Annual Review of Biomedical Data Science

From Tissues to Cell Types and Back: Single-Cell Gene Expression Analysis of Tissue Architecture

Xi Chen,[1] Sarah A. Teichmann,[1,2,3] and Kerstin B. Meyer[1]

[1]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton CB10 1SA, United Kingdom; email: km16@sanger.ac.uk

[2]European Molecular Biology Laboratory (EMBL)–European Bioinformatics Institute (EBI), Wellcome Genome Campus, Hinxton CB10 1SD, United Kingdom

[3]Theory of Condensed Matter Research Group, Cavendish Laboratory, University of Cambridge, Cambridge CB3 0HE, United Kingdom

Annu. Rev. Biomed. Data Sci. 2018. 1:29–51

First published as a Review in Advance on May 23, 2018

The *Annual Review of Biomedical Data Science* is online at biodatasci.annualreviews.org

https://doi.org/10.1146/annurev-biodatasci-080917-013452

Keywords

single-cell genomics, scRNA-seq, tissue architecture, deconvolution, spatial gene expression

Abstract

With the recent transformative developments in single-cell genomics and, in particular, single-cell gene expression analysis, it is now possible to study tissues at the single-cell level, rather than having to rely on data from bulk measurements. Here we review the rapid developments in single-cell RNA sequencing (scRNA-seq) protocols that have the potential for unbiased identification and profiling of all cell types within a tissue or organism. In addition, novel approaches for spatial profiling of gene expression allow us to map individual cells and cell types back into the three-dimensional context of organs. The combination of in-depth single-cell and spatial gene expression data will reveal tissue architecture in unprecedented detail, generating a wealth of biological knowledge and a better understanding of many diseases.

INTRODUCTION

All cells in the human body carry the same genetic information yet can differentiate into a vast array of different cell types and tissues. Modern genomic methods make it possible to identify tissue-specific patterns of gene expression, and this has led to key insights into the molecular mechanisms underlying the function of different tissues and organs. However, tissues are complex compositions of cooperating individual cells, which are the basic functional unit of biology and of gene regulation. To fully understand an organ's function, researchers need to identify all the different cell types that make up this tissue and to examine how individual cells or cell type composition can change dynamically over time or in response to external stimuli.

Recently, projects such as GTEx (Genotype-Tissue Expression; **https://www.gtexportal.org**) (1) and Illumina Human Body Map 2.0 have generated large-scale gene expression data sets from different organs. A key challenge now is to harness these data sets for a better understanding of tissue function. Here we review how tissue gene expression data sets can shed light on the presence, abundance, and function of distinct cell types; how individual cells can now be profiled; and how these data can be placed back into the organ context to understand function (**Figure 1**).

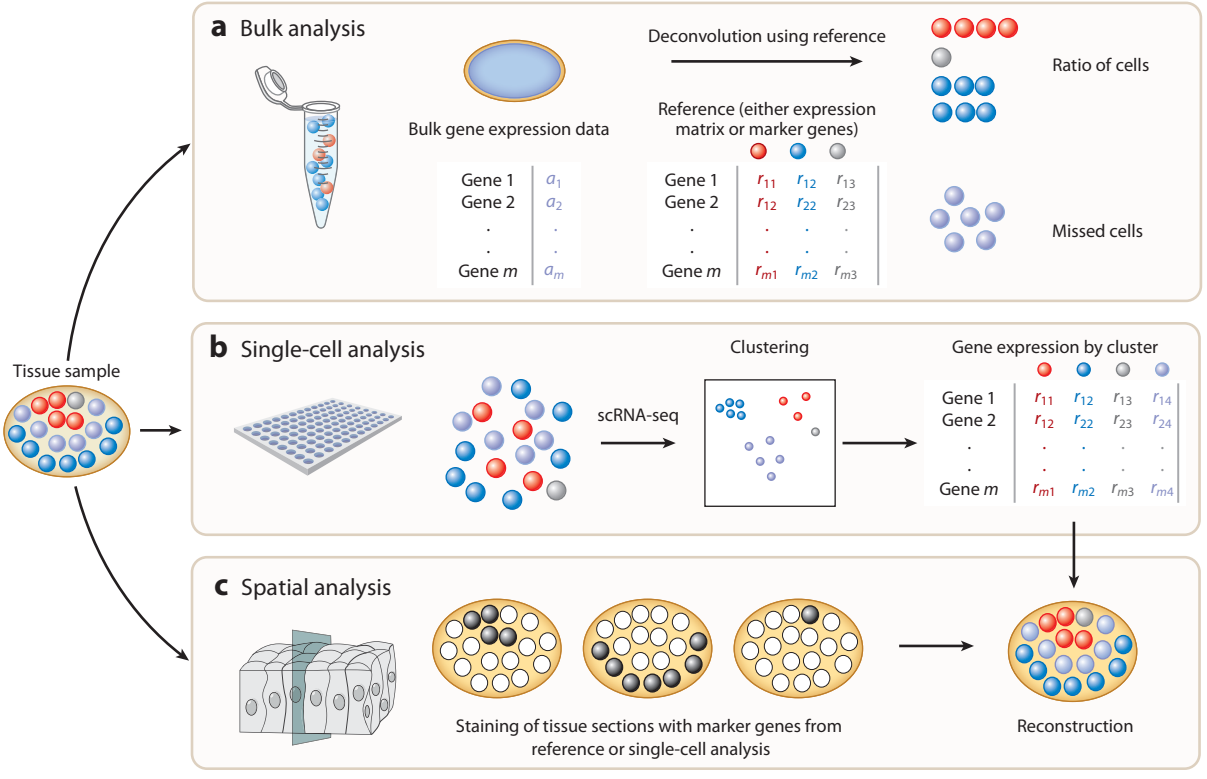


Figure 1

Overview of approaches to identify cell types present in tissues and place them back into their tissue context. (*a*) Bulk analysis obtains information from homogenized tissue and relates this back to known reference cell types. (*b*) In single-cell analysis, tissues are first dissociated to a single-cell suspension and then profiled individually by single-cell RNA sequencing (scRNA-seq). Cell types, including previously unknown cell types, can be identified but no positional information is retained. (*c*) In spatial analysis, sections are cut from tissue blocks and the position of cell types within the tissue can be identified using marker genes. This technique generates two-dimensional (2D) information but often for only a few genes. Once spatial data are available, scRNA-seq data can be mapped back onto the tissue and it is possible to reconstruct a 2D image of the tissue.

STUDYING TISSUE ARCHITECTURE AT THE SINGLE-CELL LEVEL

Early single-cell research used quantitative polymerase chain reaction (PCR), fluorescence-activated cell sorting (FACS), and immunofluorescence or fluorescence in situ hybridization (FISH) to investigate a limited number of genes or proteins, but technical improvements have allowed significantly more genes or proteins to be investigated. Mass cytometry [or cytometry by time-of-flight (CyTOF)] now allows up to 40 proteins to be assayed in parallel using metal-conjugated antibodies (2). These techniques have contributed to the delineation of cell types and subtypes, but they rely on a panel of preselected genes or proteins. This may bias the study and requires prior knowledge of, for example, the identity of key cell surface markers, which is not always available.

In contrast, single-cell genomics offers the chance to profile tissues at the single-cell level in an unbiased way. It is now possible to generate successful sequencing libraries from RNA (3) and DNA (4) from a subnanogram of material. In 2009, Tang et al. (3) presented the first single-cell transcriptome based on previous improvements in complementary DNA (cDNA) amplification (5, 6). Subsequently, full-length cDNA libraries from several mouse oocytes and a single blastomere were generated and sequenced (3). Two years later, Navin et al. (4) generated the first single-cell genomes from breast cancer samples. It was these pioneering studies that opened the gateway to the new field of single-cell genomics. Over the past decade, rapid experimental and computational developments together with massive improvements in sensitivity and throughput have generated fundamental new insights into many biological systems.

CELL-TYPE DECONVOLUTION

We should not forget that although single-cell technology is transforming the way we study tissues, there is already a very large amount of bulk data (obtained from a complex mixture of cells) that can be mined for more facts about tissue composition. Furthermore, bulk tissue or biopsy samples remain the primary source of material for biomedical research. The abundance of bulk data has led to the development of computational techniques that infer cell type composition by assuming that gene expression patterns in complex tissues are a linear combination of those in different cell types (7). Different studies have used regression and nonnegative matrix factorization on bulk data to identify distinct cell types in colon cancer samples (8), reveal cell cycle dynamics in yeast (9), and estimate the relative proportions of predefined cell types in the mouse mammary gland (10).

Cell type–specific significance analysis of microarrays (11) and population-specific expression analysis, a tool for RNA sequencing (RNA-seq) data, have been developed for deconvolution but are limited to a relatively small number of cell types. Altboum et al. (12) improved on this number, generating a reference compendium of over 200 immune cells. They also applied digital cell quantification to reveal the dynamics of up to 70 immune cell types in influenza infection (12). CIBERSORT (13), a widely adopted algorithm, has improved methods further by using a machine learning approach (support vector regression) to reveal, for example, a complex relationship between tumor-infiltrating immune cells and cancer survival (14). For a detailed survey of available deconvolution tools, readers are referred to References 7 and 15. Although these computational tools are very valuable, they require prior knowledge and do not identify new cell types. Single-cell RNA sequencing (scRNA-seq) can overcome both of these limitations.

