OXFORD

Briefings in Functional Genomics, 17(1), 2018, 64–76

doi: 10.1093/bfgp/elx019
Advance Access Publication Date: 1 September 2017
Review paper

A brief review of single-cell transcriptomic technologies

Tomer Kalisky*, Sarit Oriel, Tali Hana Bar-Lev, Nissim Ben-Haim, Ariel Trink, Yishay Wineberg, Itamar Kanter, Shlomit Gilad, and Saumyadipta Pyne*

Corresponding authors: Tomer Kalisky, Faculty of Engineering and Bar-Ilan Institute of Nanotechnology and Advanced Materials (BINA), Bar-Ilan University, Ramat Gan 5290002, Israel. E-mail: Tomer.bioengineering@gmail.com; Saumyadipta Pyne, Indian Institute of Public Health, Kavuri Hills, Madhapur, Hyderabad, India. E-mail: saumyadipta@yahoo.com.
*These authors share equal senior co-authorship and correspondence.

Abstract

In recent years, there has been an effort to develop new technologies for measuring gene expression and sequence information from thousands of individual cells. Large data sets that were obtained using these 'single cell' technologies have allowed scientists to address fundamental questions in biomedicine ranging from stems cells and development to cancer and immunology. Here, we provide a brief review of recent developments in single-cell technology. Our intention is to provide a quick background for newcomers to the field as well as a deeper description of some of the leading technologies to date.

Key words: single-cell technologies; genomics; RNA sequencing

Introduction: The rationale of single-cell analysis

Why single cells?

Over the past decade, progress in biochemistry, physics, engineering and computer science has led to the development of high-throughput technologies for measuring gene expression and sequence information from biological samples. High-throughput technologies such as RNA sequencing [1, 2] allowed, for the first time, a deeper and detailed understanding of complex biological processes such as organism development, tissue regeneration and cancer.

Library preparation procedures for RNA sequencing typically require large amounts of starting material (DNA/RNA), and thus have the serious drawback that they 'sum-up' information from thousands of cells. However, it is sometimes difficult to obtain a

Tomer Kalisky (PhD) is an assistant professor of Bioengineering in Bar-Ilan University, Israel, and a member of the Bar-Ilan Institute of Nanotechnology and Advanced Materials (BINA). His research focuses on single-cell genomics of tissues and tumors and identification of tissue-specific and cancer stem cells.
Sarit Oriel (PhD) is a postdoctoral scientist in the Department of Bioengineering, Bar-Ilan University, Israel. Her research focuses on single-cell qPCR and targeted RNA sequencing of the developing kidney.
Tali Hana Bar-Lev (PhD) is a postdoctoral scientist in the Department of Bioengineering, Bar-Ilan University, Israel. Her research focuses on single-cell RNA sequencing of the developing kidney and pediatric tumors.
Nissim Ben-Haim is an MSc student in the Department of Bioengineering, Bar-Ilan University, Israel. He is interested in computational analysis of single-cell RNA sequencing data.
Ariel Trink in an undergraduate student in the Department of Bioengineering, Bar-Ilan University, Israel. His research is focused on the genomics of Wilms' tumor—a pediatric tumor of the kidney.
Yishay Wineberg is an MSc student in the Department of Bioengineering, Bar-Ilan University, Israel. He is interested in computational analysis of single-cell RNA sequencing data and molecular tagging techniques.
Itamar Kanter (PhD) graduated from the Department of Bioengineering, Bar-Ilan University, Israel. He is interested in machine learning and in developing network-based algorithms for analysis of single-cell RNA sequencing data sets.
Shlomit Gilad (PhD) is the head of the R&D for Genomics in the Nancy and Stephen Grand Israel National Center for Personalized Medicine (G-INCPM), Israel. Her research interest lies in genomics, in particular, developing new technologies for high-throughput sequencing for both bulk samples and single cells.
Saumyadipta Pyne (PhD) is a professor at the Indian Institute of Public Health, Hyderabad, and Ramalingaswami Fellow of the Department of Biotechnology, India. His research interests include big data in life Sciences and health informatics, computational statistics and high-dimensional data modeling. He will shortly join as a faculty member of the Graduate School of Public Health, University of Pittsburgh, USA.

large number of cells, for example, when studying circulating tumor cells or an early stage embryo. Moreover, it was realized that many biological systems are composed of heterogeneous cell types in which minority cell populations play important roles. In these cases, the minority cells that often govern the overall system behavior are not well represented in 'bulk' measurements, and are therefore hard to identify and characterize. For example, in many tumors, only a small fraction (typically <1%) of molecularly distinct cells called 'Cancer Stem Cells' have the capacity for self-renewal and tumorigenesis [3–10]. Likewise, minority tissue-specific stem cell populations are responsible for tissue development, regeneration and repair [11–14]. Identification and molecular characterization of cancer stem cells and tissue-specific stem cells, as well as the gene circuits that govern their behavior and interactions with other cell types, has enormous therapeutic potential for developing targeted therapies for cancer and for the effort to reconstruct damaged tissues and organs. Another striking example is the immune system in which dynamic cellular heterogeneity is essential for fighting off the variety of attacking pathogens [15]. Therefore, it is essential to use single-cell transcriptomic technologies to fully understand these complex biological systems.

There are numerous challenges in single-cell measurements. First, as RNA molecules are mostly unstable, individual cells must be carefully collected and rapidly isolated from the tissue or tumor to keep the cells as viable as possible. Second, as a single cell contains roughly 10 pg of total RNA, which is much smaller than the nanogram amounts typically required for most gene expression and sequencing assays, a pre-amplification step is required before the actual expression measurement or sequencing can be done, which introduces additional noise, bias and sequencing errors. Third, as the number of measurements is enormous (number of cells × number of genes × number of transcripts), elaborate multiplexing strategies such as microfluidic parallelization, droplet encapsulation and molecular barcoding (tagging) must be used to perform many millions of biochemical measurements at feasible time and cost.

A typical single-cell experiment

The first step in a typical single-cell experiment is to collect the tissue or tumor and dissociate it into a single-cell suspension using digestive enzymes and/or mechanical shearing [16, 17]. Then, as an optional step, the desired cell population can be enriched using previously known markers. This is particularly useful when the sought-out cell subpopulation is extremely rare (e.g. cancer stem cells, which can constitute 0.1–1% of the tumor bulk). If there is a sufficient number of cells, cell enrichment can be done by using flow cytometry with two to three fluorescently labeled antibodies or reporter genes [18]. Another option is to mix the cells with magnetic microbeads coated with antibodies and then pass them through a magnetic separation column [19].

Individual cells are then isolated from each other in separate compartments. Cell isolation can be done by using a flow cytometer to sort single cells into separate wells [16], by 'picking' individual cells using microscopy-assisted micropipetting [20], by pushing a cell suspension through a series of microfluidic traps [19, 21, 22], by allowing individual cells to settle in microwells [23–25] or by encapsulating in nanoliter-sized droplets [26, 27]. Once each cell is isolated in its own 'partition' (well/microfluidic chamber/microwell/droplet), it is lysed and mixed with buffers, enzymes, dNTPs and primers, and the RNA is reverse transcribed and pre-amplified. Reverse transcription can be done specifically for 100–200 targets [16, 23] or globally for all polyadenylated transcripts [19, 28, 29]. Pre-amplification is usually done by polymerase chain reaction (PCR) or linear methods such as *in vitro* transcription [30, 31].

At this point, gene expression and sequence information can be obtained from the amplified complementary DNA (cDNA) of each individual cell. Expression levels of multiple genes from each cell can be directly measured by quantitative real-time PCR (qPCR) [32]. By using carefully designed primers and microfluidic devices that allow for tens of thousands of qPCR reactions to be performed in parallel, expression levels of multiple (up to ~300) genes from hundreds of individual cells can be measured [16, 33–35]. Once a reasonable number of single-cell expression profiles have been measured, computational clustering algorithms can be used to cluster them into groups—each representing a distinct cell subpopulation. Each subpopulation can be identified according to the existing scientific literature and online databases incorporating histological information and functional studies [36, 37]. From our experience, to identify cell types within a tissue or tumor, it typically is desired to have at least 20–30 representative cells from each population and to be able to observe the same subpopulation repertoire within at least two independent biological experiments.

Even more information can be obtained by RNA sequencing, which provides both expression and sequence data for all mRNA transcripts without preselecting primers. For single-cell RNA sequencing, the pre-amplified cDNA is collected, barcoded, pooled and sequenced on a next-generation sequencer such as the Illumina platform [28, 29]. The output short sequence reads are aligned to the reference genome using splice aware aligners such as TopHat [38–40] or STAR [41], and gene expression levels are inferred from the number of reads that overlap with each gene [42–45]. The underlying assumption is that genes that had many mRNA transcripts in the original cell will result in many corresponding amplified fragments and many aligned reads. Molecular tagging techniques [46–48] can be used to 'count' the original transcripts while correcting for bias caused by pre-amplification. Once single-cell expression levels from hundreds of individual cells has been obtained, single-cell profiles can be clustered to identify cell subpopulations, and data from all cells belonging to the same subpopulation can be mixed *in silico* to provide a deeper characterization of each subpopulation [30]. As RNA sequencing technology provides information about the whole transcriptome—and not only a predetermined set of genes—it can be used to discover new genes and surface markers that are uniquely expressed by each cell subpopulation [19]. Moreover, sequence data can reveal splicing patterns [28, 49–52], allele-specific expression [53], single-nucleotide variations (SNVs) [54] and copy number variations (CNVs) [54].

