Deconvolution	A process of resolving a complex mixture (e.g., tissue) into its constituent elements (e.g., underlying cell types).
Demultiplexing	The process of separating the elements of interest in a mixed or multiplexed sample.
Digital counting	The counting of RNA molecules using UMI sequences.
Doublets	Two cells that are processed together in a reaction volume (e.g., a well or droplet) and receive the same single-cell barcode.
Dropout events	Transcripts that are not detected in the final dataset even though the gene is expressed in the cell, leading to false zero values in the expression matrix.
FASTQ reads	A sequence composed of the four nucleotides (ACGT) obtained after sequencing in a specific format that represents the chain of nucleotides.
Gene expression matrix	A data matrix containing information about the level of gene expression per cell.
Imputation	The process of replacing missing data with inferred values.
Index sorting	The isolation of single cells by FACS and the retrospective assignment of fluorescence signals during scRNA-seq data analysis.
Library	DNA molecules that contain specific sequences (primers) that enable the initiation of high-throughput sequencing reactions.
Locked nucleic acids	Modified RNA nucleotides with a bridge connecting the 2′ oxygen and 4′ carbon to increase the hybridization properties of oligonucleotides.
Microtiter plates	Also known as microplates or microwell plates; flat plates with multiple wells used as individual reaction sites.
Pipeline	An analysis procedure in which inputs go through a number of processing steps chained together to produce an output.
Poisson distribution	A discrete probability distribution that expresses the probability of the number of events in specified intervals such as distance, area or volume.
Pooling	Combining molecules or cells for their joint processing.
Promoter	A DNA sequence that initiates transcription of the downstream sequence.
Pseudotime	An inferred time line of the progress cells make through a dynamic process such as cell differentiation.
Spike-in RNA	A pool of RNA transcripts of known sequence composition and quantity used to calibrate experiments.
Tagmentation	Reaction that involves the transposase-based cleaving of DNA and the tagging of the double-stranded DNA with universal overhangs.
Template-switching oligonucleotide (TSO)	A DNA oligonucleotide sequence that carries three riboguanosines (rGrGrG) at its 3′ end and binds to the cytosine extension of the cDNA molecules after RT.
Trajectory inference	Computational reconstruction of an underlying cellular developmental or differentiation path.
Unique molecular identifiers (UMIs)	Random sequences attached to transcripts and used as molecular tags to detect and quantify unique RNA molecules.
Zero-inflated data	Data with an excess of zero counts. To model zero-inflated data, a Poisson distribution is used.

through shearing effects; thus, centrifugation conditions should be optimized. Sufficient volumes should be used for cell washing and resuspension, as high concentrations can cause aggregation and clumping. Suspensions should be filtered with appropriately sized cell strainers (pore size larger than cell diameter) to remove clumps and debris. The recommended

Table 2 | Tissue-specific enzymatic treatments to prepare single-cell suspensions (from human and mouse samples)

Tissue	Digestion enzyme	Time (min)	Temperature (°C)	Final concentration	Ref.
Liver	Collagenase IV	10	37	0.16 mg/ml	126
	Liberase Blendzyme 3	5–8	37	40 µg/ml	9
	Collagenase, collagenase D and Pronase, trypsin	20, 20, 10	37	2.5 mg/ml, 10 mg/ml and 10 mg/ml, 0.05%	127
	Collagenase IV	30	37	0.05%	128
Lung	Dispase and elastase	45	37	0.33 U/ml and 3 U/ml	129
	Collagenase and dispase	45	37	0.2% solution	130
	Dispase, elastase and trypsin	60, 30, 15	4, 37 and 37	2 mg/ml, 5 U/ml plus 0.125%,	131
Skin	Trypsin	120	32	1×	132
	Liberase TL	15	37	2 mg/ml	133
Spleen	Collagenase D	45	37	2 mg/ml	134
GI tract	Dispase	20	37	0.4 mg/ml	36
	Trypsin	30	37	2 mg/ml	135
	TrypLE Express	1	37	1×	10
	Collagenase	40	37	1 mg/ml	136
	Collagenase I	60	37	2.5 mg/ml	137
	Collagenase IV	30	37	2 mg/ml	138
Pancreas	Collagenase type CLS IV	30	37	1 mg/ml	139
	Collagenase P	30	37	0.8 mM	140
	TrypLE Express	1	37	1×	141
	Accutase and TrypLE Express	10 and 5–20	37	1×	142
	Accutase	8-10	37	1×	143
	Trypsin	30	37	1×	144
Kidney	Liberase TL	15	37	2 mg/ml	133
Retina	Papain	45	37	4 U/ml	61,145
	Accutase	5	37	1×	146

cell-washing and resuspension solution is phosphate-buffered saline (calcium and magnesium free) containing bovine serum albumin to minimize cell losses and aggregation. Primary cells, stem cells and other sensitive cell types may require washing and suspension in alternative buffers to ensure viability, which also may decrease when cells are kept in suspension for a prolonged period. Cell clumps cause automated cell counters to underestimate the effective concentration of single cells, so suspensions should be processed as soon as possible after preparation, ideally within 30 min. It is important to minimize cellular aggregates, dead cells, noncellular nucleic acids and reverse-transcription (RT) inhibitors in single-cell preparations. To minimize these contaminants while maximizing the purity and unbiased recovery of different cell types, one may need to apply optimization (e.g., adjust the number of wash steps, the composition of the wash solution, centrifugation conditions and/or strainer type).

Preparation of cell suspensions. For isolation of single cells from suspensions (e.g., blood samples), samples are density centrifuged (e.g., using Ficoll-Paque or Histopaque-1077 techniques)[30], after which they can be used directly for single-cell capture. Solid tissues must first be dissociated via mechanical and enzymatic treatment. Initially, tissues are disaggregated by mechanical cutting or mincing with blades. Then enzymatic digestion is used to separate cells, with specific enzymes and digestion times used for different tissues (Table 2). Enzyme types include Accutase, elastase and collagenases, as well as commercial enzymatic mixtures such as TrypLE Express and Liberase Blendzyme 3. Elevated cell lysis can lead to cell clumping, which is reduced through treatment with DNase I during cell separation. Finally, suspensions are cleaned by filtering through a mesh or strainer before capture of single cells.

It is important to note that sample processing might introduce variation in the gene expression profile, as has been shown for the activation of stress-related genes[31]. Also, some more sensitive cell types might be damaged during sample preparation, so processing time should be kept to the minimum required. In contrast, too short digestion times could result in incomplete cell separation and the exclusion of tightly interconnected cells from subsequent single-cell analysis.

To avoid biases in cell type composition, one can use an alternative strategy that involves disruption of cellular membranes and isolation of the nuclei[25–29]. The sequencing of nuclear RNA was shown to be sufficient to deconvolute cell types[29], although this decreases the overall resolution per cell. Single-nuclei sequencing has been applied extensively for differentiated neurons, for example, as it is largely impracticable to isolate intact cells from highly interconnected adult neuronal tissue.

Single-cell capture. For transcriptome profiling in single cells, most methods require the physical isolation of cells in individual reaction volumes. Cells can be isolated by microdissection or pipetting[32], although high-throughput experiments use fluorescence-activated cell sorting (FACS)[33] or microfluidics[34] to guide cells into micro- or nanoliter reaction volumes, respectively. Microfluidic systems capture cells in integrated fluidics circuits (IFCs), droplets or nanowells, thus allowing thousands of cells to be processed simultaneously while minimizing reaction volumes and reagent use. FACS sorts cells into microtiter plates ready for library preparation by manual or automated processing, and facilitates the exclusion of dead or damaged cells, as well as the enrichment of target cell populations (e.g., through surface marker labeling). To reduce background and maximize assay performance, we also recommend FACS or magnetic-activated cell sorting (MACS) processing of single-cell solutions for microfluidic systems to remove debris, damaged/dead cells and cell aggregates.

Sample size and composition. To obtain an unbiased view of the cellular composition of a sample, one must capture all cells during the isolation process. Here attention must be paid to very small or large cells that may be excluded during FACS isolation or captured in microfluidic systems, respectively. However, for many experiments, it may be necessary to enrich for or exclude some cell types to increase the total number of cells of interest in the final scRNA-seq libraries. For example, profiling of specific immune responses requires enrichment of blood cell subtypes, whereas cancer studies might need to exclude blood cells (e.g., CD45$^+$ cells) to increase the overall number of tumor cells. Target populations can be selected by FACS and MACS with appropriate labeling (e.g., antibodies or transgenic systems). Microtiter plates and some nanowell capture systems allow index sorting, in which fluorescence intensity or cell size (FACS information) is associated with capture coordinates and subsequently with single-cell indices. The FACS device records the sorting position and intensity values of a given cell, thereby enabling the subsequent integration of transcriptome profiles with the recorded cell properties. For microfluidic systems, CITE-seq[35] provides a viable alternative that conserves information about surface markers. Here epitopes of interest are targeted with oligonucleotide-labeled antibodies. The antibody-specific sequences are poly(A)-tailed and contain barcodes that allow epitope tracking after scRNA-seq library preparation and sequencing.