IMPACT OF RECENT DEVELOPMENTS IN scRNA-seq

In the field of single-cell genomics, scRNA-seq is the most rapidly evolving technique and has proved to be a powerful tool in the study of complex tissue. **Table 1** highlights recent insights

Table 1 Recent examples of atlas-type single-cell studies of tissues

Reference(s)	Tissue	Number of cells	Method	Key results
104	Mouse spleen	>4,000	MARS-seq	1. Developed MARS-seq 2. Demonstrated the feasibility of using scRNA-seq to unbiasedly identify different cell types and subtypes in a tissue
22	Mouse retina	>44,000	Drop-seq	1. Developed Drop-seq 2. Successfully uncovered 39 cell populations in the retina 3. Determined that large cell number is important for clear distinctions among these cell types
105	Mouse somatosensory cortex and hippocampus	>3,000	STRT-seq (Fluidigm C1)	1. Successfully identified 47 distinct cell types in the cortex 2. Developed a biclustering method for cell and gene clustering
106	Human pancreas	>3,000	CEL-seq2	1. Provided cell type–specific markers for different pancreatic cells 2. Identified CD24 and TM4SF4 as reliable markers to purify alpha and beta cells
107, 108	Human pancreas	>2,000, >1,000	SMART-seq2, Fluidigm C1	1. Successfully identified various cell types and novel subtypes in the human pancreas 2. Revealed key genes that have altered expression in type 2 diabetes
27	*Caenorhabditis elegans* (L2 stage)	~50,000	sci-RNA-seq	1. Developed sci-RNA-seq 2. Was the first single-cell atlas for a whole organism 3. Was able to uncover rare cell types for as few as one cell in the worm
28	Mouse brain	>100,000	SPLiT-seq	1. Developed SPLiT-seq 2. Provided a snapshot of a whole brain in a developmental stage (postnatal day 5)
109	Human blood	>2,000	SMART-seq2	1. Unbiasedly identified six dendritic cell subtypes and four monocyte subtypes 2. Discovered a new dendritic cell type that potently activates T cells
24	Human blood	>68,000	Chromium	1. Developed the Chromium system 2. Provided a commercial platform for high-throughput droplet-based scRNA-seq
110	20 different organs and tissues from mouse	~100,000	Chromium and SMART-seq2	1. Provided an expression reference of cell types from the mouse 2. Built a useful resource that can serve as a foundation for future mouse studies
111	51 different organs and tissues from mouse	>400,000	Microwell-seq	1. Developed Microwell-seq 2. Provided an initial single-cell expression atlas for the mouse

SINGLE-CELL ISOLATION METHODS

The first step of most scRNA-seq protocols is the isolation of single cells, and this step is the primary determinant of the throughput of the method. Since the abundance of different cell types differs in tissues, it is critical to profile enough cells to capture infrequent or rare cells. A key aim in the technological development of single-cell experiments is to increase the number of cells that can be analyzed (59). Early proof-of-principle studies used low-throughput methods like manual picking and FACS to isolate single cells into plates or microfluidic chips (e.g., Fluidigm C1) to capture single cells in nanoliter chambers and subsequently generate sequencing libraries (3, 17, 18, 44, 65, 104, 112–117). These strategies can process hundreds of cells per experiment but can be laborious and error prone. To overcome this, many research groups have used robotics to automate procedures. Subsequently, droplet-based microfluidics (22–24) and nanowell-based technologies (25, 118–121) were developed to randomly capture single cells into isolated nanoliter compartments (droplets or nanowells), increasing the throughput to tens of thousands of cells while at the same time significantly reducing manual labor. scRNA-seq at this scale makes unbiased tissue profiling possible.

into tissue complexity and heterogeneity. Currently, scRNA-seq is the most sensitive and unbiased way to measure cell types and states. The process involves several important steps. First, single cells are isolated (see sidebar titled Single-Cell Isolation Methods), then sequencing libraries are prepared from these single cells. The basic principles used to produce and sequence single-cell libraries are the same as those used for bulk material. The issue for single-cell library preparation is the tiny amount of starting material. Single cells contain as little as 10–30 pg of RNA, and for lowly expressed genes, only a small number of transcripts will be present in the cell. When adapting sequencing protocols to single cells, this scarcity of input material is the key challenge and ingenious solutions have emerged to counter this.

scRNA-seq library construction methods use well-established chemistry but optimize many of the steps (e.g., removing many intermediate purification steps) to make them suitable for single cells. The study object of most scRNA-seq methods so far is polyadenylated messenger RNA (mRNA). Therefore, most scRNA-seq protocols start with capturing mRNA with oligo-dT coupled with either a sequence for PCR or a T7 promoter and synthesize the first strand of cDNA using a reverse transcriptase, with Moloney murine leukemia virus (MMLV) reverse transcriptase being the most widely used. The main differences between various scRNA-seq protocols are (*a*) how the second strand of cDNA is generated (see the sidebar titled Second Strand Generation) and (*b*) how sequencing libraries are constructed and amplified (see the sidebar titled Library Construction).

The miniscule amounts of starting material necessitate the use of multiple rounds of PCR, leading to potentially very large amplification biases. This can be overcome by using PCR or ligation to integrate a very diverse set of oligonucleotide barcodes called unique molecular identifiers (UMIs) into cDNA transcripts (see the sidebar titled Barcoding or Indexing). After PCR amplification, we can conclude that molecules with the same UMI were derived from the same original transcript, allowing the amplification bias to be removed computationally (16).

Additionally, scRNA-seq protocols are complex and every step results in a further loss of the already small amounts of RNA input material. This has been addressed by pooling samples (e.g., 17) or reducing/eliminating intermediate purification steps in the protocols—for example, by using a mixture of enzymes to catalyze several consecutive steps in the same reaction volume (e.g., 18).

Barcoding or indexing has proved critical when pooling material. Protocols mark all transcripts from the same cell with a common barcode, thus allowing all subsequent steps to be carried out in bulk (e.g., 17). Once reactions from single cells have been combined, the loss of material becomes

SECOND STRAND GENERATION

One method for synthesis of the second strand cDNA employs a terminal transferase to add a polyA tail at the 3′ end of the first strand cDNA. A polyT primer with a PCR sequence is then added, and cDNA is amplified by PCR. The method of Tang et al. (3), Quartz-seq (115), Quartz-seq2 (116), and methods used in some other early studies (5, 122–124) are all based on this method. Alternatively, the second strand synthesis can rely on the terminal transferase activity of MMLV. In the presence of Mg^{2+} or Mn^{2+}, cytosines are added to the 3′ end of the first strand cDNA (125, 126). By the addition of a template-switching oligonucleotide (TSO) with a PCR sequence and a stretch of guanines at its 3′ end, full-length cDNA can be synthesized by PCR. STRT-seq (112, 113), SMART-seq (18), SMART-seq2 (114), Drop-seq (22), Seq-Well (25), Chromium (10x Genomics; 24), and SPLiT-seq (28) use this approach. This TSO-based method provides a simple and, compared to the polyA tailing–based method, more faithful way of generating full-length cDNA. The third way of generating double-stranded cDNA is based on the combined activity of ribonuclease (RNase) H and DNA polymerase I from *Escherichia coli* (127). In this method, RNase H first cuts mRNA in the mRNA-DNA duplex. Then, the RNA-primed first strand cDNA is used as template, and second strand cDNA is synthesized by DNA polymerase I (128). CEL-seq (17), CEL-seq2 (117), MARS-seq (104), inDrop (23), and sci-RNA-seq (27) are all based on this straightforward approach.

LIBRARY CONSTRUCTION

Most scRNA-seq protocols use PCR-based methods for library amplification due to simplicity and speed. In contrast, in vitro transcription (IVT) achieves linear amplification of the library, resulting in less amplification bias but requiring more steps and time than PCR. CEL-seq (17), CEL-seq2 (117), and inDrop (23) use IVT for library amplification. Most scRNA-seq methods only sequence the 3′ end of a gene. So far, only Quartz-seq (115), SMART-seq (18), and SMART-seq2 (114) can sequence the full-length transcripts combining full-length cDNA synthesis with fragmentation or tagmentation. Of note, STRT-seq (112), STRT-seq-2i (119), Drop-seq (22), Chromium (10x Genomics; 24), Seq-Well (25), and SPLiT-seq (28) all perform full-length cDNA synthesis like SMART-seq and SMART-seq2, but STRT-seq and STRT-seq-2i only sequence the 5′ end of the transcripts, while the others focus on 3′ sequencing of the mRNA.

BARCODING OR INDEXING

Early studies prepared the cDNA and library individually for each cell, and cell barcodes were added either during second strand synthesis, for example in the STRT-seq method (112), or during the library PCR stage, for example in the modified STRT-seq (113), STRT-seq-2i (119), Quartz-seq (115), SMART-seq (18), and SMART-seq2 (114). Each single cell is converted into a separate library, and the cell barcodes are equivalent to sample barcodes. Most recent methods add cell barcodes during the reverse transcription stage as part of the oligo-dT primers to increase the experimental throughput. Cells can be pooled after reverse transcription, and downstream procedures can be performed in a single reaction. This has the benefit of reducing technical variation between different cells during library preparation. In addition, another level of barcode (sample barcode) can be added during the library amplification stage so that many samples can be multiplexed and sequenced together.