Single-cell transcriptomic technologies

In the past few years, there has been an explosion of new single-cell transcriptomic technologies [55, 56], each having its unique capabilities and limitations in terms of throughput (i.e. the number of cells and genes that can be measured), sensitivity (i.e. the ability to detect lowly expressed genes), accuracy (i.e. how close the measurement is to the true value) and precision (i.e. how well the results can be reproduced on replicate samples) [19, 57]. The field is developing quickly with new technologies coming out every few months [58]—see Table 1 for selected technologies. In this briefing, we will describe the main trends and delve deeply into a few examples.

Single-cell qPCR is used to measure expression of multiple selected genes in hundreds of individual cells with sensitivity and precision

qPCR was one of the first genomic technologies to be used for measuring the expression of selected genes from single cells [32]. Measuring multiple genes simultaneously from hundreds of individual cells requires performing a huge number of biochemical reactions in parallel (e.g. 100 genes × 100 cells = 10 000 reactions), which is infeasible in terms of cost and labor. This problem was overcome by the development of microfluidic devices such as Fluidigm Dynamic Arrays [33, 59, 60], which can achieve a high level of parallelization and combinatorial mixing.

Typically (Figure 1), individual cells are first sorted by a flow cytometer into individual wells of 96 well plates prefilled with PCR buffer [16, 33]. Cells are lysed by influx of fluids through their cell membranes because of hypotonic pressure. Then, the RNA is reverse transcribed and PCR-amplified for 14–20 cycles using a mixture of primers for a predetermined set of genes (this is referred to as ‘specific target transcript amplification’ or STA). Once the single-cell cDNA has been amplified, expression levels of selected genes can be measured using benchtop qPCR in tubes [32] or Fluidigm Dynamic Arrays [33, 59]. The Dynamic Array chip is built as a matrix of 48–96 microfluidic channels carrying different single-cell cDNA samples crossing 48–96 channels carrying gene-specific ‘detectors’ (e.g. primers) [60]. At each intersection, cDNA from a specific cell and primers designed against a specific gene are mixed in a separate microfluidic chamber and a single-real-time qPCR reaction takes place. Using this strategy, microfluidic parallelization can be used to perform thousands of independent qPCR reactions in a single experiment.

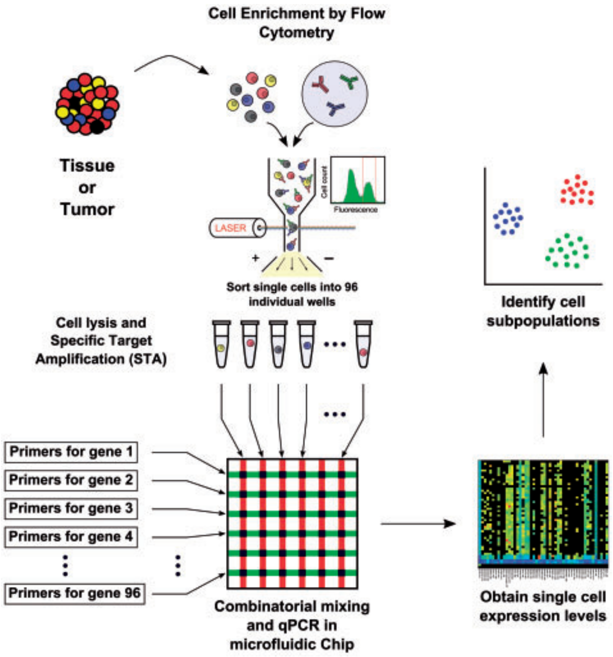

Cell Enrichment by Flow Cytometry
Tissue or Tumor
LASER
Cell count
Fluorescence
+
−
Sort single cells into 96 individual wells
Identify cell subpopulations
Cell lysis and Specific Target Amplification (STA)
Primers for gene 1
Primers for gene 2
Primers for gene 3
Primers for gene 4
Primers for gene 96
Combinatorial mixing and qPCR in microfluidic Chip
Obtain single cell expression levels

Figure 1. Sketch of a typical single-cell qPCR workflow performed with the Fluidigm Dynamic Array microfluidic chip. Combinatorial mixing on the microfluidic chip allows for performing up to 96 × 96 = 9216 qPCR reactions in parallel, thus allowing expression levels of 96 genes from 96 individual cells to be measured in a single experiment.

The resulting data are in the form of a matrix of threshold cycles (Cts) whose rows represent individual cells and whose columns represent individual genes. The single-cell gene expression matrix can be standardized (e.g. by subtracting the mean Ct of each specific gene averaged over all cells and by dividing by the SD) and partitioned into groups using clustering algorithms such as *K*-means or hierarchical clustering. Cell subpopulations are identified as clusters of single-cell profiles that express common sets of genes. From our experience, there is little to gain from normalizing to ‘housekeeping genes’ (as done in bulk gene expression measurements), as each single cell is inherently normalized.

As microfluidic single-cell qPCR is based on reverse transcription and amplification of specific preselected targets, it has high sensitivity and precision and a wide dynamic range [21, 61]. Primers for each specific gene can be carefully designed and optimized [35] to allow for detection of rare transcripts and transcription factors. Moreover, by running multiple microfluidic chips per pre-amplified material, expression of up to 280 genes from hundreds of individual cells has been obtained [35]. As a result, single-cell qPCR can provide a detailed multigene picture of the cell subpopulation repertoire from which a complex tissue or tumor is composed [16, 18, 20]. However, the major drawback is that the genes defining the various cell types have to be known in advance or ‘guessed’ from the known scientific literature. Moreover, false-positive fluorescence signals may rise because of nonspecific hybridization of primers.

Whole-transcriptome amplification technologies enable sequencing of full-length mRNA molecules from hundreds of individual cells

The development of next-generation sequencing [62], and RNA sequencing [1, 2] in particular, paved the way for the development of single-cell RNA sequencing technologies for measuring gene expression and sequence information from hundreds of individual cells. Single-cell RNA sequencing technologies can be roughly divided into two families: protocols that reverse transcribe and amplify full-length mRNA transcripts from single-cell samples to extract full sequence information, and protocols that amplify only the 5′ or 3′ ends of each transcript, with the aim of counting mRNA molecules for measuring gene expression (Table 1). We will start by describing the full transcript length sequencing protocols.

In a typically single-cell full transcript length sequencing experiment, single cells are isolated by Fluorescence-activated cell sorting (FACS) or micropipetting and inserted into individual tubes. After cell lysis, first-strand cDNA synthesis for all mRNA molecules is done by reverse transcriptase using primers that consist of a common sequence that is appended to an oligo-dT tail (Figure 2). To capture full sequence information from each mRNA molecule, it is desired to synthesize cDNA that will cover the full length of the original mRNA transcript, and that will also have well-defined flanking regions to enable priming for PCR pre-amplification. Therefore, before pre-amplification, an additional priming site has to be created at the 3′ end of the newly created first cDNA strand. This can be done either by appending a poly-A tail before second-strand synthesis—as in the Tang and Surani method [63], or by ‘template switching’ at the end of reverse transcription [64, 65]—as in the ‘SMART-seq’ [49] and ‘SMART-Seq2’ methods [29]. Then, PCR pre-amplification can be done using primers complementary to the newly created priming sites on both ends—the first one that was created by the reverse transcription primer and the second one that was