To define adequate cell numbers per experiment, one must consider sample heterogeneity and subpopulation frequency (the estimated abundance of the cell type of interest). In particular, larger cell numbers are required to resolve the structure of heterogeneous samples with many expected subpopulations. Also, the total number of cells required increases when rare cell types need to be identified. One can calculate the required cell numbers by estimating both subpopulation structure and low-frequency cell-type abundance and defining the desired cell number per group (computational tool accessible at https://satijalab.org/howmanycells). Because most experiments target poorly described systems, heterogeneity can only be estimated, so pilot experiments are recommended before large-scale data production. For comparative studies across experimental conditions, patient samples or larger population cohorts, control experiments can be used to provide information about optimal cell numbers and the need for subpopulation enrichment steps. Specifically, selected samples can be profiled with high cell numbers to comprehensively identify tissue heterogeneity. Cell numbers in subsequent data production phases can then be adapted according to the required resolution. Similarly, seemingly homogeneous samples can be initially profiled using higher cell numbers and sequencing depth to reveal yet uncharted sample complexity. Note that higher cell numbers can also be beneficial for homogeneous samples, as this increases statistical power during analysis[36].

Sample preservation. All common scRNA-seq methods were initially designed to use freshly isolated cells. However, in research and clinical practice, immediate sample processing can be challenging because of a lack of the required infrastructure or specialized equipment, such as FACS devices. Moreover, although samples may be collected at multiple time points, simultaneous sample processing may be preferred to avoid technical batch effects. Sample preservation is a viable solution because it disconnects the location and time of sampling from the downstream processing steps. In this context, cryopreservation has been established for single-cell transcriptome analysis[22]. After sample storage for up to a year at −80 °C or in liquid nitrogen and subsequent thawing, cryopreserved cells from cell lines and primary samples show complete integrity of the RNA molecules and unchanged expression profiles as compared with those of freshly prepared cells. Note that multiple freeze–thaw cycles should be avoided through the preparation of aliquots or by scraping out still-frozen cells from storage vials. Similarly, methanol fixation has been established as an alternative for droplet-based single-cell methods, and could also be used to avoid technically induced variations in gene expression triggered by prolonged sample processing time[23]. Importantly, both methods allow the archiving and transport of samples and broaden the range of applications of scRNA-seq methods, for example, to the clinical context. However, both approaches have shown a potential bias in cell-type composition, and it is strongly recommended to thoroughly evaluate preservation methods for new cell types that have not been tested. For previously archived samples, such as snap-frozen specimens, nuclei sequencing provides the only solution for scRNA-seq[25–29]. Unlike in cryopreservation, the formation of ice crystals during snap-freezing disrupts the outer cellular membrane, although the nuclei remain intact. Nevertheless, it is preferable to make an initial estimation of the RNA integrity to avoid biases related to sample quality.

Table 3 | Key features of microtiter-plate- and microfluidics-based single-cell RNA sequencing methods

Method	Capture format	Cell loading	Single-cell indexing	Molecule identifier	Additives in RT	cDNA amplification	Fragmentation	Transcript coverage	Sequencing	Ref.
Smart-seq	Plate	FACS	Tagmentation	NA	NA	PCR	Tagmentation	Full length	Paired end	47
Smart-seq2	Plate	FACS	Tagmentation	NA	Betaine	PCR	Tagmentation	Full length	Paired end	147
STRT-seq	Plate	FACS	TSO	UMI	NA	PCR	DNase I	5′ end	Single end	48
STRT-seq-2i	Nanowell	FACS/ Poisson	TSO	UMI	Betaine	PCR	Tagmentation	5′ end	Single end	58
SCRB-seq	Plate	FACS	Oligo(T) primer	UMI	NA	PCR	Tagmentation	3′ end	Paired end	49
mcSCRB-seq	Plate	FACS	Oligo(T) primer	UMI	PEG	PCR	Tagmentation	3′ end	Paired end	50
Quartz-seq	Plate	FACS	Oligo(T) primer	NA	NA	PCR	Ultrasound	Full length	Paired end	51
Quartz-seq2	Plate	FACS	Oligo(T) primer	UMI	NA	PCR	Ultrasound	3′ end	Paired end	52
CEL-seq	Plate	FACS	Oligo(T) primer	NA	NA	IVT	KOAc, MgOAc	3′ end	Paired end	32
CEL-seq2	Plate	FACS	Oligo(T) primer	UMI	NA	IVT	Random priming	3′ end	Paired end	54
MARS-seq	Plate	FACS	Oligo(T) primer	UMI	NA	IVT	Zinc	3′ end	Paired end	53
Seq-Well	Nanowell	Poisson	Oligo(T) beads	UMI	Ficoll	PCR	Tagmentation	3′ end	Paired end	59
inDrops	Droplets	Poisson	Oligo(T) beads	UMI	IGEPAL	IVT	KOAc, MgOAc	3′ end	Paired end	60
Drop-seq	Droplets	Double Poisson	Oligo(T) beads	UMI	Ficoll	PCR	Tagmentation	3′ end	Paired end	61

NA, not applicable.

Single-cell RNA sequencing

Transcriptome profiling of individual cells can be split into four major components: RNA molecule capture, RT and transcriptome amplification, sequencing library preparation, and sequencing. Various scRNA-seq methods exist, but they all apply the same underlying principles. Below we discuss these basic experimental design considerations, and highlight common and emerging microtiter-plate-based and microfluidic scRNA-seq techniques and their applications. Key features of the different scRNA-seq approaches discussed below are also summarized in Table 3. Many of these methods have undergone systematic evaluation, which confirmed their generally high accuracy, although efficiency, scalability and costs vary considerably[37,38]. This should be taken into account during the selection of methods for a given experiment.

RNA molecule capture, reverse transcription and transcriptome amplification for sequencing library preparation. Most scRNA-seq methods, including those described below, capture poly(A)-tailed RNA, although specific protocols are available for profiling total RNA[39,40] or miRNAs[41]. After cell lysis, poly (A)-tailed RNA is captured by poly(T) oligonucleotides, which exclude abundant RNA types such as rRNA and tRNA. After capture, the RNA is reverse-transcribed into stable cDNA, at which point most methods add single-cell-specific barcodes within the poly(T) oligonucleotides that allow cost-effective multiplexed processing of pooled samples. Moreover, random-nucleotide-sequence stretches in the poly(T) oligonucleotide serve as unique molecule identifiers (UMIs) that allow the user to correct for amplification biases and reduce technical noise[42]. RT is a crucial step, and different protocols have been optimized in various ways with efficient enzymes and specific additives that maximize efficiency (Box 1). cDNA can then be amplified by PCR or through in vitro transcription (IVT). For this, adaptor sequences or RNA polymerase promoter sequences are introduced during RT or second-strand synthesis. Although IVT is less prone to biases through linear amplification of molecules, it requires additional downstream steps to convert the amplified RNA into cDNA and sequencing-ready libraries. PCR-based protocols require less hands-on time, but the exponential amplification phase leads to biases in RNA composition in the final libraries. Both approaches were shown to provide interpretable results and were successfully implemented in several scRNA-seq methods (Table 3).

Full-length versus 3′- or 5′-end transcript sequencing. Single-cell transcriptome profiling can be done through full-length transcript analysis or by digital counting of 3′ or 5′ transcript ends[42]. The choice of sequencing method should be dictated by the goal of the experiment—for example, to prioritize cost-effectiveness over retention of sequence information. Digital RNA counting is a cost-effective quantification strategy, although sequence information of the transcripts is lost to a large extent. Full-length transcriptome sequencing allows the detection of splice variants and alternative transcripts, as well as genetic alterations in the transcribed fraction, such as single-nucleotide variants[19,20] and fusion transcripts[43]. Moreover, genotypes of T and B cell receptors can be obtained from full-length transcriptomes[44]. Unlike 3′- and 5′-end methods, full-length protocols do not allow the introduction of UMIs and impede early cellular barcoding and pooling, which results in higher costs for library preparation. This limitation can be overcome through the use of long-read sequencing technologies that do not need library fragmentation[45]. However, such technologies generate smaller quantities of sequencing reads, and transcriptome quantification is not yet possible.