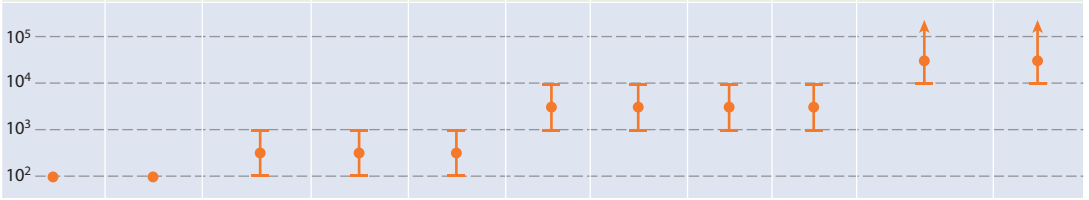
	SMART-seq2	CEL-seq2	STRT-seq	Quartz-seq2	MARS-seq	Drop-seq	inDrop	Chromium	Seq-Well	sci-RNA-seq	SPLiT-seq
Single-cell isolation	FACS, microfluidics	FACS, microfluidics	FACS, microfluidics, nanowells	FACS	FACS	Droplet	Droplet	Droplet	Nanowells	Not needed	Not needed
Second strand synthesis	TSO	RNase H and DNA pol I	TSO	PolyA tailing and primer ligation	RNase H and DNA pol I	TSO	RNase H and DNA pol I	TSO	TSO	RNase H and DNA pol I	TSO
Full-length cDNA synthesis?	Yes	No	Yes	Yes	No	Yes	No	Yes	Yes	No	Yes
Barcode addition	Library PCR with barcoded primers	Barcoded RT primers	Barcoded TSOs	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers and library PCR with barcoded primers	Ligation of barcoded RT primers
Pooling before library?	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Library amplification	PCR	In vitro transcription	PCR	PCR	In vitro transcription	PCR	In vitro transcription	PCR	PCR	PCR	PCR
Gene coverage	Full-length	3′	5′	3′	3′	3′	3′	3′	3′	3′	3′
Number of cells per assay											

Figure 2

A detailed technical comparison of popular single-cell RNA-seq protocols. For further technical details, readers are referred to the sidebars titled Second Strand Generation, Library Construction, and Barcoding or Indexing. Abbreviations: cDNA, complementary DNA; DNA pol I, DNA polymerase I; FACS, fluorescence-activated cell sorting; PCR, polymerase chain reaction; RNase H, ribonuclease H; RT, reverse transcription; TSO, template-switching oligonucleotide.

less problematic and sample handling much less labor intensive since it no longer needs to be carried out for each individual cell. Recently described scRNA-seq protocols have combined these different approaches to minimize the effect of small amounts of input material (summarized in **Figure 2**).

Another challenge in scRNA-seq is the sparsity of the data. Not all transcripts present in a single cell can be captured; therefore, the gene expression matrices in these cases contain many zeros that reflect a failure to capture relevant molecules rather than the absence of gene expression (19). There are complementary approaches to help overcome this that can be summarized as sequencing in greater depth (more genes, fewer cells) versus sequencing with higher throughput (more cells, fewer genes).

Greater-depth techniques, such as SMART-seq and SMART-seq2, can profile the full length of transcripts. As sequencing libraries contain multiple fragments from different parts of each expressed gene, the likelihood of capturing individual genes increases. An additional advantage of profiling full-length transcriptomes is the potential to identify multiple transcripts that arise from a single gene, i.e., different isoforms, and to sequence variable regions of genes that may not be located at the extreme 3′ or 5′ end. This is particularly relevant for genes such as the T cell and B cell receptor genes that are highly variable over a 300-bp region and that are unique to individual T and B lymphocytes, respectively (20, 21).

The higher-throughput approach deals with the problem of sparse data by profiling large numbers of cells and combining data from multiple cells. This has been made possible by the development of very-high-throughput techniques, such as droplet- (22–24) or microwell-based sequencing (25). In approaches such as Drop-seq (22), inDrop (23), and Chromium (24), microfluidics devices are used to generate droplets that contain a single cell together with a bead carrying barcodes that allow each transcript from that cell to be specifically labeled. In the Seq-Well method (25), droplets are replaced by nanowells that trap both cells and beads. Once first strand cDNA is complete, the droplet emulsion is broken or wells are combined to allow all subsequent steps to be carried out in bulk. This has made it possible to profile tens of thousands of single cells in parallel (**Figure 2**, bottom row). After sequencing and processing, similar cells (i.e., cells of the same type or state) can be pooled to generate pseudobulk data, significantly increasing the number of detected genes. Typically, scRNA-seq can detect 1,000–4,000 expressed genes per cell, while pseudobulk data detect more than 10,000 different genes, partially overcoming data sparsity.

The other tremendous advantage of these high-throughput techniques is the ability to detect rare cell types. Our experience with the computational analysis of such data has shown that a minimum of around 20–30 cells is required to identify new cell types or states. In a population of 10,000 cells, cell types that only represent 0.5% of the bulk population can nevertheless be identified reproducibly between individuals (26; R. Vento-Tormo, unpublished manuscript).

The recently described approach of split pool combinatorial indexing has enabled cheap single-cell profiling at a large scale without the need for single-cell isolation. This approach uses the cell itself as a reaction container, and reverse transcription (RT) is performed in situ. Cao et al. (27) developed single-cell combinatorial indexing (sci-RNA-seq), where a defined number of cells are distributed into either 96-well or 384-well plates. Each well contains a unique index that is added by using a barcoded RT primer during the RT stage. Then all cells are pooled, a limited number of cells are redistributed into wells in a new plate, and a second index is introduced. In this process, unique cells can be distinguished by different combinations of first and second indices after sequencing (27). Using this technique, 96×96 (9,216) or 384×384 (147,456) cells can be easily profiled with basic lab equipment at low cost. Similarly, SPLiT-seq (28) uses different combinations of barcoding oligonucleotides added by ligation. In other areas of genomics, the idea of combinatorial indexing has already been used successfully in many applications (e.g., haplotype-resolved genome sequencing; 29).

scRNA-seq DATA ANALYSIS

Once sequencing is complete, a series of computational steps needs to be carefully executed to convert raw sequencing reads to meaningful biological results. This process usually includes quantification of gene expression, quality control, batch correction, customized downstream analyses, and visualization. Gene expression quantification can be achieved by aligning sequencing reads to the corresponding genomes with a splice-aware aligner and counting the number of reads aligned to each gene (30–32). Transcript abundance can also be directly estimated from sequencing reads without actual alignment of individual bases, using programs such as kallisto (33) and Salmon (34).

Unavoidable technical variations introduced during sample processing and library preparation, known as batch effects, can severely hamper data interpretation and mask true biological signals. Batch correction is often necessary when integrating different single-cell experiments (35–37). Cell quality control criteria, such as the threshold number of reads or UMIs per cell, the mapping rate, the number of detected genes, etc., are often applied to filter out data from low-quality cells, empty wells, or empty droplets. Subsequently, customized analyses can be performed. These analyses usually include dimension reduction techniques to visualize the data, cell clustering based on

gene expression patterns, and differential gene expression analyses to find marker genes. Several computational toolkits, such as Cell Ranger (24), Monocle (38), Seurat (39), and Scanpy (40), have been designed to integrate different types of analysis into simple workflows for data interpretation and visualization. It should be noted that there are many computational methods available to determine differential gene expression from single-cell data. Their biases, robustness, and scalability have recently been compared by Soneson & Robinson (41).

HOW TO DECIDE WHICH SEQUENCING TECHNIQUE TO USE

From a practical point of view, deciding which scRNA-seq technology to use in research depends on various factors, such as project aims, budgets, and access to specialized equipment. If a high number of cells is important—for example, to detect rare cell types—droplet- or microwell-based technologies (such as Chromium and Seq-Well) may be preferred over FACS/plate-based methods (such as SMART-seq2). Custom or homemade equipment may be cheaper than its commercial counterparts but can be difficult to construct and often requires special engineering expertise. FACS/plate-based methods have relatively low throughput but are generally easier to implement with standard lab equipment. Methods such as SMART-seq2 offer much higher sequencing depth and detect more genes per cell (42, 43), allowing for detailed characterization of a limited number of cells.

In two recent studies, Svensson et al. (42) and Ziegenhain et al. (43) compared data generated from multiple popular scRNA-seq methods, analyzing their detection sensitivity (their ability to detect lowly expressed transcripts), quantification accuracy, and sequencing depth. All techniques tested had high accuracy but differed in sensitivity. The SMART-seq2 protocol had the best performance on the number of genes detected per cell. UMI-based protocols generated less amplification noise (43). Quantification accuracy only marginally depended on sequencing depth, but the detection limit critically relied on this (42). These two studies provide valuable benchmark information for the selection of the sequencing strategy most appropriate to the biological question under investigation.

INVESTIGATING DYNAMIC BIOLOGICAL PROCESSES USING scRNA-seq

Biological processes such as the immune response, cancer development, and embryogenesis are all highly dynamic, with different cells from the same tissue responding in different ways (e.g., cell fate changes) or at different rates (e.g., asynchronous response). One of the key advantages of single-cell over bulk RNA-seq analysis is the ability to capture these dynamic processes in an unbiased way, for example, during embryogenesis (44–47). Xue et al. (46) investigated the dynamics of gene expression in human and mouse embryonic development using scRNA-seq and identified the activity of distinct submodules that changed in a stepwise manner during the development from oocyte to morula. Comprehensive profiling of human preimplantation embryos by scRNA-seq have revealed the dynamic nature of lineage commitment and provided an atlas of early human development (47).

The temporal information (response time, developmental stage, etc.) is inherently retained in scRNA-seq data due to their single-cell resolution. The data capture a snapshot of many unsynchronized single cells from a tissue at different stages of development or in response to stimuli. It is then possible to reconstruct a pseudotime by computationally inferring the progress of single cells through a biological process (38). Trapnell et al. (38) developed an analysis technique called Monocle that is based on finding a minimal spanning tree that links individual cells based on their

similarity during a dynamic process. This allows the single cells to be placed on a trajectory or pseudotimeline (38). Monocle 2, a later development, uses reversed graph embedding for better inference of multiple developmental paths during a process (48). There are other methods for pseudotime inference, such as Wanderlust, which uses a *k*-nearest neighbor graph–based method (49), and Waterfall, which applies a minimum spanning tree on top of *k*-means clustering to infer developmental trajectories of neural stem cells (50).