Table 1. Selected single-cell transcriptomic technologies

Method	Number of cells	Number of genes	Information provided	Selected references or Web site
Measuring gene expression of specific genes				
Single-cell qPCR	100–1000 (Dynamic Array: 48 or 96 cells per chip)	Typically 48–96 (up to 280)	Ct values	Benchtop: [32] Microfluidics (Dynamic Array): [16, 18, 20, 33, 35, 59, 60]
Targeted single-cell RNA sequencing	100–1200	150–200	Short sequencing reads covering the 3’ end of preselected transcripts with cell-specific barcodes and UMIs	Benchtop: [73]
CytoSeq	1000–50 000 (100 000 microwells per chip)	150–200	Short sequencing reads covering the 3’ end of preselected transcripts with cell-specific barcodes and UMIs	Microwells: [23]
Full transcript length mRNA sequencing				
Tang and Surani single-cell RNA-seq	100–1000	Whole transcriptome (~25 000)	Short sequencing reads covering full transcript length	Benchtop: [63, 99, 100] Microfluidics: [22]
SMART-seq	100-1,000 (C1: 96 cells per chip)	Whole transcriptome (~25 000)	Short sequencing reads covering full transcript length	Template switching: [64, 65] Tagmentation: [66–69, 101] Benchtop: [28, 49, 71] Microfluidics (C1): [15, 19, 72, 102, 103]
SMART-seq2	100–1000	Whole transcriptome (~25 000)	Short sequencing reads covering full transcript length	Benchtop: [29, 54, 70, 104, 105]
MATQ-seq	100–1000	Whole transcriptome (~25 000)	Short sequencing reads covering full transcript length (with UMIs)	Benchtop: [106]
Molecular tagging and counting strategies				
STRT-Seq	100–1000	Whole transcriptome (~25 000)	Short sequencing reads covering the 5’ end of each transcript with cell-specific barcodes	Benchtop: [75, 107]
STRT/C1-Seq	100–1000 (C1: 96 cells per chip)	Whole transcriptome (~25 000)	Short sequencing reads covering the 5’ end of each transcript with cell-specific barcodes and UMIs	Microfluidics (C1): [47, 84, 87]
CEL-Seq	100–1000	Whole transcriptome (~25 000)	Short sequencing reads covering the 3’ end of each transcript with cell-specific barcodes	Benchtop: [108]
CEL-Seq2	100–1000 (C1: 96 cells per chip)	Whole transcriptome (~25 000)	Short sequencing reads covering the 3’ end of each transcript with cell-specific barcodes and UMIs	Benchtop/microfluidics (C1): [31]

(continued)

Table 1. (continued)

Method	Number of cells	Number of genes	Information provided	Selected references or Web site
MARS-Seq	1000–10 000	Whole transcriptome (~25 000)	Short sequencing reads covering the 3’ end of each transcript with cell-specific barcodes and UMIs	Benchtop: [30,85]
SCRB-Seq	1000–10 000	Whole transcriptome (~25 000)	Short sequencing reads covering the 3’ end of each transcript with cell-specific barcodes and UMI’s	Benchtop: [109]
DropSeq	1000–50 000	Whole transcriptome (~25 000)	Short sequencing reads covering the 3’ end of each transcript with cell-specific barcodes and UMIs	Droplets: [26]
Seq-Well	1000–50 000	Whole transcriptome (~25 000)	Short sequencing reads covering the 3’ end of each transcript with cell-specific barcodes and UMI’s	Microwells: [110]
inDrop	1000–50 000	Whole transcriptome (~25 000)	Short sequencing reads covering the 3’ end of each transcript with cell-specific barcodes and UMIs	Droplets: [27]
Fluidigm HT IFC	1000–10 000 (C1: 800 cells per chip)	Whole transcriptome (~25 000)	Short sequencing reads covering the 3’ end of each transcript with cell-specific barcodes	Microfluidics: www.fluidigm.com
Biorad ddSEQ	1000–50 000	Whole transcriptome (~25 000)	Short sequencing reads covering the 3’ end of each transcript with cell-specific barcodes and UMIs	Droplets: www.bio-rad.com/ddseq
Chromium Single Cell 3’ Solution	1000–50 000	Whole transcriptome (~25 000)	Short sequencing reads covering the 3’ end of each transcript with cell-specific barcodes and UMIs	Droplets: [111] https://www.10xgenomics.com/single-cell/
ICELL8	1000–50 000 (5184 cells per chip)	Whole transcriptome (~25 000)	Short sequencing reads covering the 3’ end of each transcript with cell-specific barcodes and UMIs	Microwells: [112] www.wafergen.com/products/icell8-single-cell-system
SPLiT-seq	10 000–100 000	Whole transcriptome (~25 000)	Short sequencing reads covering the 3’ end of each transcript with cell-specific barcodes and UMI’s	Split-pool barcoding: [78]
sci-RNA-seq	10 000–100 000	Whole transcriptome (~25 000)	Short sequencing reads covering the 3’ end of each transcript with cell-specific barcodes and UMI’s	Split-pool barcoding: [79]

created by template switching/poly-A tailing. The next step is pre-amplification by PCR, after which cDNA from all individual cells can be multiplexed and sequenced.

The input material—also known as 'library'—for second-generation sequencers such as the Illumina HiSeq platform consists of short DNA fragments (typically 100–10 000 bp) flanked by sequencer-specific adaptors on both ends. Standard library preparation protocols usually involve DNA fragmentation, end polishing, adaptor ligation and PCR [1, 2, 62]. As this involves multiple pipetting and cleanup steps, which is infeasible to do for each one of the hundreds of single cells, a more appropriate approach is to use Tn5 transposase-mediated library preparation (Nextera technology) [66–69], a technique in which both DNA fragmentation and adapter insertion are performed in a single ~5 min *in vitro* step ('Tagmentation'; Figure 2). This step is followed by a PCR amplification step that enriches for fragments with correctly inserted adapters and appends full-length sequencing adapters on both sides of each fragment. During this PCR step, primers are used to incorporate a unique sample-specific barcode to all fragments originating from the same cell, thus allowing for multiple (96–384) single-cell libraries to be pooled and sequenced simultaneously in a single sequencer lane (Figure 2).

Libraries can be sequenced single end or paired end. The resulting data consist of short reads (usually 50–150 bp long) covering the full length of each transcript (though there exists some 3′ bias [70]). The reads are then aligned to the reference genome of the organism studied, and a single-cell gene expression matrix can be constructed by counting the number of reads that align to each gene in each individual cell. Detailed sequence information (splice isoforms, allele-specific expression, CNVs and mutations) can be obtained for transcripts with sufficient coverage.

In the past few years, single-cell RNA sequencing has been used to transcriptionally characterize the subpopulation repertoire of complex tissues and tumors [15, 19, 49, 54, 71, 72], as well as rare populations of circulating tumor cells [28]. Two major limitations of single-cell RNA sequencing technologies are the expense of reagents and the long pipetting workflow that includes multiple cleanup steps. This makes these protocols expensive and difficult to automate. Another limitation is the low reverse transcription efficiency (with respect to targeted gene expression protocols that use primers designed for specific genes [73]), which limits the ability to detect lowly expressed genes. To overcome this, microfluidic devices such as the Fluidigm C1 integrated fluidic circuit were developed. These devices can automatically and simultaneously perform multiple (96) single-cell whole-transcription amplification workflows on nanoliter scales, thus allowing for reduced reagent consumption, relatively quick implementation in nonexpert laboratories and increased measurement sensitivity [21, 57]. For example, in the Fluidigm C1 system (Figure 2), once a single-cell suspension with optimal concentration and buoyancy is obtained, all that remains for the user is to insert it along with the appropriate reagents into the chip inlets, and the rest of the cell isolation, lysis and pre-amplification steps are done automatically by the machine. The use of microfluidic chambers of small volume allows for the product of each reaction to be diluted within a larger volume of the following reaction without multiple bead or column-based cleanup steps. After pre-amplification is complete, the single-cell cDNA is 'harvested' from the chip outlets and is ready for subsequent library-preparation steps. However, although microfluidic chips can mitigate cost and labor to some extent, they sometimes have difficulty in trapping small or unusual cells [47], especially from heterogeneous tissues and tumors.

'Molecular tagging' strategies allow for digital counting of mRNA transcripts from whole transcriptomes of many thousands of single cells

A major downside of single-cell qPCR and full transcript length mRNA sequencing is the limited number of cells that can be processed: beyond a few hundreds of cells, the required costs and labor become unfeasible. As the cells of interest (e.g. cancer stem cells) are usually a small fraction (0.1–1%) of the 'bulk' tissue or tumor, many thousands of cells are typically required to obtain a satisfactory number of representative cells. As a result, elaborate enrichment steps (e.g. by FACS) are needed to increase the proportion of the desired cell type. This, however, requires prior knowledge of putative cell markers as well as finding optimal antibodies for them. Another problem is that prior cell-type enrichment creates bias in measuring the proportions of different cells types. To overcome these limitations, 'molecular tagging' strategies were developed to increase the cell throughput to many thousands of single cells, such that even rare cell types will have sufficient representation *in silico* without need for prior cell-type enrichment.

In a typical molecular tagging workflow (Figure 3), specialized primers are used in the first reverse transcription step to append a cell-specific sequence (a 'tag' or 'barcode') to all transcripts originating from each single cell. In the following steps, all barcoded single-cell cDNAs are mixed together in a single tube for pre-amplification and library preparation using standard benchtop pipetting in 100 µl to 2 ml volumes. Libraries are sequenced paired end, where one end is aligned to the reference genome to identify the gene of origin, while the other end is used to recover the cell-specific barcode to identify the cell of origin. After sequencing, the reads of individual cells are demultiplexed *in silico* according to their separate barcodes. Using this strategy, a few thousands of cells can be processed in a single experiment.