Box 1 | Optimization of reverse transcription for single-cell transcriptome sequencing

Enzymes

Reverse transcription (RT) is one of the most critical steps in the library-preparation workflow. Despite its importance, however, relatively little has been done to improve the efficiency of the underlying enzymes. Reverse transcriptases are based on Moloney murine leukemia virus (MMLV)-derived enzymes, which originally had low processivity and high error rates due to their retroviral origins. Different point mutations have been introduced to improve processivity, resulting in enzymes that can reverse-transcribe even very long RNAs (up to 12–14 kb). SuperScript II is a commonly used enzyme that became popular in the single-cell field because of its template-switching properties, and is used in methods such as Smart-seq2[147] and STRT-seq[48,58]. Most important, SuperScript II carries point mutations that inactivate its RNase H domain, thus impairing competitive RNA degradation during cDNA synthesis. Alternative RT enzymes have been reported to have similar or superior performance, such as Maxima H (used in SCRB-seq[49,50]) and SMARTscribe in the SMARter v4 kit (Takara Bio). Protocols that do not require template switching and that generate second strands by other means, such as poly(A)-tailing or random priming[52,54], can use SuperScript III, which carries different point mutations in the RNA polymerase and has increased thermal stability.

Additives

In an attempt to overcome the limitations of MMLV-based RT enzymes, several additives have been tested over the years. The challenge of generating full-length cDNA libraries has been a constant issue in molecular biology, predating the advent of single-cell RNA sequencing. Carninci et al.[148] showed that the sugar trehalose has a thermo-stabilizing and thermo-protective effect on RT enzymes. Conducting the RT reaction at a higher temperature enhances the unfolding of secondary RNA structures that could hinder enzyme processivity. This finding was confirmed and later extended to the addition of betaine, alone or in combination with trehalose, to improve thermo-protection and related cDNA yield[149,150]. Smart-seq2[147] and STRT-seq-2i[58] use betaine in combination with magnesium chloride; use of the latter at concentrations higher than 1 mM has been suggested to have a synergic destabilizing effect in the presence of betaine[151]. However, the extra magnesium chloride could also reduce the chelating function of 1,4-dithiothreitol (DTT), which is commonly used in RT reactions to guarantee higher cDNA yields and longer transcripts. In the very first published single-cell sequencing method, Tang et al.[152] used the T4 gene 32 protein (T4g32p), a single-stranded binding protein that increases yield and processivity during RT.

Template-switching oligonucleotides

The template-switching reaction relies on 2–5 untemplated cytosine nucleotides, which are added to newly synthesized cDNA (but not to fragmented or uncapped RNAs) when the enzyme reaches the 5′ end of the RNA. The presence of a TSO carrying three complementary guanosines at its 3′ end enables the enzyme to switch templates and to add the complementary sequence of the TSO to the cDNA (including a PCR adaptor for subsequent amplification)). It has been suggested that the reduced RNA capture efficiency of single-cell RNA-seq protocols might be due to the unstable binding of TSO to the untemplated nucleotides. The Smart-seq2 protocol addresses this issue by modifying the last nucleotide of the TSO with a locked nucleic acid. Furthermore, the importance of each nucleotide in the TSO has been extensively evaluated to define its optimal composition[153].

scRNA-seq methodologies: microtiter-plate-based approaches. After isolation of single cells into microtiter plates by FACS, a full-length transcript or 3′/5′-end protocol can be applied. Smart-seq2[46] is a widely used method to reverse-transcribe and amplify full-length transcripts. After RT, the enzyme adds cytosines to the cDNA, providing the basis for a template-switching reaction. Here a template-switching oligonucleotide (TSO) binds to the extra cytosine and provides the template for the addition of PCR adaptor sequences for subsequent cDNA amplification. Compared with the original version[47], the updated protocol improves molecule-capture efficiency and yield by using locked nucleic acids in the TSO and adding betaine to the RT reaction. Sequencing libraries are prepared by tagmentation, which simultaneously fragments and indexes the cells. The Smart-seq2 protocol is highly efficient in capturing RNA molecules[37], although the late indexing step makes it more expensive than other methods. Furthermore, the absence of UMIs makes downstream data analysis more challenging. Nevertheless, the protocol provides an adequate solution if deep single-cell phenotyping is required (e.g., for homogeneous samples or for analysis of weakly expressed genes).

STRT-seq[48] uses a similar strategy for RT and template switching, but it incorporates single-cell barcodes into the TSO. This allows early pooling of cells and cost-effective multiplex processing. STRT-seq enriches 5′ transcript ends through the use of biotinylated purification and 5′-specific PCR primers. Analysis of the 5′ transcript has the advantage of providing information about transcription start sites. Moreover, cell barcodes and transcripts are obtained in a single read, which allows for cost-effective single-end sequencing. Although the original STRT-seq protocol could not correct for amplification biases, later updates for the first time included UMIs in an scRNA-seq method[42]. The SCRB-seq[49] protocol incorporates single-cell barcodes and UMIs in the poly(T) primer, thereby enabling 3′ amplification of transcripts, and, as with STRT-seq, early indexing allows cell pooling to reduce costs. The RNA capture efficiency of the original protocol was improved by an increase in the RT mix density: molecular crowding SCRB-seq (mcSCRB-seq[50]) includes polyethylene glycol to increase binding-event probabilities. In addition, the PCR enzyme was switched from KAPA to the Terra polymerase to further improve library complexity. In Quartz-seq[51], the template-switching reaction is replaced by a poly(A)-tailing step. The additional adenosines provide a template for a poly(T)-primed second-strand synthesis followed by PCR amplification. The amplified transcriptome then undergoes ultrasound fragmentation and sequencing-adaptor ligation. A later version, Quartz-seq2[52], improved the molecule-detection efficiency by using shorter RT primers and improving poly(A)-tagging efficiency.

Amplification biases during exponential PCR are addressed in CEL-seq[32], in which transcripts are copied through IVT. The linear amplification of molecules, made possible by inclusion of a T7 promoter in the poly(T) primer, results in more evenly duplicated transcriptomes. Also, transcriptome amplification by IVT does not require template switching, which improves molecule-capture efficiency. This workflow was further optimized in MARS-seq[53] by inclusion of UMIs in the poly(T) primers and upscaling of cell numbers through automation. In addition, the original CEL-seq protocol was updated in CEL-seq2[54] for more efficient RNA capture and a simplified workflow. Briefly, the CEL-seq2 protocol uses UMIs, a shorter RT primer, and more efficient RT and second-strand synthesis enzymes. Furthermore, cDNA synthesis after IVT is initiated by random priming instead of adaptor ligation.

scRNA-seq methodologies: microfluidic systems-based approaches. Microfluidics allows higher-throughput scRNA-seq workflows, thus eliminating the technical constraints on scalability associated with microtiter plates. Moreover, reducing reaction volumes from microliters to nanoliters reduces costs and technical variability[55] while improving cDNA yield[56]. There are three strategies for capturing cells: IFCs, droplets and nanowells, all of which increase the number of capture sites relative to that achieved with microtiter plates. The first microfluidics system used for scRNA-seq was designed as an automated array solution (Fluidigm C1) in which single cells enter a fluidics circuit and then are immobilized in hydrodynamic traps, lysed, and processed in consecutive nanoliter reaction chambers via a modified Smart-seq2 protocol. Although early versions could use only commercial scRNA-seq assays, a more recent open format accommodates custom scRNA-seq protocols[42] and additional applications for genetics and epigenetics single-cell experiments[57]. Costs were further reduced by an increase in throughput and cell capture from 96 to 800 sites (C1 HT-IFC), and inclusion of an early-indexing strategy that allows cell pooling. Notably, this high-throughput version switched from full-length to 3′ RNA sequencing. Also, the array formats, which are restricted to specific cell sizes (small, medium and large arrays), affect unbiased sampling from complex sample types. To further increase cell numbers, microfluidics progressed to open nanowell systems that allow better scalability. In STRT-seq-2i[58], the original protocol was applied in a nanowell platform with 9,600 sites, with cells loaded by limiting dilution or direct addressable FACS sorting. Positioning cells by FACS allows for index sorting that assigns cell properties (e.g., fluorescence signal or size) to array coordinates and barcodes. Nanowells containing cells can be specifically utilized by targeted dispensing, which substantially reduces reagent costs and contamination by ambient RNA. Moreover, the array format allows imaging to exclude doublets. To guarantee high cell viability during the time-consuming loading into nanowells, FCS can be added to the buffer and sample aliquots can be kept on ice. Alternatively, Seq-Well[59] provides a nanowell-based method that captures cells in 86,000 sub-nanoliter reactions. The underlying principle is the preloading of nanowells with barcoded beads before cells enter the capture sites through limited dilution. Subsequently, the arrays are sealed for cell lysis and RNA molecule capture on beads before the immobilized molecules are pooled for 3′-end library production. The Seq-Well system is portable, and so allows sample processing at the sampling sites, as large equipment is not required. The fact that no major investments are required makes the Seq-Well system a flexible and cost-effective alternative. However, although cells can be monitored by microscopy, the random distribution of barcoded beads does not allow the user to integrate imaging data. Also, the method requires experienced users to obtain reproducible, high-quality results.