Single-cell clustering using bifurcation analysis was developed to infer branched trajectories of different lineages derived from the same progenitor cell population (51). Similarly, Wishbone successfully pinpointed the branching point of mouse thymic T cell development, at which double-positive progenitor cells become either $CD8^+$ or $CD4^+$ T cells (52). Haghverdi et al. (53) used diffusion-like random walks to infer the timing of fate decision of red blood cells and endothelial-like cells. Gaussian processes are commonly used for nonparametric analysis of time series data and have been incorporated into several algorithms that infer branching points (54–57). Since 2014, more than 50 trajectory inference methods have been developed, and many of these have recently been compared for accuracy and robustness in a large benchmarking study (58). This analysis has identified useful benchmarking metrics and developed a decision tree to guide potential users to the most appropriate algorithm.

The explosion of techniques to reconstruct developmental trajectories from single-cell data highlights the power of scRNA-seq in demonstrating the continuous nature of many biological processes. It has also allowed researchers to define progenitor cell populations that may be important for tissue maintenance and regeneration.

TRENDS IN THE FIELD OF scRNA-seq

The invention of microfluidic and robotic devices to reduce manual labor and significantly increase the number of cells that can be analyzed per experiment (59), along with the reduction in cost and the development of computational methods to analyze data, has meant that single-cell genomic studies are now carried out routinely in many laboratories. That upsurge in research has shifted studies from proof-of-concept to work that aims to generate mechanistic insights and make new discoveries (see **Table 1**). scRNA-seq has even been combined with CRISPR/Cas9 screening technologies to investigate regulatory circuits in single cells (60–63).

The advent of very-high-throughput technologies (**Figure 2**) is also driving a shift in research. Benchmarking studies have demonstrated that low-coverage and low-depth expression data are able to discriminate different cell types (64, 65). In the future, high-throughput droplet- or nanowell-based methods may therefore be prioritized over sequencing depth for many studies interested in tissue deconvolution. Once subpopulations (either a subtype or a specific developmental stage) are identified, detailed characterization, using more targeted and sensitive methods like SMART-seq2, combined with FACS or CyTOF, can specifically investigate subpopulations.

Another trend relates to the amount of single-cell data being generated. More efficient computational methods have been developed to reduce the time needed to identify patterns from the high-dimensional single-cell data. In fact, the computational side of single-cell genomics is probably the fastest evolving area of the field. In 2015, Stegle et al. (19) listed only a few available tools for scRNA-seq analysis. Now, just three years later, dozens of computational tools and packages are listed on the GitHub page "Awesome Single Cell" (**https://github.com/seandavi/awesome-single-cell**), many of which utilize machine learning techniques. We anticipate that more and more computational tools will become available not only for scRNA-seq but also for single-cell genome and epigenome analysis.

SPATIAL TECHNIQUES FOR GENE EXPRESSION PROFILING OF SINGLE CELLS

Despite the efficacy with which scRNA-seq can robustly and sensitively identify distinct cell types and cell states from complex tissues, scRNA-seq requires tissue dissociation and so is unable to retain information about the position of cells in the tissue context. This spatial information captures the local microenvironment that determines how the cell functions, which cell types it may differentiate into, or what cell state it may be in. Techniques for measuring gene expression within the tissue context range from microdissection, in situ RNA hybridization, and immunohistochemistry to in situ sequencing of genes, all of which have greatly increased in scale over recent years. These techniques are frequently combined with computational rendering of the three-dimensional (3D) space to obtain final results.

Tissue Dissection

Gene expression studies that retain spatial information have to balance the analysis of large numbers of cells for only a few markers against the much more in-depth profiling of a more limited number of cells. Two dissection methods have become popular. The first uses laser capture microdissection (LCM) to collect specific cell types that are then profiled by RT-PCR or bulk sequencing or combined with single-cell sequencing (Geo-seq) (66). This has allowed spatially resolved transcriptomes to be generated for the early mouse embryo, brain, and other tissues (67, 68). The drawback with LCM is that it is very labor intensive, severely limiting the number of samples and cells that can be analyzed. An alternative method serially slices an anatomical structure and profiles the slices using RNA-seq or microarrays. By taking slices in three planes, researchers can reconstruct a 3D image of gene expression as demonstrated for the mouse brain (69).

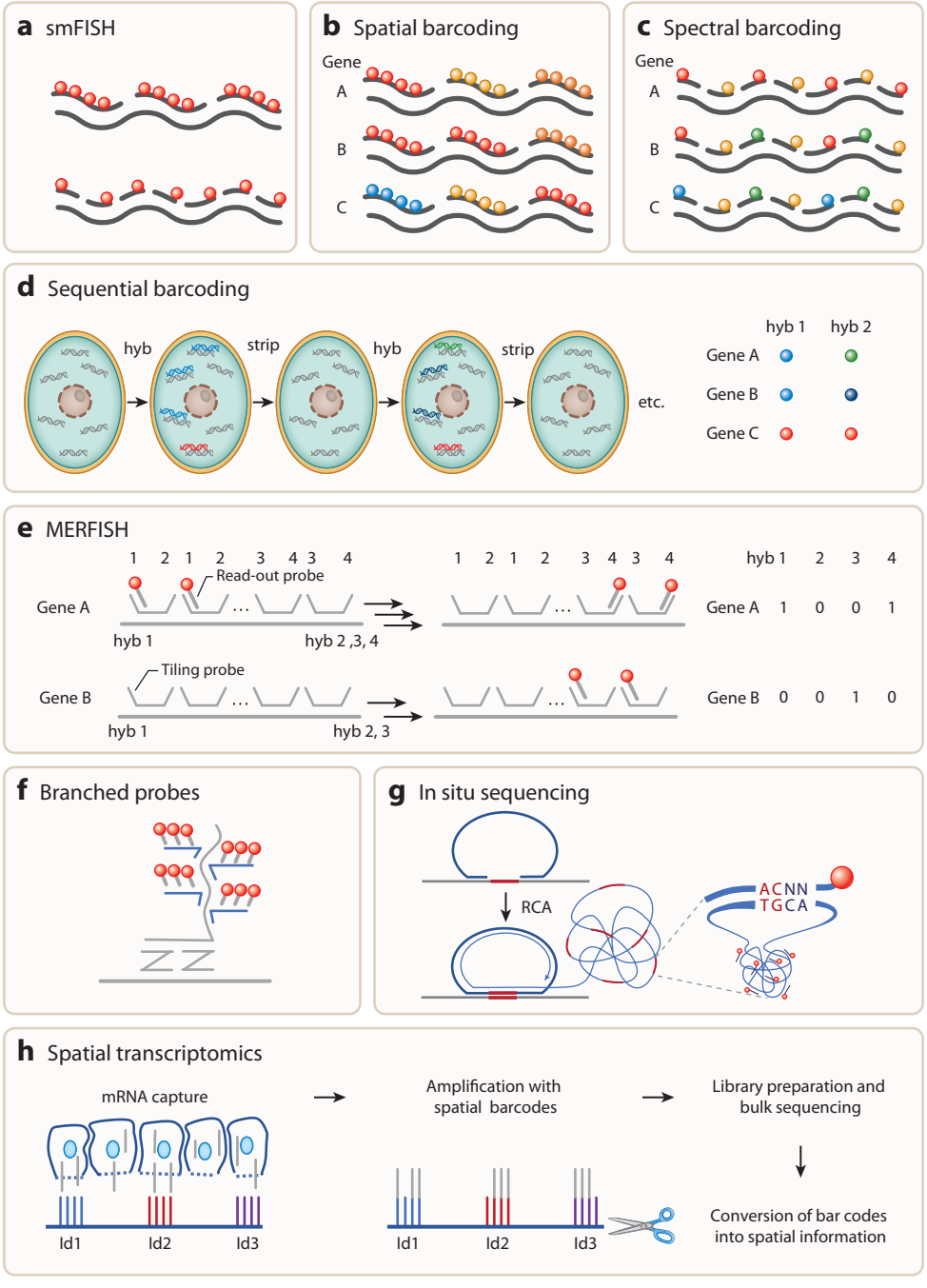
Single-Molecule RNA Fluorescent in Situ Hybridization

In situ hybridization of RNA with labeled sequence-specific probes followed by microscopic detection of the labels has been used for many decades to map gene expression onto tissues. Frequently probes are linked to enzymes that catalyze a chromogenic reaction or to fluorogenic molecules. A leap toward quantitative detection of RNA came with single-molecule RNA FISH (smFISH) (**Figure 3*a***). Individual transcripts are visualized using fluorescent probes and high-resolution microscopy (70). The use of shorter probes, each labeled with a single fluorophore (70, 71) at the 3′ terminus of each probe, has made this approach cheaper and more quantitative. Optimized hybridization conditions have greatly speeded up the protocols (70–72), and the use of different color fluorophores has allowed multiple genes to be detected in a single sample. smFISH is a highly quantitative method with near 100% detection sensitivity for multiple genes in parallel (63).

Multiplexed Single-Molecule RNA Fluorescent in Situ Hybridization

Spatial and spectral barcoding have led to further advances in multiplexing of probes (**Figure 3*b*,*c***) (73, 74). In spatial barcoding, mRNA molecules are hybridized to sets of different colored fluorophores along the length of the molecule, generating color barcodes detected by super-resolution microscopy. Spatial barcoding is technically challenging, requires high-resolution microscopes, and is limited by the need to compress the tissue to linearize the mRNA molecules.