There are a number of differences between full transcript length mRNA sequencing and molecular tagging technologies. The first difference is at the physical level: as the number of cells that can be analyzed is much larger in molecular tagging protocols, single-cell capture and isolation is usually done in microfluidic chambers, microwells [23, 24] or droplets [26, 27, 74], which offer much higher cell throughput (although cell picking or FACS sorting into 96 or 384 well plates can also be done [30, 75]). Each compartment contains a single cell and reverse transcription primers, all having the same cell-specific tag. In droplets, this is usually done by encapsulating, within each droplet, a single DNA-barcoded bead along with a single cell. In micro-wells, this can be set up either by loading with DNA-barcoded beads (single bead per well) or by preprinting barcoded primers into each microwell. While it is somewhat faster to create many thousands of droplets, microwells are more suitable for optical imaging, short-term culturing or perturbation of single cells [24].

The second difference between full transcript length mRNA sequencing and molecular tagging technologies is at the molecular level. Both methods contain three main stages: reverse transcription, pre-amplification and library preparation. However, in molecular tagging protocols, an additional cell-specific barcode is appended to the 3′ end of the transcript during reverse transcription, and therefore only library fragments that contain the 3′ end of the original transcript can be used. As

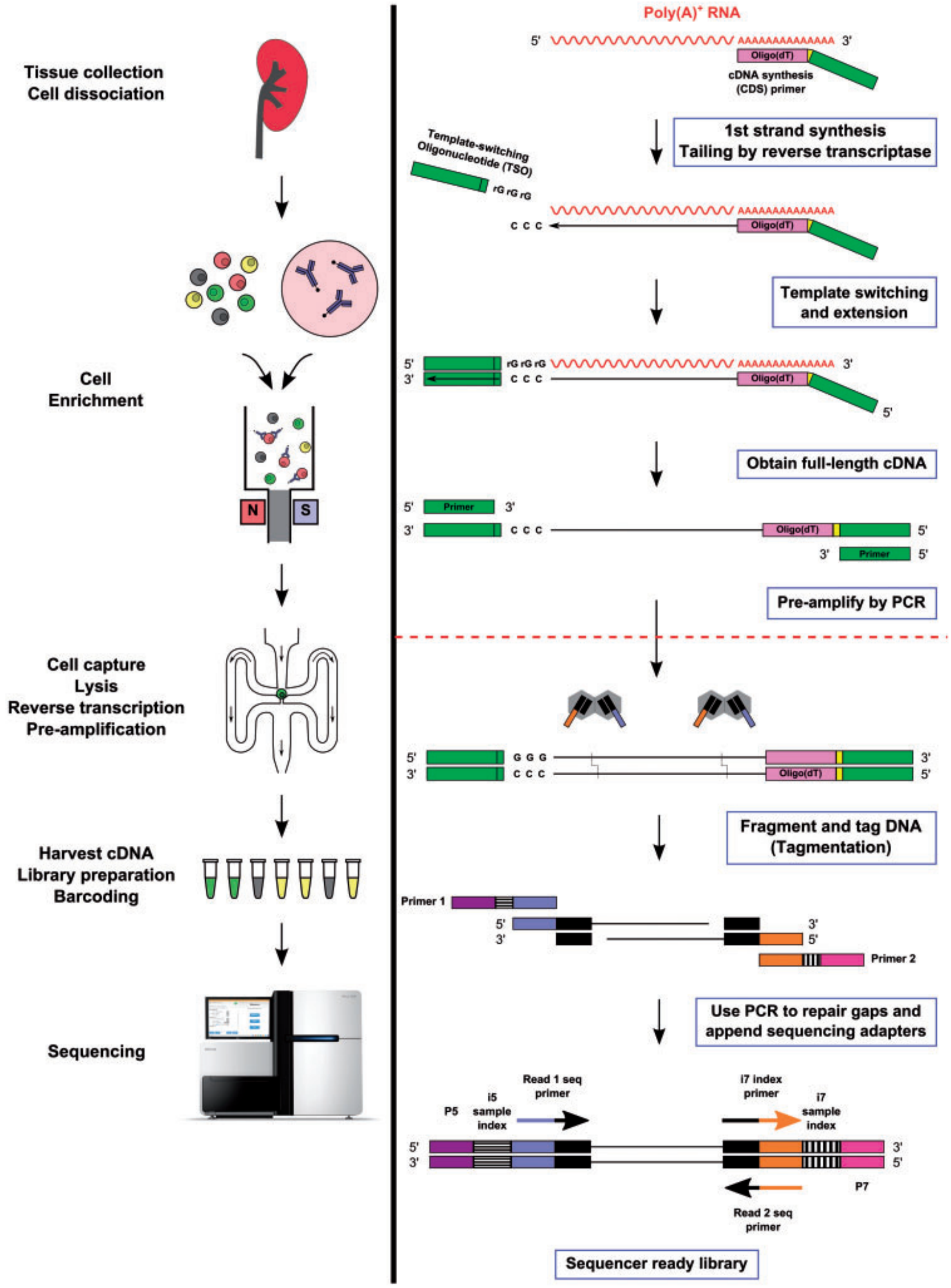


Figure 2. An example for a typical 'SMART-Seq' protocol for full transcript length mRNA sequencing from 96 individual cells using the Fludigm C1 microfluidic system. Left panel: Workflow sketch. Tissues are collected and enzymatically and/or mechanically dissociated. The desired cell type is enriched using cell-type-specific antibodies labeled with magnetic beads. The cell suspension is then inserted into a microfluidic chip, where it is pushed through a series of 96 butterfly-shaped microfluidic traps. Each trap is designed such that once an individual cell is captured, the rest of the suspension flows on to the next trap through a pair of bypass wing-shaped channels. Then, individual cells are isolated from each other, and each cell is lysed. SMART-Seq technology is used to reverse transcribe and pre-amplify the mRNA, after which the amplified cDNA is harvested from the chip into 96 tubes. Library preparation and cell barcoding are done off-chip using the Nextera protocol with 96 different combinations of i5 and i7 barcodes. All 96 libraries are combined and sequenced on a single Illumina HiSeq lane. Right panel: The SMART-Seq protocol at the molecular level. First-strand synthesis is carried out using the MMLV reverse transcriptase in the presence two primers: a cDNA synthesis (CDS) primer that contains an oligo-dT segment, and a template-switching oligo (TSO). When the reverse transcriptase reaches the 5′ end of the mRNA, it adds a few (2–5) C nucleotides to the 3′ end of the newly synthesized cDNA. The TSO, which contains three rG nucleotides at its 3′ end, base pairs with the C-rich tail and the reverse transcriptase 'switches templates' and continues to replicate the TSO. The resulting cDNA strand contains well-defined flanking regions for PCR priming and pre-amplification. After PCR amplification, the cDNA is 'tagmented' using Tn5 transposase-mediated fragmentation and adapter insertion (Nextera), followed by PCR, which appends the full sequencing adapters and incorporates cell-specific barcodes. Typically, 96 single cells are sequenced on a single Illumina Hi-Seq lane at 1–2 million reads per cell (Note: the location of the i5 index primer varies between different Illumina machines).