Although they are scalable to higher throughputs, the IFC and nanowell approaches are intrinsically constrained by the number of reaction sites. Droplet-based systems overcome this by encapsulating cells in nanoliter microreactor droplets. Here, cell numbers scale linearly with the emulsion volume, and large numbers of droplets are produced at high speed, which facilitates large-scale scRNA-seq experiments. Furthermore, droplet size can be adjusted to reduce potential biases during cell capture. Because barcodes are introduced into droplets randomly, this approach does not allow the assignment of cell barcodes to images and so precludes the visual detection of doublets and the integrative analysis of cell properties (e.g., fluorescent signals) with transcriptome profiles. Two droplet-based methods, inDrops[60] and Drop-seq[61], were developed in parallel, with related commercial systems allowing straightforward implementation. inDrops[60,62] encapsulates cells by using hydrogel beads bearing poly(T) primers with defined barcodes, after which the photo-releasable primers are detached from the beads to improve molecule-capture efficiency and initiate in-drop RT reactions. The barcoded cDNAs are then pooled for linear amplification (IVT) and 3′-end sequencing-library preparation. The technique has extremely high cell-capture efficiency (>75%) owing to the synchronized delivery of deformable beads, allowing near-perfect loading of droplets. Therefore, the system is most suitable for experiments with limited total numbers of cells. The inDrops system is licensed to 1CellBio, and a variant protocol has been commercialized as the Chromium Single Cell 3′ Solution (10x Genomics)[63]. The Chromium system is straightforward to implement and standardize, although library-preparation costs are considerably higher than those of the original system. Unlike inDrops protocols, Drop-seq[61] uses beads with random barcodes. After cell lysis and RNA capture, the drops are broken and pooled, covalent binding is carried out through cDNA synthesis, the cDNA is amplified by PCR, and 3′-end sequencing libraries are produced by tagmentation. Drop-seq has lower cell-capture efficiency than inDrops methods because beads and cells are delivered by double limiting dilution (double Poisson distribution), which results in 2–4% barcoded cells. The Drop-seq system is commercially available through Dolomite Bio, and a similar system is provided by Illumina (ddSEQ).

scRNA-seq methodologies: split-pool barcoding-based approaches. Conceptually different from the above techniques are methods based on combinatorial barcoding. Here, cells are not processed as individual units but isolated in pools. These pools are split and mixed, with each round integrating pool-specific barcodes. The combination of such pool indices results in unique barcode combinations for each cell through their random assignment during consecutive pooling processes. Both split-pooling methods, SPLiT-seq (split-pool ligation-based transcriptome sequencing)[64] and sci-RNA-seq (single-cell combinatorial-indexing RNA-seq)[12], were shown to reliably produce single-cell transcriptomes and to be scalable to hundreds of thousands of cells per experiment. SPLiT-seq includes four rounds of indexing, resulting in >20 million possible barcode combinations. After initial indexing during RT, two rounds of index ligation and a final PCR indexing step create cell-specific barcoded 3′-transcript libraries. During the second ligation round, UMIs are incorporated for the subsequent correction of amplification biases. Additional rounds of barcoding or a switch from 96-well to 384-well microtiter formats could further scale up cell numbers. The original sci-RNA-seq protocol includes a two-step indexing workflow with the first index and UMI introduced during RT and a second index during PCR amplification (after tagmentation). The use of indexed tagmentation sequences could further scale up possible barcode combinations and increase cell numbers per experiment. Formaldehyde- and methanol-based fixation of cells, used in SPLiT-seq and sci-RNA-seq, respectively, allows sample storage, thereby providing additional flexibility to the experimental designs. Both methods allow the processing of nuclei and consequently the analysis of more challenging cell types, such as neurons. The split-pool strategy used in sci-RNA-seq was further shown to be applicable in different single-cell epigenomic analysis approaches, including open chromatin (sci-ATAC-seq[65]), chromatin conformation (sci-Hi-C[66]) and DNA methylation (sci-MET[67]) approaches.

Library preparation and sequencing. In library preparation for short-read sequencing applications, the amplified cDNA (PCR) or RNA (IVT) is fragmented before sequencing adaptors are added. Fragmentation can be achieved enzymatically (with tagmentase or DNase), chemically (with zinc, KOAc or MgOAc) or through mechanic forces (e.g., ultrasound) (Table 3). 3′- or 5′-based libraries are subsequently amplified with primers specific for the transcript end or start, respectively. During this step of the protocol, a pool-specific index can be introduced that allows the multiplexed sequencing of multiple experiments. Full-length methods introduce the cell-specific barcodes only after fragmentation, thus impeding pooled processing of cells at earlier stages of the protocol. Apart from STRT-seq, scRNA-seq libraries require paired-end sequencing, in which one read provides information about the transcripts while the other reads the single-cell barcodes and UMI sequences. STRT-seq incorporates the cell barcode and UMI at the 5′-transcript end, which allows cell, molecule and transcript information to be captured in a single read, as no poly(T) stretch separates the respective sequences. High-throughput microfluidics-based experiments generally involve sequencing to lower depths (<100,000 reads per cell), whereas higher read numbers (~500,000 reads per cell) are optimal for many microtiter-plate formats[38]. Nevertheless, single-cell libraries are usually not sequenced to saturation, and the phenotyping resolution (detection of more genes and of those expressed at lower levels) can benefit from further increases in the sequencing depth. Annotation of splice variants from full-length transcriptomes requires deeper sequencing to better resolve the expression levels of transcript variants.

Further technical considerations

Cell doublets. An intrinsic problem for most microfluidics-based methods is that two cells can be captured per reaction site (nanowell or droplet), both receiving identical barcodes. Doublet rates can be experimentally determined in species-mixture experiments, but otherwise can only be estimated. They occur when cells are positioned randomly in reaction sites by limiting dilution and can be controlled by the cell suspension concentration. The relationship between cell loading and doublet rate was systematically quantified for the Chromium system[63]. Up to the maximal recommended loading of 10,000 cells per droplet lane, the doublet rates showed a linear relationship (in line with the Poisson loading of cells into droplets), with inferred rates ranging from 2% (2,500 cells) to 8% (10,000 cells). Other microfluidics approaches yield similar numbers: Drop-seq, 0.36–11.3% (12.5–100 cells/µl; ref. [61]); InDrops, 4% (ref. [60]); and Seq-Well, 1.6% (ref. [59]). The doublet rate decreases at higher dilutions, with a resulting increase in reagent costs per cell, as fewer total cells are captured per experiment. Researchers can partially overcome this handicap by jointly capturing samples from different individuals, such that genotype differences allow the user to distinguish between donors and thereby reliably identify doublets[68]. Specifically, single-nucleotide polymorphisms identified from the RNA sequencing reads are used to determine the donor origin of the cells and to discriminate samples that were processed in a single batch. However, such a workflow is practicable only when the experimental design includes different human individuals or model organisms with distinct genetic backgrounds. Currently, there is no computational method for credibly identifying doublets, so doublet rates must be minimized by experimental design. Doublets can have dramatic consequences for data interpretation, as artifactual mixed transcriptomes can easily be mistaken for intermediate cell states in dynamic systems.

Cell-capture efficiency. Cell-capture efficiency is an important consideration, especially in experiments involving primary or rare samples. The number of cells that receive barcodes is directly related to the proportion of sample that enters downstream analysis. The capture efficiency of FACS-based methods is constrained by the time the device requires to move between wells. To maximize capture rates of FACS-based methods, one can dilute and sort cell suspensions at low speed (e.g., 100 cells/s). Microfluidics technologies differ markedly in capture efficiency, mainly as a result of cell and bead loading mechanics. The HT-IFC system captures a maximum of 800 out of 6,000 injected cells. In nanowell systems that use limiting dilution for cell loading (no sorting), cells enter reaction sites by gravity, with generally high efficiency. For example, 10,000 cells are added to the surface of a Seq-Well array, and around 3,000 cells are captured. For droplet-based systems, the rate at which cells enter the analysis is directly related to the loading efficiency of the beads. When most droplets contain barcoded beads, cell capture is optimal (inDrops). In contrast, if beads and cells are encapsulated by limiting dilution, most cells do not enter a bead-containing droplet, which results in lower capture efficiency (Drop-seq; discussed above).

Costs. The total cost of scRNA-seq experiments is determined by three main components: equipment, reagents and sequencing. For most methods, the cost of scRNA-seq library preparation scales linearly with cell numbers; an exception is custom droplet methods. The actual costs per cell vary widely across methods and institutes, with microfluidic systems being generally cheaper (<$0.30 per cell) than early-indexing plate-based 3′ digital counting methods (~$1–2 per cell). Late-indexing full-length transcriptome profiling is costlier, even with small volumes (~$8–12 per cell). However, costs can be reduced through the use of non-commercial tagmentase[69] or minimum reaction volumes and automated workflows for plate-based formats[70]. Importantly, microtiter plates can be shipped and stored, which disconnects sampling sites from scRNA-seq processes such that expensive devices can be centralized in core units, thus optimizing resource management. Custom microfluidics methods further decrease costs per cell. Commercialized microfluidics methods are more expensive ($0.50–2.00 per cell) than custom systems (<$0.30 per cell), although their automated design reduces hands-on time and personnel costs.