The spectral barcoding approach uses unique combinations of different color fluorophores distributed across the mRNA, generating a pseudocolor that identifies the mRNA. This is done



(Caption appears on following page)

Figure 3 (*Figure appears on preceding page*)

Schematic representation of gene expression assays that retain spatial information in tissues. (*a–c*) smFISH can detect individual RNA transcripts at single-cell resolution. Either single or multiple fluorophores can be linked per probe (*a*) or used to generate color barcodes (*b*) or novel colors in a technique known as spectral barcoding (*c*). (*d*) Sequential barcoding relies on multiple rounds of hybridization and stripping of probes. Sequences of distinct fluorophores can mark specific genes. (*e*) MERFISH relies on multiple rounds of hybridization, generating an error-robust barcode that is used to identify genes. Arrows depict multiple rounds of hybridization. Further details are given in the text. (*f*) Branched probes allow for local signal amplification, thereby increasing the signal-to-noise ratio. (*g*) In situ sequencing relies on RCA to generate multiple identical copies of DNA that can then be profiled using sequencing by ligation with subsequent imaging. (*h*) Spatial transcriptomics is a technique in which a tissue section is overlaid onto a glass slide that carries locally distinct, barcoded, and identifiable polyA capture probes (Id1–3). After RNA is captured and transcribed, all further library preparation steps can be carried out in bulk, and sequenced reads can be mapped back onto tissue sections computationally. Abbreviations: hyb, hybridization; MERFISH, multiplexed error-robust fluorescence in situ hybridization; mRNA, messenger RNA; RCA, rolling circle amplification; smFISH, single-molecule in situ hybridization; strip, stripping.

using lower-resolution microscopy with no need to linearize the mRNA (73, 75, 76). Using spectral barcoding in combination with fluorescence resonance energy transfer between emitter and activator fluorophores, Lubeck & Cai (74) measured the single-cell gene expression profiles of 32 calcium-responsive genes in single yeast cells, finding clusters of coregulated genes. Subsequently it was realized that multiplexing could be achieved through sequential rounds of hybridization, imaging, and probe stripping, generating a temporal barcode (73). With this sequential FISH (seqFISH) approach, four dyes and eight rounds of hybridization ($4^8 = 65{,}536$) should in principle resolve an entire transcriptome. In a proof-of-principle experiment, 12 genes were analyzed with four dyes and two rounds of hybridization (73).

In a more recent iteration of the technique, multiplexed error-robust FISH (MERFISH) (76, 77) detected thousands of genes using a two-step hybridization protocol (**Figure 3*e***). In MERFISH, genes are tiled with hybridizing oligonucleotides, each containing a nonhomologous readout sequence at each end. Then, sequential hybridization of fluorophores to the readout sequences are visualized. In theory, 2^{16} (65,536) genes could be encoded. This technique is very sensitive to false negative (lack of hybridization) signals. However, the error-robust coding system, known as Hamming distance (78), can be applied to identify and correct errors. Using this approach, it was possible to measure 140 distinct genes with an 80% detection efficiency (77). The generation of complex, predefined probe sets is a challenge that remains in delivering this technology.

The MERFISH technique has been scaled using a two-color encoding scheme, enlarging the field of view, improving the performance of the readout probes, and increasing the speed of the image analysis (78, 79). This development means that MERFISH is fast approaching the throughput of droplet-based sequencing techniques, analyzing 130 genes in up to 40,000 cells. Although it does not profile at the same depth as droplet-based techniques, it does maintain spatial information and it can identify novel subpopulations of cells.

Branched Probes

Branched probes can generate high-contrast images from in situ hybridization. This is a highly sensitive technique that can be applied to standard histological sections and has been widely implemented and commercialized, for example by Advanced Cell Diagnostics with RNAscope® or Affymetrix, which use branched RNA and DNA probes, respectively, that allow for signal amplification. These probes are easily combined with antibody or DNA stains, aiding interpretation of the

results. In addition, methods to amplify the fluorescent signal by single-molecule hybridization chain reaction (smHCR) have been developed (78–80). By combining sequential hybridization with smHCR (74, 81) and implementing a simple but efficient error correction, researchers applied seqFISH to sections of the mouse hippocampus, profiling 249 genes in over 16,000 cells with only four rounds of hybridization (78–80, 82). Although this technique relies on predefined gene sets, potentially biasing results (83), the detection efficiency of seqFISH is high compared to the gold standard smFISH.

All FISH techniques are relatively easy to apply to single-cell layers but more difficult in tissue sections, as autofluorescence and light scatter increase with the sample thickness. Tissue clearing and hydrogel embedding, which remove lipids and proteins and replace them with a porous hydrogel, can be applied to any FISH-based protocol. This has been successfully applied to improve MERFISH (78, 79) and combined with seqFISH in whole-mount zebrafish embryos and 250-μm mouse brain sections (82). Technological advances in imaging and microscopy technology also help to improve data acquisition but are beyond the scope of this review.

Sequencing-Based Approaches

Two rolling circle amplification (RCA) methods have been described for spatial in situ sequencing (**Figure 2*g***). In fluorescence in situ sequencing of RNA, cells are fixed onto a glass slide. The RNA is then reverse transcribed into amine-modified cDNA and circularized (82, 84). Each cDNA is then linearly amplified and sequenced by well-established SOLiD (sequencing by oligonucleotide ligation and detection) sequencing technology, which relies on multiple rounds of capturing fluorescent dinucleotide pairs by ligation (for an overview see References 82 and 84). Fluorescence is then imaged and converted into sequence information. Ke et al. (85) combined RCA with padlock probes and sequencing-by-ligation chemistry to demonstrate that short fragments of RNA could be sequenced and mutations detected in subregions of cancer tissue (82, 84, 85). These techniques have great potential, but the number of transcripts that can be sequenced is still low and may be limited by the physical size of rolling circle amplicons.

An alternative approach termed spatial transcriptomics (**Figure 3*b***) (82, 84–86) uses histological tissue sections and spatial barcoding to analyze gene expression. Tissue sections are placed on a glass slide with positional barcoded oligo-dT capture probes. Cells are permeabilized, and mRNA is captured by the barcoded oligonucleotides and reverse transcribed. The cDNA can then be released from the slides, all further steps are carried out in bulk, and standard sequencing protocols can be applied. Sequencing reads are computationally mapped back to the tissue sections using the positional barcodes. Currently, spatial features have a diameter of 100 μm. Depending on the tissue, this represents 10–100 cells per barcode. By applying machine learning algorithms for dimensionality reduction followed by hierarchical clustering, researchers can generate sample clusters that correspond to well-defined morphological features and allow for unbiased identification of marker genes.

The ease and high throughput of spatial transcriptomics are likely to ensure its widespread adoption. Future developments are expected to produce smaller barcoded features, which will ensure that each feature captures fewer cells, further increasing the resolution of this technique. Compared to LCM, spatial transcriptomics is much less labor intensive and has greater detection sensitivity, but it is less sensitive than smFISH.

All of these spatial techniques are evolving at a rapid pace. Unlike the sequencing data, which are relatively simple strings of bases, several spatial techniques produce large, complex imaging data that are difficult to analyze (e.g., 73, 77). Currently, there are fewer computational methods for automatic high-throughput imaging analysis than those for sequencing analysis. The major challenge for the future is integrating these methods with other omics technologies.

RECONSTRUCTION OF THREE-DIMENSIONAL EXPRESSION PATTERNS AND THEIR ANALYSIS

As described earlier, LCM techniques have allowed gene expression patterns to be mapped into a 3D context, which in turn can be used as a zip-code to map single-cell data into their spatial context, as illustrated in a study of mouse development (68). Knowledge of in situ hybridization patterns for a relatively small number of marker genes is sufficient for the spatial assignment of cells. This computational assignment is based on the segregation of gene expression patterns from scRNA-seq data as obtained by Seurat (39) and other algorithms (87). In RNA tomography, mathematical image reconstruction can be used to generate 3D images (66, 88).

Expression patterns among genes are highly correlated in tissues; therefore, larger clusters of cells can be determined from as few as 10–12 genes, while a finer classification will require the combinatorial analysis of many genes (82). To date, the most advanced reconstruction of an organ is probably that of the mouse brain, as developed at the Allen Institute (**http://www.brain-map.org/**). The description of complex 3D data sets will require new tools that allow us to identify spatial changes in gene expression occurring at different scales and that are able to find recurring patterns of correlated expression (89).

Even with the wealth of new methods and tools to help with tissue reconstruction, there are still fundamental issues to overcome. There is evidence that tissue disaggregation itself may lead to the activation of many genes (90). Lovatt et al. (91) used caged, photoactivatable tags to analyze single-cell transcriptomes in vivo after laser activation and found that up to 30% fewer genes were expressed in the intact tissue context. A challenge for the future is the analysis of gene expression in living tissue. Currently, in vivo monitoring is restricted to model organisms, and the dynamic imaging of human cells has largely concentrated on in vitro cell cultures (92). Additional gene expression patterns and functional characteristics of cell types will become apparent once it is possible to monitor dynamic changes in live cells.

COMBINING DIFFERENT APPROACHES AND ITERATIVE EXPERIMENTAL DESIGN

Each of the described methods has its unique strengths and weaknesses: RNA-seq can profile very large numbers of cells in an unbiased manner and is well suited to define cell types and states (39, 93–97). However, scRNA-seq requires the disaggregation of the organ, leading to loss of spatial information, in order to generate single-cell suspensions that may not contain all the cell types present in the original tissue. MERFISH and seqFISH maintain the spatial information of cells and provide highly accurate measurements of fewer genes, but this analysis is not genome wide and genes to be analyzed must be predefined. FISH techniques are less well suited to finding and cataloguing new cell types but are superior to RNA-seq when trying to understand the interaction between differing cell types.

RNA-seq and FISH techniques provide different but complementary transcriptome resolutions. Multiplexed FISH can profile hundreds of preselected genes with great accuracy. scRNA-seq can in principle measure all expressed genes, but at lower quantitative accuracy and with a higher dropout rate. Spatial transcriptomics can generate full transcriptomes in an unbiased manner and maintains spatial information, but currently without single-cell resolution. In terms of its applications, it sits somewhere between RNA-seq and FISH.