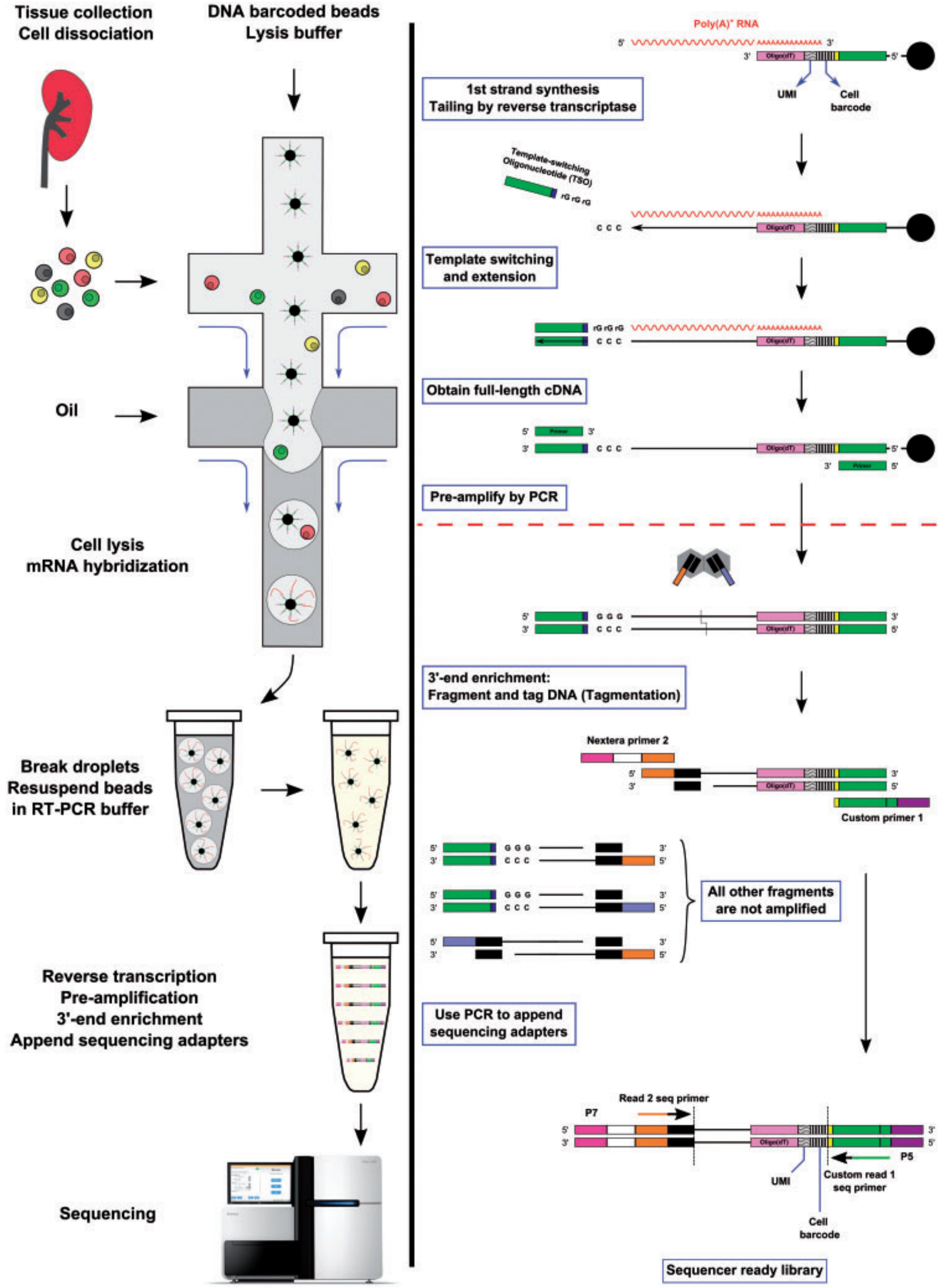
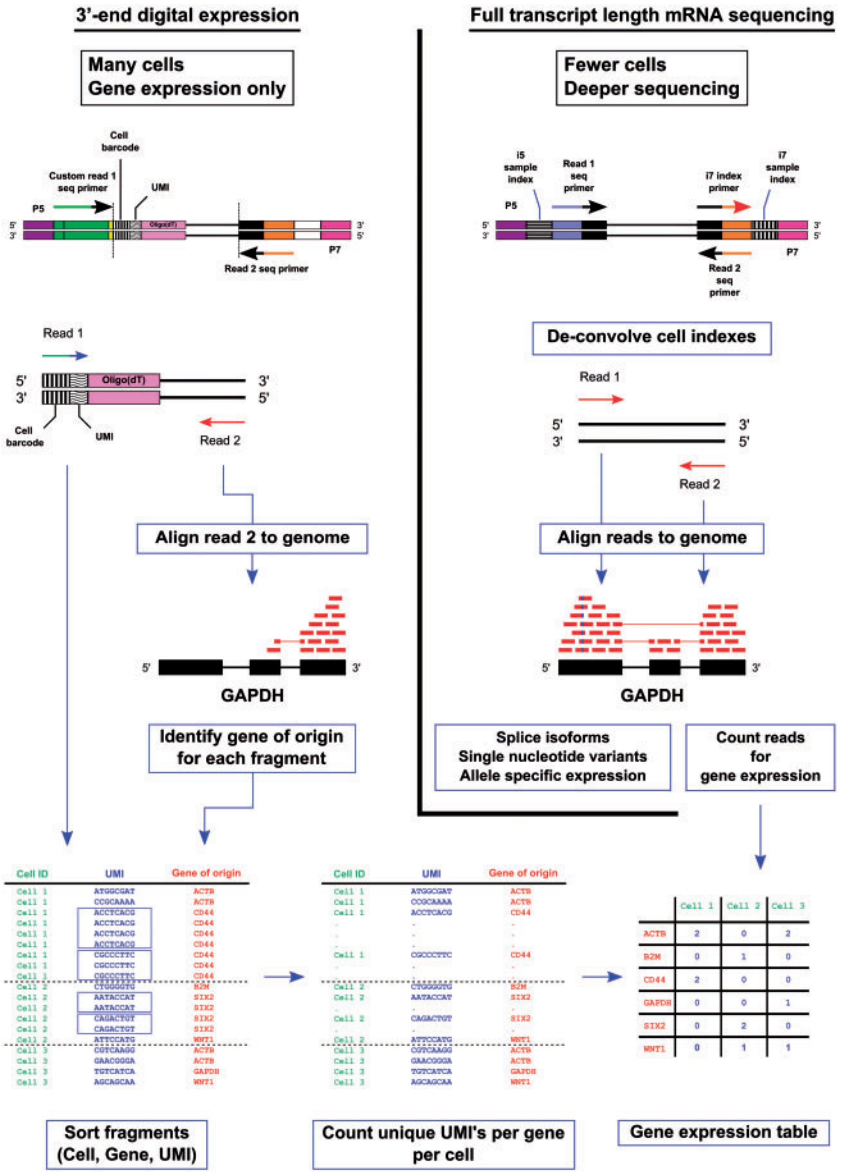


Figure 3. A sketch of the DropSeq protocol [26]—a 3′ tag counting technology. Left panel: Workflow. After tissue dissociation, the cell suspension is injected into a microfluidic device, where it is joined by another flow containing DNA-barcoded beads that are suspended in lysis buffer. Each bead is linked to primers containing a cell barcode (a DNA sequence of 12 bp, which are the same for all primers linked to the same bead), a subsequent UMI (a random sequence of 8 bp) and an oligo-dT segment. Joining a third flow of oil creates an emulsion in which thousands of droplets, many of which contain a single bead and a single cell, are dispersed within the oil. Thus, each such droplet is a distinct compartment in which a single cell is lysed and its mRNA is hybridized to the beads. The droplets are broken by adding a demulsifier to disrupt the water–oil interface, and the beads (along with the hybridized mRNA) are separated from the oil by centrifugation and are resuspended in reverse transcription buffer. The mRNA is then reverse transcribed and pre-amplified. To prepare sequencing-ready cDNA fragments that also contain the cell barcode and UMI, Tn5 transposase-mediated fragmentation and adapter insertion ('tagmentation') is done, followed by PCR, which selects for the '3'-end' fragments and appends the sequencing adapters. Paired-end sequencing is done to extract both the sequence of the gene from which the mRNA was transcribed (Read 2) as well as the cell barcode and UMI (Read 1). Right panel: Molecular biology of the DropSeq protocol. Note that instead of using a standard library preparation step (that is intended to create adapter flanked fragments that cover full transcript length), here a 3′ end enrichment step is performed to choose only those fragments that contain the cell-specific barcode and UMI that were inserted in the first reverse transcription step. This requires using a custom read 1 sequencing primer that contains segments of the reverse transcription primer.

a result, all sequence information further than a few hundreds of bases from the 3′ end is lost. For example, in the library-preparation step of the drop-seq protocol [26], only the 3′ ends of the cDNA fragments are prepared for sequencing. Hence, molecular barcoding protocols aim for counting transcripts (‘3′ tag counting’) for expression measurements rather than for sequencing the entire transcript length. Note that some barcoding methods append the molecular tag to the 5′ end of the transcript rather than the 3′ end [75].

Molecular barcoding can be taken one step further to attach a unique barcode to every single mRNA molecule in each single-cell sample [46–48, 76]. It is thought that most genes are expressed at <1000 mRNA transcripts per single cell. By randomly appending a unique barcode (also called ‘Unique Molecular Identifier’ or UMI) to each transcript at the reverse transcription step, it is possible to discern between sequencing reads originating from different transcripts to those originating from copies of the same transcript that were created by PCR replication (Figures 3–4). In this way, it is possible to directly count the original number of barcoded transcripts and avoid selective sequence-specific bias and inaccuracy at low copy numbers caused by the pre-amplification and library preparation steps. In practice, UMIs are usually appended by inserting a sequence of random bases after the cell-specific sequence tag within the



Cell ID	UMI	Gene of origin
Cell 1	ATGGCGAT	ACTB
Cell 1	CCGCAAAA	ACTB
Cell 1	ACCTCACG	CD44
Cell 1	ACCTCACG	CD44
Cell 1	ACCTCACG	CD44
Cell 1	ACCTCACG	CD44
Cell 1	CGCCCTTC	CD44
Cell 1	CGCCCTTC	CD44
Cell 1	CGCCCTTC	CD44
Cell 2	CTGGGGTG	B2M
Cell 2	AATACCAT	SIX2
Cell 2	AATACCAT	SIX2
Cell 2	CAGACTGT	SIX2
Cell 2	CAGACTGT	SIX2
Cell 2	ATTCCATG	WNT1
Cell 3	CGTCAAGG	ACTB
Cell 3	GAACGGGA	ACTB
Cell 3	TGTCATCA	GAPDH
Cell 3	AGCAGCAA	WNT1