Although the cost of library preparation is decreasing rapidly, sequencing costs are becoming a major factor. Methods with higher molecule-capture efficiency produce more complex sequencing libraries, which makes them informative at low sequencing depths. Consequently, more efficient scRNA-seq methods can compensate for higher library preparation costs by decreasing overall sequencing costs.

Data processing

Data processing includes all the steps necessary to convert raw sequencing reads into gene expression matrices, using workflows similar to those used for bulk RNA-seq. After FASTQ reads have been generated and their quality has been checked (with tools such as FastQC; http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), the next important step is de-multiplexing of reads using cell barcodes. Whereas Smart-seq libraries can be directly de-multiplexed using the index reads, the 3′-end-based methods require a dedicated processing step to identify the single-cell indexes in the sequencing reads. De-multiplexed reads are then mapped to reference genomes with alignment tools such as TopHat[71] and STAR[72], the latter of which offers proven accuracy and splice-variant sensitivity. Recent alignment tools were optimized for fast handling of large-scale datasets without loss of accuracy. For example, Kallisto[73] reduces the alignment time by two orders of magnitude through pseudo-alignment, as opposed to alignment of individual bases. In a final processing step, mapped reads are quantified to create a transcript expression matrix. RSEM[74], Cufflinks[75] and HTSeq[76] can be used for full-length transcript datasets, whereas special tools, such as UMI-tools[77], which accounts for sequencing errors in UMI sequences, are available for counting UMI-tagged data types.

In addition to the specific tools available for individual processing steps, single-cell data processing pipelines have been developed that combine mapping and quantification steps and include quality control measures for reads and cells. A pipeline developed by Ilicic et al.[78] supports various mapping and quantification tools, and includes modules for filtering low-quality cells. Scater provides an organized workflow for converting raw sequencing reads into a ‘single-cell expression set’ (SCESet) class, a data structure that facilitates data handling and analysis[79]. Other available pipelines are either protocol specific (e.g., zUMI[80], scPIPE[81] and SEQC[82] for UMI data) or technology specific (e.g., Cell Ranger for Chromium systems). The scRNA-tools database (http://www.scRNA-tools.org) provides a comprehensive list of available computational tools for data processing and analysis[83]. Methods are categorized by analysis task, and researchers can select tools according to the required analysis type.

Normalization. Single-cell RNA-seq datasets show high levels of noise and variability related to nonbiological technical effects, including dropout events due to stochastic RNA loss during sample preparation, biased amplification and incomplete library sequencing. Technical variation also results from batch effects on processing units (e.g., plates or arrays), time points, facilities and other sources. Moreover, natural variability complicates analysis because of, for example, variable cell size and RNA content, different cell cycle stages and gender differences. Therefore, dataset normalization becomes an important step for meaningful data analysis. This can be guided by the addition of artificial spike-in RNA, which is used to model technical noise, as implemented in BASiCS[84]. However, it is not clear whether artificial RNA sufficiently reflects the behavior of endogenous RNA, or whether cellular RNA influences spike-in detection. Recent high-throughput methods distribute cells by limiting dilution, which makes the use of spike-in RNA impracticable because of the high number of otherwise empty reaction volumes. Alternative normalization methods originally developed for bulk RNA sequencing, such as log-expression[85], trimmed mean M-values[86] and upper-quartiles[87], can also be used in scRNA-seq, although more-specialized normalization methods are being developed that can better handle many aspects of this specific type of data. Recent single-cell approaches apply between-sample normalization (SCnorm[88]) or normalize on cell-based factors after pool-based size factor deconvolution (SCRAN[89]). However, for correction for large-scale sources of variation, a recommended and standard procedure is data modeling with the correct distribution. Here, confounding factors can be incorporated as covariates into the model and regressed out. Whereas batch effects are usually detected by visual inspection of reduced-space representations (e.g., principal components), kBET[90] is a batch-effect test based on *k* nearest neighbors. It quantitatively measures batch effects within and between datasets without directly correcting the data. This approach concludes that a combination of log normalization or SCRAN pooling with ComBat[91] or limma[92] regression provides the best batch-corrected dataset while preserving the biological structure. The batch effect problem is magnified when datasets from different time points, individuals or scRNA-seq methods are integrated. In this case, Haghverdi et al.[93] propose an approach based on mutual nearest neighbors in which a shared subset of populations is sufficient to correct for batch effects across experiments, although predefined or equal population compositions are required. Alternatively, by inferring cell clusters from gene expression similarities and coexpression patterns, Biscuit (Bayesian inference for single-cell clustering and imputing)[82] identifies and corrects for technical variation per cell. Also, the commonly used scRNA-seq package Seurat provides a solution for integrating datasets based on common sources of variation[94], with a new feature that allows the identification of shared populations and facilitates comparative analysis across datasets.

Imputation and gene selection. In addition to having a high noise level, scRNA-seq datasets are also very sparse, which poses further challenges to cellular phenotyping and data interpretation. Non-expressed genes and technical shortcomings, such as dropout events (unsequenced transcripts), result in many zeros in the expression matrix, and thus an incomplete description of a single cell's transcriptome. To reduce sparsity, missing transcript values can be computationally inferred by imputation, for example, with MAGIC[95], which uses diffusion maps to find data structures and restore missing information. Alternatively, scImpute[96] learns a gene's dropout probability by fitting a mixture model and then imputes probable dropout events by borrowing information from similar cells (selected on the basis of genes that are not severely affected).

A common strategy for determining heterogeneity in a sample is to analyze highly variable genes across datasets. A thorough feature-selection step to remove uninformative or noisy genes increases the signal-to-noise ratio but also reduces the computational complexity. Commonly used strategies for extracting variable genes in scRNA-seq tools exploit the relationship between the mean transcript abundance and a measure of dispersion such as the coefficient of variation[97], the dispersion parameter of the negative binomial distribution[98] or the proportion of total variability[84].

Data analysis

Some of the major applications of scRNA-seq experiments include assessment of sample heterogeneity and identification of novel cell types and states. This is achieved through determination of coexpression patterns and clustering of cells by similarity. Cell clusters can subsequently be interpreted through annotation of gene sets that drive clusters (marker genes). A common way to visually inspect cellular subpopulation structures is to carry out dimensionality reduction (DR) and project cells into a two- or three-dimensional space. Principal component analysis (PCA) and *t*-distributed stochastic neighbor embedding (t-SNE) are commonly used approaches for data representation[99,100]. Diffusion components[101] and uniform manifold approximation and projection (UMAP)[102] are viable alternatives that overcome some limitations of PCA and t-SNE by preserving the global structures and pseudo-temporal ordering of cells, as well as being faster[103]. Even though DR techniques can guide the initial data inspection, more-robust clustering algorithms are needed to define subpopulations among cells.

Although prior assumptions and canonical population markers allow supervised clustering (e.g., with Monocle2[104]), hypothesis-free unsupervised clustering is preferred in most cases. A commonly used unsupervised algorithm is hierarchical clustering, which provides consistent results without a predefined number of clusters. Hierarchical clustering can be conducted in an agglomerative (bottom-up) or divisive (top-down) manner, with consecutive merging or splitting of clusters, respectively. Tools such as PAGODA[105], SINCERA[106] and bigSCale[7] implement hierarchical clustering. Another suitable unsupervised clustering algorithm is *k*-means, which estimates *k* centroids (centers of the clusters), assigns cells to the nearest centroid, recomputes centroids on the basis of the mean of cells in the centroid clusters, and then reiterates these steps. SC3, for example, integrates both *k*-means and hierarchical clustering to provide accurate and robust clustering of cells[107]. Other unsupervised approaches, such as SNN-Cliq[108] and Seurat[94], use graph-based clustering, which builds graphs with nodes representing cells and edges indicating similar expression, and then partitions the graphs into interconnected 'quasi-cliques' or 'communities'. Clustering can be done directly on the basis of expression values or more processed data types, such as principal components or similarity matrices, the latter of which shows improved yield in cluster separation. Cluster stability is measured via resampling methods (e.g., bootstrapping) or on the basis of cell similarities within assigned clusters (e.g., silhouette index). To support cluster reproducibility, different algorithms can be compared using adjusted Rand indexes[107]. Clusters can be represented by color-coding in a low-dimensional space produced by the DR algorithms discussed above (e.g., PCA, t-SNE).