The different approaches reviewed here provide information that can be combined and tailored to specific tissues and different biological questions (**Figure 4**). For instance, studies could combine RNA-seq with spatial gene expression for cataloguing the cell types found in different organs and add multiplexed FISH for understanding the functional dependencies of the same cells.

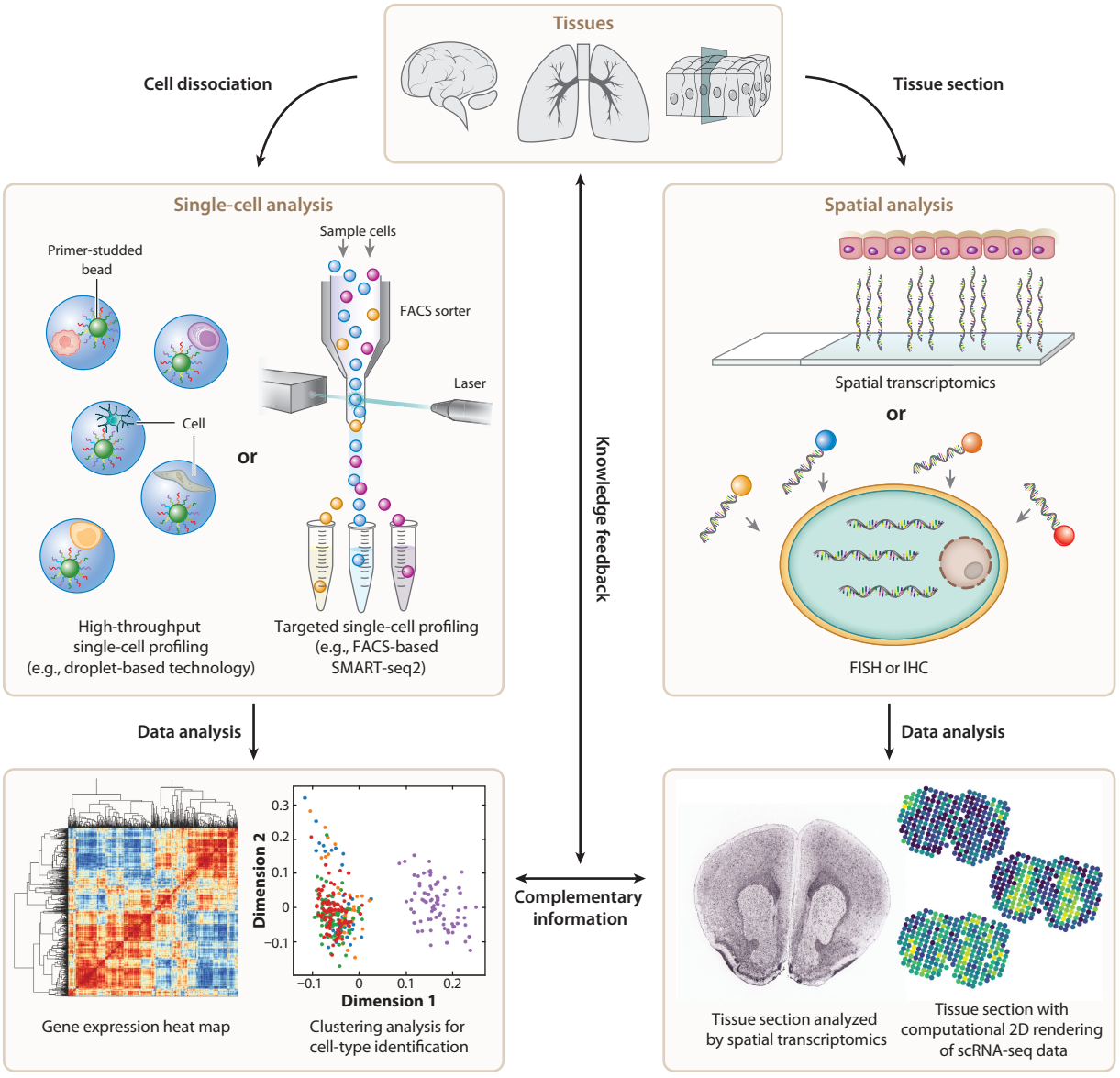


Figure 4

A hypothetical future workflow which combines single-cell genomics and spatial methods to understand tissue architecture. Complementary information obtained in different analysis techniques can be used to inform the experimental design through knowledge feedback. For example, genes that mark specific cell types identified through scRNA-seq can be used as probes in spatial analysis. Alternatively, spatial methods may indicate the presence of specific cell types that may have been lost during the cell dissociation process. Abbreviations: FACS, fluorescence-activated cell sorting; FISH, fluorescence in situ hybridization; IHC, immunohistochemistry; scRNA-seq, single-cell RNA sequencing.

An iterative experimental design is most likely to be successful in achieving a robust deconvolution of cell types within tissues. A high-throughput single-cell technique followed by clustering analysis will define subtypes of cells and identify markers that can then be used to obtain spatial data or enrich rare cell types (**Figure 4**). Integration of bulk RNA-seq data, single-cell data, and spatial data can determine whether certain cell types are underrepresented or even missing from the single-cell data. This is relevant when cell types are highly sensitive to tissue dissociation, requiring improved experimental protocols, and is particularly relevant for rare cell types. In this context, alternative sequencing strategies such as single-nucleus RNA-seq (98–100) can also be employed. In this approach, tissue is frozen and then mechanically dissociated before nuclei are isolated for sequencing. Several studies have already adapted droplet-based methods for high-throughput single-nucleus RNA-seq profiling (101, 102). Nuclear RNA-seq can capture all cells, but a disadvantage is its inability to enrich rare cell types based on cell surface markers.

Setting experimental standards that can be used in benchmarking studies to identify biases, which are undoubtedly associated with each of the different single-cell profiling methods, would greatly facilitate the integration of data sets derived from distinct experimental techniques.

CONCLUSIONS AND OUTLOOK

The last decade has seen an explosion in the field of single-cell genomics. The application of scRNA-seq, improved deconvolution techniques, and developments in spatial gene expression analysis mean that surprising new discoveries are on the horizon. The field is now at a stage where it can move from mapping individual tissue systems or model organisms to profiling the whole human body. To achieve this massive undertaking, we and others have recently initiated an international, multidisciplinary consortium to deliver the Human Cell Atlas (103), which aims to create "a comprehensive reference map of the types and properties of all human cells. . .as a basis for understanding, diagnosing, monitoring, and treating health and disease." Delivering this project will require tightly integrated interdisciplinary studies, where researchers with different expertise ranging from engineering to physics and medicine work together. The massive amount of data being generated will require input from information science, image analysis, deep learning, and others and will be a catalyst to developing new data analysis tools. One challenge for the future in this evolving landscape will be integrating multiple different data types, including multiomics data, into a comprehensive framework that is easily accessible to researchers from a wide range of fields. Addressing these challenges will help us better understand both how the human body functions, and ultimately, how this knowledge can be used to improve human health.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We thank Dr. Valentine Svensson and Dr. Kylie James from the Teichmann group for critical reading of the manuscript and Carmen Pryce for helpful suggestions and copyediting. We also thank Dr. Wenyi Wang for helpful comments.

LITERATURE CITED

1. GTEx Consort. 2013. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* 45(6):580–85
2. Bandura DR, Baranov VI, Ornatsky OI, Antonov A, Kinach R, et al. 2009. Mass cytometry: technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry. *Anal. Chem.* 81(16):6813–22
3. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, et al. 2009. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* 6(5):377–82
4. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, et al. 2011. Tumour evolution inferred by single-cell sequencing. *Nature* 472(7341):90–94
5. Kurimoto K, Yabuta Y, Ohinata Y, Ono Y, Uno KD, et al. 2006. An improved single-cell cDNA amplification method for efficient high-density oligonucleotide microarray analysis. *Nucleic Acids Res.* 34(5):e42
6. Kurimoto K, Yabuta Y, Ohinata Y, Saitou M. 2007. Global single-cell cDNA amplification to provide a template for representative high-density oligonucleotide microarray analysis. *Nat. Protoc.* 2(3):739–52
7. Mohammadi S, Zuckerman N, Goldsmith A, Grama A. 2017. A critical survey of deconvolution methods for separating cell types in complex tissues. *Proc. IEEE.* 105(2):340–66
8. Venet D, Pecasse F, Maenhaut C, Bersini H. 2001. Separation of samples into their constituents using gene expression data. *Bioinformatics* 17(Suppl. 1):S279–87
9. Lu P, Nakorchevskiy A, Marcotte EM. 2003. Expression deconvolution: a reinterpretation of DNA microarray data reveals dynamic changes in cell populations. *PNAS* 100(18):10370–75
10. Wang M, Master SR, Chodosh LA. 2006. Computational expression deconvolution in a complex mammalian organ. *BMC Bioinform.* 7:328
11. Shen-Orr SS, Tibshirani R, Khatri P, Bodian DL, Staedtler F, et al. 2010. Cell type–specific gene expression differences in complex tissues. *Nat. Methods* 7(4):287–89
12. Altboum Z, Steuerman Y, David E, Barnett-Itzhaki Z, Valadarsky L, et al. 2014. Digital cell quantification identifies global immune cell dynamics during influenza infection. *Mol. Syst. Biol.* 10:720
13. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, et al. 2015. Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* 12(5):453–57
14. Gentles AJ, Newman AM, Liu CL, Bratman SV, Feng W, et al. 2015. The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nat. Med.* 21(8):938–45
15. Shen-Orr SS, Gaujoux R. 2013. Computational deconvolution: extracting cell type-specific information from heterogeneous samples. *Curr. Opin. Immunol.* 25(5):571–78
16. Kivioja T, Vähärautio A, Karlsson K, Bonke M, Enge M, et al. 2011. Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* 9(1):72–74
17. Hashimshony T, Wagner F, Sher N, Yanai I. 2012. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep.* 2(3):666–73
18. Ramsköld D, Luo S, Wang Y-C, Li R, Deng Q, et al. 2012. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* 30(8):777–82
19. Stegle O, Teichmann SA, Marioni JC. 2015. Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.* 16(3):133–45
20. Stubbington MJT, Lönnberg T, Proserpio V, Clare S, Speak AO, et al. 2016. T cell fate and clonality inference from single-cell transcriptomes. *Nat. Methods* 13(4):329–32
21. Lindeman I, Emerton G, Sollid LM, Teichmann S. 2017. BraCeR: reconstruction of B-cell receptor sequences and clonality inference from single-cell RNA-sequencing. bioRxiv 185504. **https://doi.org/10.1101/185504**
22. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, et al. 2015. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 161(5):1202–14
23. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, et al. 2015. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* 161(5):1187–201
24. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, et al. 2017. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* 8:14049