Cell ID	UMI	Gene of origin
Cell 1	ATGGCGAT	ACTB
Cell 1	CCGCAAAA	ACTB
Cell 1	ACCTCACG	CD44
.	.	.
.	.	.
.	.	.
Cell 1	CGCCCTTC	CD44
.	.	.
.	.	.
Cell 2	CTGGGGTG	B2M
Cell 2	AATACCAT	SIX2
.	.	.
Cell 2	CAGACTGT	SIX2
.	.	.
Cell 2	ATTCCATG	WNT1
Cell 3	CGTCAAGG	ACTB
Cell 3	GAACGGGA	ACTB
Cell 3	TGTCATCA	GAPDH
Cell 3	AGCAGCAA	WNT1

	Cell 1	Cell 2	Cell 3
ACTB	2	0	2
B2M	0	1	0
CD44	2	0	0
GAPDH	0	0	1
SIX2	0	2	0
WNT1	0	1	1

Figure 4. Single-cell RNA sequencing data analysis. Right panel: Full transcript length sequencing—libraries covering the full transcript length are sequenced either single end or paired end. Reads are demultiplexed according to cell-specific barcodes on both ends. For each cell, reads are aligned to the reference genome. The number of reads that align to a particular gene locus is an indication for its expression level. Similarly, splice isoform expression and other sequence information can be inferred. Left panel: An example for 3′ end digital expression. Sequencing library fragments covering the 3′ end of the transcript are sequenced paired end. Read 1 contains a cell-specific barcode (or ‘tag’) and a unique transcript identifier (UMI). Read 2 is aligned to the reference genome to identify the specific gene from which the transcript originated. In a simplified algorithm, each library fragment is represented by a textual string containing a unique cell ID and UMI (from read 1) and a gene of origin (from read 2). The fragments are lexicographically sorted in three levels: first according to the cell ID, then according to the gene of origin and then according to the UMI. Then, for each individual cell, the number of unique UMIs for each gene is counted. This better represents the original number of transcripts.

reverse transcription primer [26] (or template-switch oligo [75]). The expression level of each gene is estimated by counting the number of reads with distinct UMIs that align to the position of that gene within the genome (Figure 4). To properly count the number of molecules, the number of distinct UMIs must be larger than the number of molecules that are to be counted. For example, a 5 bp long UMI can distinguish between 4^5 = 1024 molecules, which is more than the typical number of transcripts originating from most genes. UMIs provide a more 'digital' way for measuring gene expression (often called '3'-end digital expression'), and were found to significantly increase precision [47] (in the sense of reproducibility on replicate samples) and reduce technical variability in single-cell RNA sequencing experiments [77].

The combination of 'molecular tagging' strategies and automated droplet/microwell technologies has led to an increase in the number of individual cells that can be analyzed. Current protocols can process 10 000–50 000 individual cells within a few days. Recently developed 'split-pool' barcoding strategies that use the cells themselves as 'containers' can be used to increase throughput to millions of cells [78, 79]. This can provide unbiased counting of cell types in complex tissues and tumors, which usually contain cell subpopulations whose sizes span several orders of magnitude. This capability is critical for the discovery of new rare cell types such as cancer stem cells without requiring prior knowledge of markers for prior cell enrichment. The drawback of molecular tagging techniques is that with short-read sequencing technology, most transcript sequence information such as splice isoforms, SNVs or allele-specific expression, is lost (Figure 4).

One difficulty encountered in molecular tagging techniques is the occurrence of cell doublets within a single compartment (chamber/microwell/droplet), which may occur because of cells sticking together or because of imperfections in the mechanism of partitioning [26]. While cell doublets are encountered at low frequency (~1%) in almost all single-cell technologies, the numbers can become significant when processing many thousands of cells. This may be mitigated by careful calibration of the input cell concentration, as well as by automatic scanning microscopy and image processing algorithms to identify and exclude doublets from the analysis. In addition, contaminating RNA, presumably originating from cells that were damaged during the preparation of the single-cell suspension, can confound single-cell measurements. This may be overcome by careful tissue collection and dissociation, by using digestive enzymes optimized for each particular tissue and, when possible, by using a wash step after cell trapping. Finally, the number of UMIs is often overestimated because of PCR amplification errors. This can be solved by merging together similar UMIs and treating them as a single count [80].

Data analysis: inferring meaningful biological insight from single-cell RNA sequencing measurements

A major challenge in analyzing single-cell RNA sequencing data is the large technical noise due to the small amount of input material and low transcript detection efficiencies. This results in significant variability because of random Poisson 'sampling' of mRNA molecules, as well as variability in sequencing efficiencies across different cells. Using spike-in RNA standards [81, 82], these inherent noise sources can be measured, modeled and de-convolved [77, 83].

Unlike methods that are designed to target specific transcripts (such as microfluidic single-cell qPCR), single-cell RNA sequencing data are often sparse because of the fact that many genes—especially those expressed at low levels—are not detected. As a result, it can be computationally challenging to identify the repertoire of cell subpopulations as well as the genes that differentiate between them. A straightforward method is to choose a subset of previously known genes or transcription factors relevant to the biology of the specific tissue or tumor [30], and to use clustering algorithms to identify meaningful cell types according to the expression of these genes across the different cells. A more unbiased approach is to perform principal component analysis and choose genes with maximal loadings in the first principal components, or to use the representation of the gene expression matrix along the first principal components - those that explain most of the cellular variability - for further analysis [19, 26]. Yet, another approach (which can be used in combination with the rest) is to choose only genes whose variance exceeds that of a baseline Poisson-like noise level, with the underlying assumption that the variability in these genes arises because of active upregulation or downregulation between the different cell types rather than random sampling noise [26, 84]. It is sometimes also helpful to filter out genes whose expression does not correlate to any other gene [84].

Some notable clustering algorithms that can be used for identifying cell populations from single-cell expression data are *K*-means clustering [85], hierarchical clustering [19], biclustering [84, 86, 87], affinity propagation [84, 88], density-based spatial clustering (e.g. DBSCAN) [26, 89], modeling as a mixture of multivariate distributions [90, 91] and network-based community detection algorithms [75, 92]. In many cases, it is useful to use methods such as tSNE [93] for embedding high-dimensional data in two or three dimensions for visualization and further analysis [26]. Once cell subpopulations have been identified, their transcriptomes can be combined to form 'pooled' transcriptomes for deeper characterization [30] to identify new markers and gene circuits.

Often, there are many cells whose type cannot be completely determined and these create a continuum between 'seed' cell types [30, 94]. For example, the cell cycle state is often inferred from correlating to gene sets known to be upregulated in different phases of the cell cycle [26, 54]. In developing or differentiating biological systems, where time dependence is relevant, algorithms have been developed to infer a 'pseudo-time'—a temporal trajectory according to which the cells can be ordered *in silico* [95].

Outlook

During the past decade, there has been an 'explosion' of new single-cell technologies, each having its own capabilities and limitations. The main challenges that lie ahead are improving the detection yields—mainly through improvement of reverse transcription and pre-amplification efficiencies, as well as finding ways to retain spatial information, i.e. the arrangement of cells within the tissue/tumor and the location of each transcript within each cell [96–98]. Moreover, there are considerable challenges in profiling single cells from paraffin-embedded tissues. There are also considerable challenges in developing computational tools for analysis and visualization of the huge single-cell data sets that are being produced.

To date, we believe that a combined approach is most appropriate for dissecting tissues and tumors: large-scale 'molecular tagging' techniques should be used to identify cell types from many thousands of cells and find markers for further enrichment. Then, medium-scale methods for sequencing full transcript lengths from hundreds of cells can be used for deeper analysis of selected cell populations, for example, to identify novel splice isoforms, allele-specific expression, CNVs, RNA editing events and point mutations. Finally, target-specific methods such as single-cell qPCR and mRNA-FISH (Fluorescence *in situ* hybridization) can be used for more exact measurement, validation and spatial localization of specific targets of interest such as transcription factors.

Key Points

- Recently developed single-cell transcriptomic technologies enable high-resolution analysis of complex biological systems such as a developing embryo, a regenerating tissue or a tumor.
- Single-cell qPCR can be used to measure expression of multiple selected genes in hundreds of individual cells with high sensitivity and precision.
- Whole-transcriptome amplification technologies enable sequencing of full-length mRNA molecules from hundreds of individual cells.
- 'Molecular tagging' strategies allow for digital counting of mRNA transcripts from whole transcriptomes of many thousands of single cells.
- The main challenges that lie ahead are improving transcript detection yields and retaining spatial information.

Acknowledgements

We wish to thank Aaron Streets, Paul Blainey, Josh Weinstein, Achia Urbach, and all members of our labs for fruitful comments and suggestions.