Marker genes that discriminate subpopulations can be identified by differential gene expression analysis of clusters using, for example, model-based approaches such as SCDE[109], MAST[110] and scDD[111], which account for data bimodality by using a mixture model. Individual genes can be evaluated to serve as binary classifiers for cell identity with, for example, ROC or LRT tests based on the zero-inflated data[94,107]. A recent publication comprehensively compared differential expression analysis methods for scRNA-seq and can be referred to as a guide for the selection of appropriate differential expression tools[112].

Another important application of scRNA-seq is trajectory inference, which estimates dynamic processes by ordering cells along a predicted differentiation path (pseudotime) using algorithms such as reversed graph embedding (Monocle2[113]) and minimum spanning tree (TSCAN[114]). Also, trajectory inference methods have been comprehensively benchmarked through tests of their accuracy and overall performance[115]. To further facilitate the interpretation of results, tools such as SCENIC[116] provide the opportunity to investigate active regulatory networks in subpopulations of cells. The analysis guides the identification of active transcription factors, eventually providing insights into the cellular mechanisms that drive heterogeneity. For cluster annotation, scmap facilitates comparison of data across experiments by projecting cells from one dataset onto cell types or individual cells from another scRNA-seq experiment[117]. With cell convolution tools such as bigSCale[7], scRNA-seq analysis can be expanded to millions of cells. Eventually, single cells can be mapped back to the spatial tissue context via experimental approaches[118,119] or pseudo-spatial ordering of cells[2,9,94].

To make scRNA-seq data publicly available, one can use data storage and sharing repositories. The Gene Expression Omnibus (GEO) (https://www.ncbi.nlm.nih.gov/geo/) is commonly used to provide access to raw data and more-processed formats, such as gene expression quantification matrices. Large-scale projects, such as the Human Cell Atlas, set up specific data coordination platforms to further ease data query and accessibility. For data analysis, many researchers provide free open access to their computational pipelines through public databases such as GitHub (https://github.com/) or offer ready-to-use packages through, for example, Bioconductor (https://www.bioconductor.org/).

Summary

Although it is challenging to define broadly applicable designs for scRNA-seq experiments, we here provide general guidelines to support the production of high-quality datasets and their meaningful interpretation. Thoroughly planned and conducted sample preparation is critical to preserve cellular and RNA integrity and allow unbiased representation of the sample composition. The selection of downstream scRNA-seq techniques is driven by the complexity of the underlying sample and the desired resolution per cell. Although large numbers of cells, processed in microfluidic systems, might better represent the composition of heterogeneous samples, an in-depth analysis of smaller samples could be more appropriate for resolving subtle differences in homogeneous mixtures. Budget restraints and reduced library complexity generally lead to the shallow sequencing of high numbers of cells, whereas cell-type-focused experiments with sensitive methods can benefit from deeper sequencing. Eventually, the analysis and interpretation of single-cell transcriptomes is enabled by a wealth of computational methods specifically tailored to answer biological questions in a hypothesis-free manner or guided by previous knowledge. Despite technical challenges, scRNA-seq experiments are a powerful tool that can be used to fully resolve sample heterogeneity and dynamic cellular systems or to identify perturbation effects at high resolution.

Future directions of the single-cell field

Single-cell transcriptomics technologies are advancing rapidly. Cell numbers that can be analyzed are increasing to hundreds of thousands of cells per experiment, markedly improving statistical power and resolution for detecting rare and transient cell types. However, high-throughput techniques come with the expense of decreased molecule capture rates, and future methods need to better balance cell numbers with cell resolution. This will be accompanied by decreased sequencing costs, eventually allowing comprehensive, high-resolution snapshots of complex tissues to be achieved. Today, tissue- and organism-level projects use ‘sky-dive’ experimental strategies, initially creating a low-resolution atlas with thousands of cells to estimate sample heterogeneity, and then zooming in on target cell types by means of efficient scRNA-seq methods to achieve higher per-cell resolution. In the future, high-resolution maps will allow users to zoom in on the existing data, circumventing costly and time-consuming sample reprocessing. Microfluidics methods have already driven a paradigm shift in experimental designs, and conceptually different alternative methods such as combinatorial barcoding[12,64] might push the barrier back even farther. Because they do not require physical separation of individual cells, these approaches allow for cost-effective parallel processing of cells, which will make it possible for cell numbers to be scaled up even further.

An additional future avenue of intense investigation will be based on advances in monitoring of transcriptional profiles in spatial contexts. scRNA-seq relies on disconnection of cells from their natural environment, but spatial methods, including in situ sequencing[120] and single-molecule (smFISH[118]) and multiplexed error-robust (MERFISH[119]) fluorescence in situ hybridization, profile gene expression in the tissue context. Although current methods have low transcriptome resolution or require prior marker selection, they are extremely powerful in resolving tissue complexity[9,121]. Future spatial methods should allow the field to advance from the current combinatory experimental designs[122], or pseudo-space analysis[2,94], to a full tissue expression profile in three dimensions. Eventually, phenotype heterogeneity and dynamics in living multicellular systems will be resolved by the fusion of unbiased transcriptome profiling in spatial and temporal dimensions with the combined profiling of additional layers of molecular information, such as genetic variation[123] and gene regulatory marks (e.g., DNA methylation[124] and open chromatin[125]), from the very same cell.

References

1. Regev, A. et al. The Human Cell Atlas. *eLife* **6**, e27041 (2017).
2. Ibarra-Soria, X. et al. Defining murine organogenesis at single-cell resolution reveals a role for the leukotriene pathway in regulating blood progenitor formation. *Nat. Cell Biol.* **20**, 127–134 (2018).
3. Grün, D. et al. De novo prediction of stem cell identity using single-cell transcriptome data. *Cell Stem Cell* **19**, 266–277 (2016).
4. Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C. & Teichmann, S. A. The technology and biology of single-cell RNA sequencing. *Mol. Cell* **58**, 610–620 (2015).
5. Bendall, S. C. et al. Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* **157**, 714–725 (2014).
6. Kelsey, G., Stegle, O. & Reik, W. Single-cell epigenomics: recording the past and predicting the future. *Science* **358**, 69–75 (2017).
7. Iacono, G. et al. bigSCale: an analytical framework for big-scale single-cell data. Preprint at *bioRxiv* https://doi.org/10.1101/197244 (2017).
8. *Tabula Muris* Consortium et al. Single-cell transcriptomics of 20 mouse organs creates a *Tabula Muris*. *Nature* **562**, 367–372 (2018).
9. Halpern, K. B. et al. Single-cell spatial reconstruction reveals global division of labour in the mammalian liver. *Nature* **542**, 352–356 (2017).