25. Gierahn TM, Wadsworth MH 2nd, Hughes TK, Bryson BD, Butler A, et al. 2017. Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nat. Methods* 14(4):395–98
26. Arvanti E, Claasen M. 2017. Sensitive detection of rare disease-associated cell subsets via representation learning. *Nat. Commun.* 8:14825
27. Cao J, Packer JS, Ramani V, Cusanovich DA, Huynh C, et al. 2017. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* 357(6352):661–67
28. Rosenberg AB, Roco C, Muscat RA, Kuchina A, Mukherjee S, et al. 2017. Scaling single cell transcriptomics through split pool barcoding. bioRxiv 105163. **https://doi.org/10.1101/105163**
29. Amini S, Pushkarev D, Christiansen L, Kostem E, Royce T, et al. 2014. Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. *Nat. Genet.* 46(12):1343–49
30. Engström PG, Steijger T, Sipos B, Grant GR, Kahles A, et al. 2013. Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat. Methods* 10(12):1185–91
31. Everaert C, Luypaert M, Maag JLV, Cheng QX, Dinger ME, et al. 2017. Benchmarking of RNA-sequencing analysis workflows using whole-transcriptome RT-qPCR expression data. *Sci. Rep.* 7(1):1559
32. Teng M, Love MI, Davis CA, Djebali S, Dobin A, et al. 2016. A benchmark for RNA-seq quantification pipelines. *Genome Biol.* 17:74
33. Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* 34(5):525–27
34. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. 2017. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* 14(4):417–19
35. Tung P-Y, Blischak JD, Hsiao CJ, Knowles DA, Burnett JE, et al. 2017. Batch effects and the effective design of single-cell gene expression studies. *Sci. Rep.* 7:39921
36. Haghverdi L, Lun ATL, Morgan MD, Marioni JC. 2017. Correcting batch effects in single-cell RNA sequencing data by matching mutual nearest neighbours. bioRxiv 165118. **https://doi.org/10.1101/165118**
37. Buttner M, Miao Z, Wolf A, Teichmann SA, Theis FJ. 2017. Assessment of batch-correction methods for scRNA-seq data with a new test metric. bioRxiv 200345. **https://doi.org/10.1101/200345**
38. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, et al. 2014. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* 32(4):381–86
39. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. 2015. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* 33(5):495–502
40. Wolf FA, Angerer P, Theis FJ. 2018. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* 19(1):15
41. Soneson C, Robinson MD. 2018. Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods* 15:225–61
42. Svensson V, Natarajan KN, Ly L-H, Miragaia RJ, Labalette C, et al. 2017. Power analysis of single-cell RNA-sequencing experiments. *Nat. Methods* 14(4):381–87
43. Ziegenhain C, Vieth B, Parekh S, Reinius B, Guillaumet-Adkins A, et al. 2017. Comparative analysis of single-cell RNA sequencing methods. *Mol. Cell* 65(4):631–43.e4
44. Tang F, Barbacioru C, Bao S, Lee C, Nordman E, et al. 2010. Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis. *Cell Stem Cell* 6(5):468–78
45. Yan L, Yang M, Guo H, Yang L, Wu J, et al. 2013. Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nat. Struct. Mol. Biol.* 20(9):1131–39
46. Xue Z, Huang K, Cai C, Cai L, Jiang C-Y, et al. 2013. Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature* 500(7464):593–97
47. Petropoulos S, Edsgärd D, Reinius B, Deng Q, Panula SP, et al. 2016. Single-cell RNA-Seq reveals lineage and X chromosome dynamics in human preimplantation embryos. *Cell* 165(4):1012–26
48. Qiu X, Mao Q, Tang Y, Wang L, Chawla R, et al. 2017. Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* 14:979–82
49. Bendall SC, Davis KL, Amir E-AD, Tadmor MD, Simonds EF, et al. 2014. Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* 157(3):714–25

50. Shin J, Berg DA, Zhu Y, Shin JY, Song J, et al. 2015. Single-cell RNA-seq with Waterfall reveals molecular cascades underlying adult neurogenesis. *Cell Stem Cell* 17(3):360–72
51. Marco E, Karp RL, Guo G, Robson P, Hart AH, et al. 2014. Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *PNAS* 111(52):E5643–50
52. Setty M, Tadmor MD, Reich-Zeliger S, Angel O, Salame TM, et al. 2016. Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat. Biotechnol.* 34(6):637–45
53. Haghverdi L, Büttner M, Wolf FA, Buettner F, Theis FJ. 2016. Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* 13(10):845–48
54. Macaulay IC, Svensson V, Labalette C, Ferreira L, Hamey F, et al. 2016. Single-cell RNA-sequencing reveals a continuous spectrum of differentiation in hematopoietic cells. *Cell Rep.* 14(4):966–77
55. Lönnberg T, Svensson V, James KR, Fernandez-Ruiz D, Sebina I, et al. 2017. Single-cell RNA-seq and computational analysis using temporal mixture modelling resolves T_H1/T_{FH} fate bifurcation in malaria. *Sci. Immunol.* 2(9):eaal2192
56. Boukouvalas A, Hensman J, Rattray M. 2017. BGP: branched Gaussian processes for identifying gene-specific branching dynamics in single cell data. bioRxiv 166868. **http://doi:10.1101/166868**
57. Penfold CA, Sybirna A, Reid J, Huang Y, Wernisch L, et al. 2017. Nonparametric Bayesian inference of transcriptional branching and recombination identifies regulators of early human germ cell development. bioRxiv 167684. **http://doi:10.1101/167684**
58. Saelens W, Cannoodt R, Todorov H, Saeys Y. 2018. A comparison of single-cell trajectory inference methods: towards more accurate and robust tools. bioRxiv 276907. **http://doi:10.1101/276907**
59. Svensson V, Vento-Tormo R, Teichmann SA. 2018. Exponential scaling of single-cell RNA-seq in the last decade. *Nat Protoc.* 13(4):599–604
60. Dixit A, Parnas O, Li B, Chen J, Fulco CP, et al. 2016. Perturb-seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* 167(7):1853–66.e17
61. Adamson B, Norman TM, Jost M, Cho MY, Nuñez JK, et al. 2016. A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell* 167(7):1867–82.e21
62. Jaitin DA, Weiner A, Yofe I, Lara-Astiaso D, Keren-Shaul H, et al. 2016. Dissecting immune circuits by linking CRISPR-pooled screens with single-cell RNA-seq. *Cell* 167(7):1883–96.e15
63. Datlinger P, Rendeiro AF, Schmidl C, Krausgruber T, Traxler P, et al. 2017. Pooled CRISPR screening with single-cell transcriptome readout. *Nat. Methods* 14(3):297–301
64. Guo G, Huss M, Tong GQ, Wang C, Li Sun L, et al. 2010. Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. *Dev. Cell* 18(4):675–85
65. Pollen AA, Nowakowski TJ, Shuga J, Wang X, Leyrat AA, et al. 2014. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.* 32(10):1053–58
66. Chen J, Suo S, Tam PPL, Han J-DJ, Peng G, Jing N. 2017. Spatial transcriptomic analysis of cryosectioned tissue samples with Geo-seq. *Nat. Protoc.* 12(3):566–80
67. Butler AE, Matveyenko AV, Kirakossian D, Park J, Gurlo T, Butler PC. 2016. Recovery of high-quality RNA from laser capture microdissected human and rodent pancreas. *J. Histotechnol.* 39(2):59–65
68. Peng G, Suo S, Chen J, Chen W, Liu C, et al. 2016. Spatial transcriptome for the molecular annotation of lineage fates and cell identity in mid-gastrula mouse embryo. *Dev. Cell* 36(6):681–97
69. Okamura-Oho Y, Shimokawa K, Takemoto S, Hirakiyama A, Nakamura S, et al. 2012. Transcriptome tomography for brain analysis in the web-accessible anatomical space. *PLOS ONE* 7(9):e45373
70. Femino AM, Fay FS, Fogarty K, Singer RH. 1998. Visualization of single RNA transcripts in situ. *Science* 280(5363):585–90
71. Raj A, van den Bogaard P, Rifkin SA, van Oudenaarden A, Tyagi S. 2008. Imaging individual mRNA molecules using multiple singly labeled probes. *Nat. Methods* 5(10):877–79
72. Shaffer SM, Wu M-T, Levesque MJ, Raj A. 2013. Turbo FISH: A method for rapid single molecule RNA FISH. *PLOS ONE* 8(9):e75120
73. Lubeck E, Coskun AF, Zhiyentayev T, Ahmad M, Cai L. 2014. Single-cell in situ RNA profiling by sequential hybridization. *Nat. Methods* 11(4):360–61