Funding

T.K., S.O., T.H.B.L, N.B.H, A.T., Y.W., and I.K. are supported by the Israel Science Foundation (ICORE no. 1902/12 and Grants no. 1634/13 and 2017/13), the Israel Cancer Association (Grant no. 20150911), the Israel Ministry of Health (Grant no. 3-10146), and the EU-FP7 (Marie Curie International Reintegration Grant no. 618592). S.P. is supported by Ramalingaswami Fellowship from the Department of Biotechnology, India.

References

1. Mortazavi A, Williams BA, McCue K, *et al*. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 2008;**5**:621–8.
2. Nagalakshmi U, Wang Z, Waern K, *et al*. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 2008;**320**:1344–9.
3. Li C, Heidt DG, Dalerba P, *et al*. Identification of pancreatic cancer stem cells. *Cancer Res* 2007;**67**:1030–7.
4. Al-Hajj M, Wicha MS, Benito-Hernandez A, *et al*. Prospective identification of tumorigenic breast cancer cells. *Proc Natl Acad Sci USA* 2003;**100**:3983–8.
5. Dalerba P, Dylla SJ, Park I-K, *et al*. Phenotypic characterization of human colorectal cancer stem cells. *Proc Natl Acad Sci USA* 2007;**104**:10158–63.
6. Prince ME, Sivanandan R, Kaczorowski A, *et al*. Identification of a subpopulation of cells with cancer stem cell properties in head and neck squamous cell carcinoma. *Proc Natl Acad Sci USA* 2007;**104**:973–8.
7. Pode-Shakked N, Shukrun R, Mark-Danieli M, *et al*. The isolation and characterization of renal cancer initiating cells from human Wilms' tumour xenografts unveils new therapeutic targets. *EMBO Mol Med* 2013;**5**:18–37.
8. Singh SK, Hawkins C, Clarke ID, *et al*. Identification of human brain tumour initiating cells. *Nature* 2004;**432**:396–401.
9. Bonnet D, Dick JE. Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. *Nat Med* 1997;**3**:730–7.
10. Bussolati B, Dekel B, Azzarone B, *et al*. Human renal cancer stem cells. *Cancer Lett* 2013;**338**:141–6.
11. Shackleton M, Vaillant F, Simpson KJ, *et al*. Generation of a functional mammary gland from a single stem cell. *Nature* 2006;**439**:84–8.
12. Stingl J, Eirew P, Ricketson I, *et al*. Purification and unique properties of mammary epithelial stem cells. *Nature* 2006; **439**:993–7.
13. Barker N, van Es JH, Kuipers J, *et al*. Identification of stem cells in small intestine and colon by marker gene Lgr5. *Nature* 2007;**449**:1003–7.
14. Kobayashi A, Valerius MT, Mugford JW, *et al*. Six2 Defines and Regulates a Multipotent Self-Renewing Nephron Progenitor Population throughout Mammalian Kidney Development. *Cell Stem Cell* 2008;**3**:169–81.
15. Proserpio V, Piccolo A, Haim-Vilmovsky L, *et al*. Single-cell analysis of CD4+ T-cell differentiation reveals three major cell states and progressive acceleration of proliferation. *Genome Biol* 2016;**17**:103.
16. Dalerba P, Kalisky T, Sahoo D, *et al*. Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nat Biotechnol* 2011;**29**:1120–7.
17. Brunskill EW, Park J-S, Chung E, *et al*. Single cell dissection of early kidney development: multilineage priming. *Development* 2014;**141**:3093–101.
18. Rothenberg ME, Nusse Y, Kalisky T, *et al*. Identification of a cKit(+) colonic crypt base secretory cell that supports Lgr5(+) stem cells in mice. *Gastroenterology* 2012;**142**: 1195–1205.e6.
19. Treutlein B, Brownfield DG, Wu AR, *et al*. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* 2014;**509**:371–5.
20. Guo G, Huss M, Tong GQ, *et al*. Resolution of Cell Fate Decisions Revealed by Single-Cell Gene Expression Analysis from Zygote to Blastocyst. *Dev Cell* 2010;**18**:675–85.
21. Wu AR, Neff NF, Kalisky T, *et al*. Quantitative assessment of single-cell RNA-sequencing methods. *Nat Methods* 2014;**11**: 41–6.
22. Streets AM, Zhang X, Cao C, *et al*. Microfluidic single-cell whole-transcriptome sequencing. *Proc Natl Acad Sci USA* 2014;**111**:7048–53.
23. Fan HC, Fu GK, Fodor SP a. Combinatorial labeling of single cells for gene expression cytometry. *Science* 2015;**347**: 1258367.
24. Yuan J, Sims PA. An Automated Microwell Platform for Large-Scale Single Cell RNA-Seq. *Sci Rep* 2016;**6**:33883.
25. Bose S, Wan Z, Carr A, *et al*. Scalable microfluidics for single-cell RNA printing and sequencing. *Genome Biol* 2015;**16**:120.
26. Macosko EZ, Basu A, Satija R, *et al*. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* 2015;**161**:1202–14.

27. Klein AM, Mazutis L, Akartuna I, *et al*. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* 2015;**161**:1187–201.
28. Ramsköld D, Luo S, Wang Y-C, *et al*. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat Biotechnol* 2012;**30**:777–82.
29. Picelli S, Faridani OR, Björklund AK, *et al*. Full-length RNA-seq from single cells using Smart-seq2. *Nat Protoc* 2014;**9**: 171–81.
30. Jaitin DA, Kenigsberg E, Keren-Shaul H, *et al*. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* 2014;**343**:776–9.
31. Hashimshony T, Senderovich N, Avital G, *et al*. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol* 2016;**17**:77.
32. Bengtsson M, Ståhlberg A, Rorsman P, *et al*. Gene expression profiling in single cells from the pancreatic islets of Langerhans reveals lognormal distribution of mRNA levels. *Genome Res* 2005;**15**:1388–92.
33. Sanchez-Freire V, Ebert AD, Kalisky T, *et al*. Microfluidic single-cell real-time PCR for comparative analysis of gene expression patterns. *Nat Protoc* 2012;**7**:829–38.
34. Wills QF, Livak KJ, Tipping AJ, *et al*. Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments. *Nat Biotechnol* 2013;**31**:748–52.
35. Guo G, Luc S, Marco E, *et al*. Mapping cellular hierarchy by single-cell analysis of the cell surface repertoire. *Cell Stem Cell* 2013;**13**:492–505.
36. Harding SD, Armit C, Armstrong J, *et al*. The GUDMAP database–an online resource for genitourinary research. *Development* 2011;**138**:2845–53.
37. Uhlén M, Fagerberg L, Hallström BM, *et al*. Tissue-based map of the human proteome. *Science* 2015;**347**:1260419.
38. Trapnell C, Pachter L, Salzberg SL. TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* 2009;**25**: 1105–11.
39. Kim D, Pertea G, Trapnell C, *et al*. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* 2013;**14**:R36.
40. Langmead B, Trapnell C, Pop M, *et al*. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 2009;**10**:R25.
41. Dobin A, Davis CA, Schlesinger F, *et al*. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* 2013;**29**:15–21.
42. Anders S, Pyl PT, Huber W. HTSeq-A Python framework to work with high-throughput sequencing data. *Bioinformatics* 2015;**31**:166–69.
43. Anders S, McCarthy DJ, Chen Y, *et al*. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nat Protoc* 2013;**8**:1765–86.
44. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. bioRxiv 2014;
45. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol* 2010;**11**:R106.
46. Fu GK, Hu J, Wang P-H, *et al*. Counting individual DNA molecules by the stochastic attachment of diverse labels. *Proc Natl Acad Sci USA* 2011;**108**:9026–31.
47. Islam S, Zeisel A, Joost S, *et al*. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat Methods* 2014;**11**: 163–6.
48. Kivioja T, Vähärautio A, Karlsson K, *et al*. Counting absolute numbers of molecules using unique molecular identifiers. *Nat Methods* 2011;**9**:72–4.
49. Shalek AK, Satija R, Adiconis X, *et al*. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* 2013;**498**:236–40.
50. Anders S, Reyes A, Huber W. Detecting differential usage of exons from RNA-seq data. *Genome Res* 2012;**22**:2008–17.
51. Katz Y, Wang ET, Airoldi EM, *et al*. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods* 2010;**7**:1009–15.
52. Shen S, Park JW, Lu Z, *et al*. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc Natl Acad Sci USA* 2014;**111**:E5593-601.
53. Deng Q, Ramsköld D, Reinius B, *et al*. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* 2014;**343**:193–6.
54. Tirosh I, Venteicher AS, Hebert C, *et al*. Large-scale single-cell RNA-seq reveals a developmental hierarchy in human oligodendroglioma. *Nature* 2016;**539**:309–13.
55. Picelli S. Single-cell RNA-sequencing: The future of genome biology is now. *RNA Biol* 2016;**6286**:1–14.
56. Kolodziejczyk AA, Kim JK, Svensson V, *et al*. The Technology and Biology of Single-Cell RNA Sequencing. *Mol Cell* 2015;**58**: 610–20.
57. Svensson V, Natarajan KN, Ly L-H, *et al*. Power analysis of single-cell RNA-sequencing experiments. *Nat Methods* 2017;
58. Svensson V, Roser V-T, Teichmann SA. Moore's Law in Single Cell Transcriptomics. 2017; arXiv:1704.01379v1 [q-bio.GN].
59. Livak KJ, Wills QF, Tipping AJ, *et al*. Methods for qPCR gene expression profiling applied to 1440 lymphoblastoid single cells. *Methods* 2013;**59**:71–9.
60. Liu J, Hansen C, Quake SR. Solving the 'World-to-Chip' interface problem with a microfluidic matrix. *Anal Chem* 2003;**75**: 4718–23.
61. Heid CA, Stevens J, Livak KJ, *et al*. Real time quantitative PCR. *Genome Res* 1996;**6**:986–94.
62. Bentley DR, Gormley NA, Balasubramanian S, *et al*. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 2008;**456**:53–9.
63. Tang F, Barbacioru C, Nordman E, *et al*. RNA-Seq analysis to capture the transcriptome landscape of a single cell. *Nat Protoc* 2010;**5**:516–35.
64. Zhu YY, Machleder EM, Chenchik A, *et al*. Reverse transcriptase template switching: A SMART approach for full-length cDNA library construction. *Biotechniques* 2001;**30**:892–97.
65. Matz M, Shagin D, Bogdanova E, *et al*. Amplification of cDNA ends based on template-switching effect and step-out PCR. *Nucleic Acids Res* 1999;**27**:1558–60.
66. Adey A, Morrison HG, Asan, *et al*. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biol* 2010;**11**:R119.
67. Reznikoff WS. Tn 5 as a model for understanding DNA transposition. *Mol Biol* 2003;**47**:1199–206.
68. Picelli S, Björklund AK, Reinius B, *et al*. Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res* 2014;**24**:2033–40.
69. Reznikoff WS. Transposon Tn 5. *Annu Rev Genet* 2008;**42**:269–86.
70. Picelli S, Björklund ÅK, Faridani OR, *et al*. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat Methods* 2013;**10**:1096–8.
71. Patel AP, Tirosh I, Trombetta JJ, *et al*. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* 2014;**344**:1396–401.
72. Chen S, Brunskill EW, Potter SS, *et al*. Intrinsic Age-Dependent Changes and Cell-Cell Contacts Regulate Nephron Progenitor Lifespan. *Dev Cell* 2015;**35**:49–62.