10. Haber, A. L. et al. A single-cell survey of the small intestinal epithelium. *Nature* **551**, 333–339 (2017).
11. Karaiskos, N. et al. The *Drosophila* embryo at single-cell transcriptome resolution. *Science* **358**, 194–199 (2017).
12. Cao, J. et al. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**, 661–667 (2017).
13. Fincher, C. T., Wurtzel, O., de Hoog, T., Kravarik, K. M. & Reddien, P. W. Cell type transcriptome atlas for the planarian *Schmidtea mediterranea*. *Science* **360**, eaaq1736 (2018).
14. Davie, K. et al. A single-cell transcriptome atlas of the aging *Drosophila* brain. *Cell* **174**, 982–998 (2018).
15. Han, X. et al. Mapping the mouse cell atlas by Microwell-seq. *Cell* **172**, 1091–1107 (2018).
16. Shahbazi, M. N. et al. Pluripotent state transitions coordinate morphogenesis in mouse and human embryos. *Nature* **552**, 239–243 (2017).
17. Enge, M. et al. Single-cell analysis of human pancreas reveals transcriptional signatures of aging and somatic mutation patterns. *Cell* **171**, 321–330 (2017).
18. Calon, A. et al. Stromal gene expression defines poor-prognosis subtypes in colorectal cancer. *Nat. Genet.* **47**, 320–329 (2015).
19. Tirosh, I. et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196 (2016).
20. Tirosh, I. et al. Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. *Nature* **539**, 309–313 (2016).
21. Puram, S. V. et al. Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell* **171**, 1611–1624 (2017).
22. Guillaumet-Adkins, A. et al. Single-cell transcriptome conservation in cryopreserved cells and tissues. *Genome Biol.* **18**, 45 (2017).
23. Alles, J. et al. Cell fixation and preservation for droplet-based single-cell transcriptomics. *BMC Biol.* **15**, 44 (2017).
24. Wang, W., Penland, L., Gokce, O., Croote, D. & Quake, S. R. High fidelity hypothermic preservation of primary tissues in organ transplant preservative for single cell transcriptome analysis. *BMC Genomics* **19**, 140 (2018).
25. Lacar, B. et al. Nuclear RNA-seq of single neurons reveals molecular signatures of activation. *Nat. Commun.* **7**, 11022 (2016).
26. Krishnaswami, S. R. et al. Using single nuclei for RNA-seq to capture the transcriptome of postmortem neurons. *Nat. Protoc.* **11**, 499–524 (2016).
27. Habib, N. et al. Div-Seq: single-nucleus RNA-seq reveals dynamics of rare adult newborn neurons. *Science* **353**, 925–928 (2016).
28. Habib, N. et al. Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nat. Methods* **14**, 955–958 (2017).
29. Bakken, T. E. et al. Equivalent high-resolution identification of neuronal cell types with single-nucleus and single-cell RNA sequencing. Preprint at *bioRxiv* https://doi.org/10.1101/239749 (2017).
30. Villani, A.-C. et al. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* **356**, eaah4573 (2017).
31. van den Brink, S. C. et al. Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nat. Methods* **14**, 935–936 (2017).
32. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: single-cell RNA-seq by multiplexed linear amplification. *Cell Rep.* **2**, 666–673 (2012).
33. Paul, F. et al. Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell* **163**, 1663–1677 (2015).
34. Prakadan, S. M., Shalek, A. K. & Weitz, D. A. Scaling by shrinking: empowering single-cell 'omics' with microfluidic devices. *Nat. Rev. Genet.* **18**, 345–361 (2017).
35. Stoeckius, M. et al. Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017).
36. Barriga, F. M. et al. Mex3a marks a slowly dividing subpopulation of Lgr5^{+} intestinal stem cells. *Cell Stem Cell* **20**, 801–816 (2017).
37. Ziegenhain, C. et al. Comparative analysis of single-cell RNA sequencing methods. *Mol. Cell* **65**, 631–643 (2017).
38. Svensson, V. et al. Power analysis of single-cell RNA sequencing experiments. *Nat. Methods* **14**, 381–387 (2017).
39. Avital, G. et al. scDual-Seq: mapping the gene regulatory program of *Salmonella* infection by host and pathogen single-cell RNA sequencing. *Genome Biol.* **18**, 200 (2017).
40. Hayashi, T. et al. Single-cell full-length total RNA sequencing uncovers dynamics of recursive splicing and enhancer RNAs. *Nat. Commun.* **9**, 619 (2018).
41. Faridani, O. R. et al. Single-cell sequencing of the small-RNA transcriptome. *Nat. Biotechnol.* **34**, 1264–1266 (2016).
42. Islam, S. et al. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* **11**, 163–166 (2014).
43. Giustacchini, A. et al. Single-cell transcriptomics uncovers distinct molecular signatures of stem cells in chronic myeloid leukemia. *Nat. Med.* **23**, 692–702 (2017).
44. Stubbington, M. J. T. et al. T cell fate and clonality inference from single-cell transcriptomes. *Nat. Methods* **13**, 329–332 (2016).
45. Karlsson, K. & Linnarsson, S. Single-cell mRNA isoform diversity in the mouse brain. *BMC Genomics* **18**, 126 (2017).
46. Picelli, S. et al. Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).
47. Ramsköld, D. et al. Full-length mRNA-seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* **30**, 777–782 (2012).
48. Islam, S. et al. Highly multiplexed and strand-specific single-cell RNA 5′ end sequencing. *Nat. Protoc.* **7**, 813–828 (2012).
49. Soumillon, M., Cacchiarelli, D., Semrau, S., van Oudenaarden, A. & Mikkelsen, T. S. Characterization of directed differentiation by high-throughput single-cell RNA-Seq. Preprint at *bioRxiv* https://doi.org/10.1101/003236 (2014).
50. Bagnoli, J. W. et al. Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq. Preprint at *bioRxiv* https://doi.org/10.1101/188367 (2017).
51. Sasagawa, Y. et al. Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. *Genome Biol.* **14**, R31 (2013).
52. Sasagawa, Y. et al. Quartz-Seq2: a high-throughput single-cell RNA sequencing method that effectively uses limited sequence reads. *Genome Biol.* **19**, 29 (2018).
53. Jaitin, D. A. et al. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776–779 (2014).
54. Hashimshony, T. et al. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.* **17**, 77 (2016).
55. Wu, A. R. et al. Quantitative assessment of single-cell RNA sequencing methods. *Nat. Methods* **11**, 41–46 (2014).
56. Streets, A. M. et al. Microfluidic single-cell whole-transcriptome sequencing. *Proc. Natl Acad. Sci. USA* **111**, 7048–7053 (2014).
57. Buenrostro, J. D. et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).
58. Hochgerner, H. et al. STRT-seq-2i: dual-index 5′ single cell and nucleus RNA-seq on an addressable microwell array. *Sci. Rep.* **7**, 16327 (2017).
59. Gierahn, T. M. et al. Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nat. Methods* **14**, 395–398 (2017).
60. Klein, A. M. et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).

61. Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
62. Zilionis, R. et al. Single-cell barcoding and sequencing using droplet microfluidics. *Nat. Protoc.* **12**, 44–73 (2017).
63. Zheng, G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
64. Rosenberg, A. B. et al. Scaling single cell transcriptomics through split pool barcoding. Preprint at *bioRxiv* https://doi.org/10.1101/105163 (2017).
65. Cusanovich, D. A. et al. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910–914 (2015).
66. Ramani, V. et al. Massively multiplex single-cell Hi-C. *Nat. Methods* **14**, 263–266 (2017).
67. Mulqueen, R. M. et al. Scalable and efficient single-cell DNA methylation sequencing by combinatorial indexing. Preprint at *bioRxiv* https://doi.org/10.1101/157230 (2017).
68. Kang, H. M. et al. Multiplexed droplet single-cell RNA sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018).
69. Picelli, S. et al. Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res.* **24**, 2033–2040 (2014).
70. Mora-Castilla, S. et al. Miniaturization technologies for efficient single-cell library preparation for next-generation sequencing. *J. Lab Autom.* **21**, 557–567 (2016).
71. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
72. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
73. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
74. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
75. Trapnell, C. et al. Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
76. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
77. Smith, T., Heger, A. & Sudbery, I. UMI-tools: modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. *Genome Res.* **27**, 491–499 (2017).
78. Ilicic, T. et al. Classification of low quality cells from single-cell RNA-seq data. *Genome Biol.* **17**, 29 (2016).
79. McCarthy, D. J., Campbell, K. R., Lun, A. T. L. & Wills, Q. F. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* **33**, 1179–1186 (2017).
80. Parekh, S., Ziegenhain, C., Vieth, B., Enard, W. & Hellmann, I. zUMIs—a fast and flexible pipeline to process RNA sequencing data with UMIs. *Gigascience* **7**, giy059 (2018).
81. Tian, L. et al. scPipe: a flexible R/Bioconductor preprocessing pipeline for single-cell RNA sequencing data. *PLoS Comput. Biol.* **10**, e1006361 (2018).
82. Azizi, E. et al. Single-cell map of diverse immune phenotypes in the breast tumor microenvironment. Preprint at *bioRxiv* https://doi.org/10.1101/221994 (2018).
83. Zappia, L., Phipson, B. & Oshlack, A. Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. *PLoS Comput. Biol.* **14**, e1006245 (2018).
84. Vallejos, C. A., Marioni, J. C. & Richardson, S. BASiCS: Bayesian analysis of single-cell sequencing data. *PLoS Comput. Biol.* **11**, e1004333 (2015).
85. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).
86. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).
87. Bullard, J. H., Purdom, E., Hansen, K. D. & Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* **11**, 94 (2010).
88. Bacher, R. et al. SCnorm: robust normalization of single-cell RNA-seq data. *Nat. Methods* **14**, 584–586 (2017).
89. Lun, A. T., Bach, K. & Marioni, J. C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.* **17**, 75 (2016).
90. Buttner, M., Miao, Z., Wolf, A., Teichmann, S. A. & Theis, F. J. Assessment of batch-correction methods for scRNA-seq data with a new test metric. Preprint at *bioRxiv* https://doi.org/10.1101/200345 (2017).
91. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
92. Ritchie, M. E. et al. limma powers differential expression analyses for RNA sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
93. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**, 421–427 (2018).
94. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
95. van Dijk, D. et al. Recovering gene interactions from single-cell data using data diffusion. *Cell* **174**, 716–729 (2018).
96. Li, W. V. & Li, J. J. scImpute: an accurate and robust imputation method for single-cell RNA-seq data. Preprint at *bioRxiv* https://doi.org/10.1101/141598 (2017).
97. Brennecke, P. et al. Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* **10**, 1093–1095 (2013).
98. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* **40**, 4288–4297 (2012).
99. Abdi, H. & Williams, L. J. Principal component analysis. *Wiley Interdiscip. Rev. Comput. Stat.* **2**, 433–459 (2010).
100. van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
101. Haghverdi, L., Buettner, F. & Theis, F. J. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics* **31**, 2989–2998 (2015).
102. McInnes, L. & Healy, J. UMAP: uniform manifold approximation and projection for dimension reduction. Preprint at https://arxiv.org/abs/1802.03426 (2018).
103. Becht, E. et al. Evaluation of UMAP as an alternative to t-SNE for single-cell data. Preprint at *bioRxiv* https://doi.org/10.1101/298430 (2018).
104. Qiu, X. et al. Single-cell mRNA quantification and differential analysis with Census. *Nat. Methods* **14**, 309–315 (2017).
105. Fan, J. et al. Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nat. Methods* **13**, 241–244 (2016).
106. Guo, M., Wang, H., Potter, S. S., Whitsett, J. A. & Xu, Y. SINCERA: a pipeline for single-cell RNA-seq profiling analysis. *PLoS Comput. Biol.* **11**, e1004575 (2015).