74. Lubeck E, Cai L. 2012. Single-cell systems biology by super-resolution imaging and combinatorial labeling. *Nat. Methods* 9(7):743–48
75. Levsky JM, Shenoy SM, Pezo RC, Singer RH. 2002. Single-cell gene expression profiling. *Science* 297(5582):836–40
76. Levesque MJ, Raj A. 2013. Single-chromosome transcriptional profiling reveals chromosomal gene expression regulation. *Nat. Methods* 10(3):246–48
77. Chen KH, Boettiger AN, Moffitt JR, Wang S, Zhuang X. 2015. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* 348(6233):aaa6090
78. Moon TK. 2005. *Error Correction Coding: Mathematical Methods and Algorithms*. Hoboken, NJ: Wiley
79. Moffitt JR, Hao J, Wang G, Chen KH, Babcock HP, Zhuang X. 2016. High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *PNAS* 113(39):11046–51
80. Shah S, Lubeck E, Schwarzkopf M, He T-F, Greenbaum A, et al. 2016. Single-molecule RNA detection at depth by hybridization chain reaction and tissue hydrogel embedding and clearing. *Development* 143(15):2862–67
81. Choi HMT, Chang JY, Trinh LA, Padilla JE, Fraser SE, Pierce NA. 2010. Programmable in situ amplification for multiplexed imaging of mRNA expression. *Nat. Biotechnol.* 28(11):1208–12
82. Shah S, Lubeck E, Zhou W, Cai L. 2016. In situ transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus. *Neuron* 92(2):342–57
83. Cembrowski MS, Spruston N. 2017. Integrating results across methodologies is essential for producing robust neuronal taxonomies. *Neuron* 94(4):747–51.e1
84. Lee JH, Daugharthy ER, Scheiman J, Kalhor R, Ferrante TC, et al. 2015. Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. *Nat. Protoc.* 10(3):442–58
85. Ke R, Mignardi M, Pacureanu A, Svedlund J, Botling J, et al. 2013. In situ sequencing for RNA analysis in preserved tissue and cells. *Nat. Methods* 10(9):857–60
86. Ståhl PL, Salmén F, Vickovic S, Lundmark A, Navarro JF, et al. 2016. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* 353(6294):78–82
87. Achim K, Pettit J-B, Saraiva LR, Gavriouchkina D, Larsson T, et al. 2015. High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. *Nat. Biotechnol.* 33(5):503–9
88. Junker JP, Noël ES, Guryev V, Peterson KA, Shah G, et al. 2014. Genome-wide RNA tomography in the zebrafish embryo. *Cell* 159(3):662–75
89. Svensson V, Teichmann SA, Stegle O. 2017. SpatialDE: identification of spatially variable genes. bioRxiv 143321. **http://doi:10.1101/143321**
90. van den Brink SC, Sage F, Vértesy Á, Spanjaard B, Peterson-Maduro J, et al. 2017. Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nat. Methods* 14(10):935–36
91. Lovatt D, Ruble BK, Lee J, Dueck H, Kim TK, et al. 2014. Transcriptome in vivo analysis (TIVA) of spatially defined single cells in live tissue. *Nat. Methods* 11(2):190–96
92. Skylaki S, Hilsenbeck O, Schroeder T. 2016. Challenges in long-term imaging and quantification of single-cell dynamics. *Nat. Biotechnol.* 34(11):1137–44
93. Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, et al. 2017. SC3: consensus clustering of single-cell RNA-seq data. *Nat. Methods* 14(5):483–86
94. Wang B, Zhu J, Pierson E, Ramazzotti D, Batzoglou S. 2017. Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nat. Methods* 14(4):414–16
95. Qiu X, Hill A, Packer J, Lin D, Ma Y-A, Trapnell C. 2017. Single-cell mRNA quantification and differential analysis with Census. *Nat. Methods* 14(3):309–15
96. Jiang L, Chen H, Pinello L, Yuan G-C. 2016. GiniClust: detecting rare cell types from single-cell gene expression data with Gini index. *Genome Biol.* 17(1):144
97. Kharchenko PV, Silberstein L, Scadden DT. 2014. Bayesian approach to single-cell differential expression analysis. *Nat. Methods* 11(7):740–42
98. Grindberg RV, Yee-Greenbaum JL, McConnell MJ, Novotny M, O'Shaughnessy AL, et al. 2013. RNA-sequencing from single nuclei. *PNAS* 110(49):19802–7
99. Lacar B, Linker SB, Jaeger BN, Krishnaswami S, Barron J, et al. 2016. Nuclear RNA-seq of single neurons reveals molecular signatures of activation. *Nat. Commun.* 7:11022

100. Habib N, Li Y, Heidenreich M, Swiech L, Avraham-Davidi I, et al. 2016. Div-Seq: single-nucleus RNA-Seq reveals dynamics of rare adult newborn neurons. *Science* 353(6302):925–28
101. Lake BB, Chen S, Sos BC, Fan J, Kaeser GE, et al. 2018. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat. Biotechnol.* 36(1):70–80
102. Habib N, Avraham-Davidi I, Basu A, Burks T, Shekhar K, et al. 2017. Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nat. Methods* 14(10):955–58
103. Regev A, Teichmann S, Lander ES, Amit I, Benoist C, et al. 2017. The Human Cell Atlas. bioRxiv 121202. **http://doi:10.1101/121202**
104. Jaitin DA, Kenigsberg E, Keren-Shaul H, Elefant N, Paul F, et al. 2014. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* 343(6172):776–79
105. Zeisel A, Muñoz-Manchado AB, Codeluppi S, Lönnerberg P, La Manno G, et al. 2015. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* 347(6226):1138–42
106. Muraro MJ, Dharmadhikari G, Grün D, Groen N, Dielen T, et al. 2016. A single-cell transcriptome atlas of the human pancreas. *Cell Syst.* 3(4):385–94.e3
107. Segerstolpe Å, Palasantza A, Eliasson P, Andersson E-M, Andréasson A-C, et al. 2016. Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell Metab.* 24(4):593–607
108. Xin Y, Kim J, Okamoto H, Ni M, Wei Y, et al. 2016. RNA sequencing of single human islet cells reveals type 2 diabetes genes. *Cell Metab.* 24(4):608–15
109. Villani A-C, Satija R, Reynolds G, Sarkizova S, Shekhar K, et al. 2017. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* 356(6335):eaah4573
110. Wyss-Coray T, Darmanis S, Muris Consortium T. 2017. Transcriptomic characterization of 20 organs and tissues from mouse at single cell resolution creates a Tabula Muris. bioRxiv 237446. **https://doi.org/10.1101/237446**
111. Han X, Wang R, Zhou Y, Fei L, Sun H, et al. 2018. Mapping the mouse cell atlas by Microwell-seq. *Cell* 172(5):1091–107.e17
112. Islam S, Kjällquist U, Moliner A, Zajac P, Fan J-B, et al. 2011. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.* 21(7):1160–67
113. Islam S, Zeisel A, Joost S, La Manno G, Zajac P, et al. 2014. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* 11(2):163–66
114. Picelli S, Björklund ÅK, Faridani OR, Sagasser S, Winberg G, Sandberg R. 2013. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* 10(11):1096–98
115. Sasagawa Y, Nikaido I, Hayashi T, Danno H, Uno KD, et al. 2013. Quartz-Seq: a highly reproducible and sensitive single-cell RNA-Seq reveals non-genetic gene expression heterogeneity. *Genome Biol.* 14(4):R31
116. Sasagawa Y, Danno H, Takada H, Ebisawa M, Hayashi T, et al. 2017. Quartz-Seq2: a high-throughput single-cell RNA-sequencing method that effectively uses limited sequence reads. bioRxiv 159384. **http://doi:10.1101/159384**
117. Hashimshony T, Senderovich N, Avital G, Klochendler A, de Leeuw Y, et al. 2016. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.* 17:77
118. Fan HC, Fu GK, Fodor SPA. 2015. Combinatorial labeling of single cells for gene expression cytometry. *Science* 347(6222):1258367
119. Hochgerner H, Lönnerberg P, Hodge R, Mikes J, Heskol A, et al. 2017. STRT-seq-2i: dual-index 5′ single cell and nucleus RNA-seq on an addressable microwell array. *Sci. Rep.* 7:16327
120. Bose S, Wan Z, Carr A, Rizvi AH, Vieira G, et al. 2015. Scalable microfluidics for single-cell RNA printing and sequencing. *Genome Biol.* 16:120
121. Yuan J, Sims PA. 2016. An automated Microwell platform for large-scale single cell RNA-Seq. *Sci. Rep.* 6:33883
122. Brady G, Barbara M, Iscove NN. 1990. Representative in vitro cDNA amplification from individual hemopoietic cells and colonies. *Methods Mol. Cell. Biol.* 2:17–25
123. Tietjen I, Rihel JM, Cao Y, Koentges G, Zakhary L, Dulac C. 2003. Single-cell transcriptional analysis of neuronal progenitors. *Neuron* 38(2):161–75
124. Chiang M-K, Melton DA. 2003. Single-cell transcript analysis of pancreas development. *Dev. Cell* 4(3):383–93

125. Schmidt WM, Mueller MW. 1999. CapSelect: a highly sensitive method for 5′ CAP- dependent enrichment of full-length cDNA in PCR-mediated analysis of mRNAs. *Nucleic Acids Res.* 27(21):e31
126. Zhu YY, Machleder EM, Chenchik A, Li R, Siebert PD. 2001. Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. *Biotechniques* 30(4):892–97
127. Gubler U, Hoffman BJ. 1983. A simple and very efficient method for generating cDNA libraries. *Gene* 25(2–3):263–69
128. Okayama H, Berg P. 1982. High-efficiency cloning of full-length cDNA. *Mol. Cell. Biol.* 2(2):161–70

Annual Review
of Biomedical Data
Science

Volume 1, 2018

Contents

Errata

An online log of corrections to *Annual Review of Biomedical Data Science* articles may be found at http://www.annualreviews.org/errata/biodatasci