73. Fu GK, Wilhelmy J, Stern D, *et al*. Digital encoding of cellular mRNAs enabling precise and absolute gene expression measurement by single-molecule counting. *Anal Chem* 2014; **86**:2867–70.
74. Guo MT, Rotem A, Heyman JA, *et al*. Droplet microfluidics for high-throughput biological assays. *Lab Chip* 2012;**12**:2146–55.
75. Islam S, Kjällquist U, Moliner A, *et al*. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res* 2011;**21**:1160–67.
76. Shiroguchi K, Jia TZ, Sims PA, *et al*. Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. *Proc Natl Acad Sci USA* 2012;**109**:1347–52.
77. Grün D, Kester L, van Oudenaarden A. Validation of noise models for single-cell transcriptomics. *Nat Methods* 2014;**11**: 637–40.
78. Rosenberg AB, Roco C, Muscat RA, *et al*. Scaling single cell transcriptomics through split pool barcoding. bioRxiv 2017; 105163.
79. Heitland I, Hermann A, Kuhn M, *et al*. Comprehensive single cell transcriptional profiling of a multicellular organism by combinatorial indexing. bioRxiv 2017; 1–90.
80. Smith TS, Sudbery I, Heger A, *et al*. UMI-tools: Modelling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. bioRxiv 2016.
81. Jiang L, Schlesinger F, Davis CA, *et al*. Synthetic spike-in standards for RNA-seq experiments. *Genome Res* 2011;**21**: 1543–51.
82. Hardwick SA, Chen WY, Wong T, *et al*. Spliced synthetic genes as internal controls in RNA sequencing experiments. *Nat Meth* 2016;**13**:792–98.
83. Brennecke P, Anders S, Kim JK, *et al*. Accounting for technical noise in single-cell RNA-seq experiments. *Nat Methods* 2013; **10**:1093–5.
84. Zeisel A, Manchado a. BM, Codeluppi S, *et al*. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* 2015;**347**:1138–42.
85. Matcovitch-Natan O, Winter DR, Giladi A, *et al*. Microglia development follows a stepwise program to regulate brain homeostasis. *Science* 2016;**353**:aad8670.
86. Getz G, Levine E, Domany E. Coupled two-way clustering of DNA microarray data. *Proc Natl Acad Sci USA* 2000;**97**: 12079–84.
87. La Manno G, Gyllborg D, Codeluppi S, *et al*. Molecular diversity of midbrain development in mouse, human and stem cells. *Cell* 2016;**167**:566–580.e19.
88. Frey BJ, Dueck D. Clustering by passing messages between data points. *Science* 2007;**315**:972–76.
89. Ester M, Kriegel HP, Sander J, *et al*. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Proc 2nd Int Conf Knowl Discov Data Min* 1996; 226–31.
90. Pyne S, Hu X, Wang K, *et al*. Automated high-dimensional flow cytometric data analysis. *Proc Natl Acad Sci USA* 2009; **106**:8519–24.
91. Pyne S, Lee SX, Wang K, *et al*. Joint modeling and registration of cell populations in cohorts of high-dimensional flow cytometric data. *PLoS One* 2014;**9**.
92. Levine JH, Simonds EF, Bendall SC, *et al*. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell* 2015;**162**:184–97.
93. Van Der Maaten LJP, Hinton GE. Visualizing high-dimensional data using t-SNE. *J Mach Learn Res* 2008;**9**:2579–605.
94. Korem Y, Szekely P, Hart Y, *et al*. Geometry of the Gene Expression Space of Individual Cells. *PLOS Comput Biol* 2015; **11**:e1004224.
95. Trapnell C, Cacchiarelli D, Grimsby J, *et al*. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* 2014;**32**: 381–6.
96. Satija R, Farrell JA, Gennert D, *et al*. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* 2015;**33**: 495–502.
97. Lee JHJH, Daugharthy ER, Scheiman J, *et al*. Highly Multiplexed Subcellular RNA Sequencing in Situ. *Science* 2014;**343**:1360–63.
98. Lee JH, Daugharthy ER, Scheiman J, *et al*. Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. *Nat Protoc* 2015;**10**:442–58.
99. Tang F, Barbacioru C, Wang Y, *et al*. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* 2009;**6**: 377–82.
100. Tang F, Barbacioru C, Bao S, *et al*. Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-seq analysis. *Cell Stem Cell* 2010;**6**:468–78.
101. Gertz J, Varley KE, Davis NS, *et al*. Transposase mediated construction of RNA-seq libraries. *Genome Res* 2012;**22**:134–41.
102. Pollen AA, Nowakowski TJ, Shuga J, *et al*. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat Biotechnol* 2014;**32**:1053–8.
103. Martinez-jimenez CP, Eling N, Chen H, *et al*. Aging increases cell-to-cell transcriptional variability upon immune stimulation. *Science* 2017;**1436**:1433–36.
104. Villani A-C, Satija R, Reynolds G, *et al*. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* 2017;**356**:eaah4573.
105. Venteicher AS, Tirosh I, Hebert C, *et al*. Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq. *Science* 2017;**355**:eaai8478.
106. Sheng K, Cao W, Niu Y, *et al*. Effective detection of variation in single-cell transcriptomes using MATQ-seq. *Nat Methods* 2017;**14**:267–70.
107. Islam S, Kjällquist U, Moliner A, *et al*. Highly multiplexed and strand-specific single-cell RNA 5' end sequencing. *Nat Protoc* 2012;**7**:813–28.
108. Hashimshony T, Wagner F, Sher N, *et al*. CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification. *Cell Rep* 2012; **2**:666–73.
109. Soumillon M, Cacchiarelli D, Semrau S, *et al*. Characterization of directed differentiation by high-throughput single-cell RNA-Seq. bioRxiv 2014; 3236.
110. Gierahn TM, Wadsworth MH, Hughes TK, *et al*. Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nat Methods* 2017;**14**:395–98.
111. Zheng GXY, Terry JM, Belgrader P, *et al*. Massively parallel digital transcriptional profiling of single cells. bioRxiv 2016; 8:65912.
112. Goldstein LD, Chen Y-JJ, Dunne J, *et al*. Massively parallel nanowell-based single-cell gene expression profiling. *BMC Bioinformatics* 2017; 1–10.