107. Kiselev, V. Y. et al. SC3: consensus clustering of single-cell RNA-seq data. *Nat. Methods* **14**, 483–486 (2017).
108. Xu, C. & Su, Z. Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics* **31**, 1974–1980 (2015).
109. Kharchenko, P. V., Silberstein, L. & Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nat. Methods* **11**, 740–742 (2014).
110. Finak, G. et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16**, 278 (2015).
111. Korthauer, K. D. et al. A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biol.* **17**, 222 (2016).
112. Soneson, C. & Robinson, M. D. Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods* **15**, 255–261 (2018).
113. Qiu, X. et al. Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **14**, 979–982 (2017).
114. Ji, Z. & Ji, H. TSCAN: pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Res.* **44**, e117 (2016).
115. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods: towards more accurate and robust tools. Preprint at *bioRxiv* https://doi.org/10.1101/276907 (2018).
116. Aibar, S. et al. SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).
117. Kiselev, V. Y., Yiu, A. & Hemberg, M. scmap: projection of single-cell RNA-seq data across datasets. *Nat. Methods* **15**, 359–362 (2018).
118. Ji, N. & van Oudenaarden, A. Single molecule fluorescent in situ hybridization (smFISH) of *C. elegans* worms and embryos. *WormBook* http://www.wormbook.org/chapters/www_smFISH/smFISH.html (2012).
119. Moffitt, J. R. et al. High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proc. Natl Acad. Sci. USA* **113**, 11046–11051 (2016).
120. Ke, R., Mignardi, M., Hauling, T. & Nilsson, M. Fourth generation of next-generation sequencing technologies: promise and consequences. *Hum. Mutat.* **37**, 1363–1367 (2016).
121. Ke, R. et al. In situ sequencing for RNA analysis in preserved tissue and cells. *Nat. Methods* **10**, 857–860 (2013).
122. Lein, E., Borm, L. E. & Linnarsson, S. The promise of spatial transcriptomics for neuroscience in the era of molecular cell typing. *Science* **358**, 64–69 (2017).
123. Macaulay, I. C. et al. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat. Methods* **12**, 519–522 (2015).
124. Angermueller, C. et al. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat. Methods* **13**, 229–232 (2016).
125. Clark, S. J. et al. scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat. Commun.* **9**, 781 (2018).
126. Liu, W. et al. Sample preparation method for isolation of single-cell types from mouse liver for proteomic studies. *Proteomics* **11**, 3556–3564 (2011).
127. Dorrell, C. et al. Surface markers for the murine oval cell response. *Hepatology* **48**, 1282–1291 (2008).
128. Su, X. et al. Single-cell RNA-seq analysis reveals dynamic trajectories during mouse liver development. *BMC Genomics* **18**, 946 (2017).
129. Treutlein, B. et al. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509**, 371–375 (2014).
130. Chapman, H. A. et al. Integrin $\alpha_6\beta_4$ identifies an adult distal lung epithelial population with regenerative potential in mice. *J. Clin. Invest.* **121**, 2855–2862 (2011).
131. Xu, Y. et al. Single-cell RNA sequencing identifies diverse roles of epithelial cells in idiopathic pulmonary fibrosis. *JCI Insight* **1**, e90558 (2016).
132. Joost, S. et al. Single-cell transcriptomics reveals that differentiation and spatial signatures shape epidermal and hair follicle heterogeneity. *Cell Syst.* **3**, 221–237 (2016).
133. Der, E. et al. Single cell RNA sequencing to dissect the molecular heterogeneity in lupus nephritis. *JCI Insight* **2**, e93009 (2017).
134. Autengruber, A., Gereke, M., Hansen, G., Hennig, C. & Bruder, D. Impact of enzymatic tissue disintegration on the level of surface molecule expression and immune cell function. *Eur. J. Microbiol. Immunol. (Bp.)* **2**, 112–120 (2012).
135. Grün, D. et al. Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* **525**, 251–255 (2015).
136. Glass, L. L. et al. Single-cell RNA sequencing reveals a distinct population of proglucagon-expressing cells specific to the mouse upper small intestine. *Mol. Metab.* **6**, 1296–1303 (2017).
137. Herring, C. A. et al. Unsupervised trajectory analysis of single-cell RNA-seq and imaging data reveals alternative tuft cell origins in the gut. *Cell Syst.* **6**, 37–51 (2018).
138. Merlos-Suárez, A. et al. The intestinal stem cell signature identifies colorectal cancer stem cells and predicts disease relapse. *Cell Stem Cell* **8**, 511–524 (2011).
139. Wollny, D. et al. Single-cell analysis uncovers clonal acinar cell heterogeneity in the adult pancreas. *Dev. Cell* **39**, 289–301 (2016).
140. Baron, M. et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. *Cell Syst.* **3**, 346–360 (2016).
141. Segerstolpe, Å. et al. Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell Metab.* **24**, 593–607 (2016).
142. Petersen, M. B. K. et al. Single-cell gene expression analysis of a human ESC model of pancreatic endocrine development reveals different paths to β-cell differentiation. *Stem. Cell Rep.* **9**, 1246–1261 (2017).
143. Muraro, M. J. et al. A single-cell transcriptome atlas of the human pancreas. *Cell Syst.* **3**, 385–394 (2016).
144. Li, D. et al. Complete disassociation of adult pancreas into viable single cells through cold trypsin-EDTA digestion. *J. Zhejiang Univ. Sci. B* **14**, 596–603 (2013).
145. Shekhar, K. et al. Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. *Cell* **166**, 1308–1323 (2016).
146. Daniszewski, M. et al. Single cell RNA sequencing of stem cell-derived retinal ganglion cells. *Sci. Data* **5**, 180013 (2018).
147. Picelli, S. et al. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10**, 1096–1098 (2013).
148. Carninci, P. et al. Thermostabilization and thermoactivation of thermolabile enzymes by trehalose and its application for the synthesis of full length cDNA. *Proc. Natl Acad. Sci. USA* **95**, 520–524 (1998).
149. Spiess, A.-N. & Ivell, R. A highly efficient method for long-chain cDNA synthesis using trehalose and betaine. *Anal. Biochem.* **301**, 168–174 (2002).
150. Pinto, F. L. & Lindblad, P. A guide for in-house design of template-switch-based 5′ rapid amplification of cDNA ends systems. *Anal. Biochem.* **397**, 227–232 (2010).
151. Lambert, D. & Draper, D. E. Effects of osmolytes on RNA secondary and tertiary structure stabilities and RNA–Mg^{2+} interactions. *J. Mol. Biol.* **370**, 993–1005 (2007).

152. Tang, F. et al. mRNA-seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).
153. Zajac, P., Islam, S., Hochgerner, H., Lönnerberg, P. & Linnarsson, S. Base preferences in non-templated nucleotide incorporation by MMLV-derived reverse transcriptases. *PLoS One* **8**, e85270 (2013).

Acknowledgements

H.H. is a Miguel Servet (CP14/00229) researcher funded by the Spanish Institute of Health Carlos III (ISCIII). This work received funding from the European Union's Horizon 2020 research and innovation program under Marie Skłodowska-Curie grant agreement no. H2020-MSCA-ITN-2015-675752 (SINGEK; A.L.) and the Ministerio de Ciencia, Innovación y Universidades (SAF2017-89109-P; AEI/FEDER, UE; H.H.). This project has been made possible in part by grant no. 2018-182827 (H.H.) from the Chan Zuckerberg Initiative DAF, an advised fund of the Silicon Valley Community Foundation. We thank ThePaperMill for critical reading and scientific editing services. Core funding was provided by the ISCIII and the Generalitat de Catalunya.

Author contributions

The authors contributed to the various sections of this tutorial as follows: A.L., Data processing and Data analysis; C.M., Sample preparation; S.P., Optimization (Box 1); H.H., Design, Sample preparation, Single-cell RNA sequencing, Further technical considerations and Future directions. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to H.H.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Published online: 16 November 2018