A BRIEF REVIEW OF SINGLE CELL TRANSCRIPTOMIC TECHNOLOGIES

Tomer Kalisky¹, Sarit Oriel¹, Tali Hana Bar-Lev¹, Nissim Ben-Haim¹, Ariel Trink¹, Yishay Wineberg¹, Itamar Kanter¹, Shlomit Gilad², and Saumyadipta Pyne³

¹Faculty of Engineering and Bar-Ilan Institute of Nanotechnology and Advanced Materials (BINA), Bar-Ilan University, Ramat Gan 5290002, Israel.

²The Israel National Center for Personalized Medicine (INCPM), Weizmann Institute of Science, Rehovot, Israel.

³Indian Institute of Public Health, Kavuri Hills, Madhapur, Hyderabad, India.

*These authors share equal senior co-authorship and correspondence

**Corresponding authors:
Tomer Kalisky
Email: Tomer.bioengineering@gmail.com
Affiliation: Faculty of Engineering and Bar-Ilan Institute of Nanotechnology and Advanced Materials (BINA), Bar-Ilan University, Ramat Gan 5290002, Israel.

Saumyadipta Pyne
Email: spyne@iiphh.org
Affiliation: Indian Institute of Public Health, Kavuri Hills, Madhapur, Hyderabad, India.
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 datasets 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.

SUMMARY 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.

KEY WORDS:
Single cell technologies, Genomics, RNA sequencing
AUTHOR BIOGRAPHY:

Tomer Kalisky (Ph.D) 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 (Ph.D) 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 Bar-Lev (Ph.D) 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. His 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 (Ph.D) graduated from the Department of Bioengineering, Bar-Ilan University, Israel. He is interested in machine learning and in developing networks based algorithms for analysis of single cell RNA sequencing datasets.

Shlomit Gilad (Ph.D) is Head of the R&D for Genomics in the Nancy & Stephen Grand Israel National Center for Personalized Medicine (G-INCPM), Israel. Her research interest lie in genomics, in particular, developing new technologies for high throughput sequencing for both bulk samples and single cells.

Saumyadipta Pyne (Ph.D) 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.
INTRODUCTION: THE RATIONALE OF SINGLE CELL ANALYSIS

Why single cells?

Over the last 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 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 less than 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, since RNA molecules are mostly very unstable, individual cells must be carefully collected and rapidly isolated from the tissue or tumor in order to keep the cells as viable as possible. Second, since a single cell contains roughly 10pg of total RNA, which is much smaller than the nano-gram 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, since the number of measurements is enormous (number of cells x number of genes x number of transcripts), elaborate multiplexing strategies such as microfluidic parallelization, droplet encapsulation, and molecular barcoding (tagging) must be employed 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 sub-population is extremely rare (for example, 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 2-3 fluorescently labeled
antibodies or reporter genes [18]. Another option is to mix the cells with magnetic micro-beads 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 micro-wells [23–25], or by encapsulating in nanolitter 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, dNTP’s, 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 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 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, in order 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 sub-population 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 pre-selecting 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 sub-populations, and data from all cells belonging to the same sub-population can be mixed in-silico to provide a deeper characterization of each sub-population [30]. Since 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 sub-population [19]. Moreover, sequence data can reveal splicing patterns [28,49–52], allele-specific expression [53], single nucleotide variations (SNV’s) [54], and copy number variations (CNV’s) [54].
SINGLE CELL TRANSCRIPTOMIC TECHNOLOGIES

In the last 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

Quantitative real-time PCR (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 (for example, 100 genes x 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 (Fig. 1), individual cells are first sorted by a flow cytometer into individual wells of 96 well plates pre-filled with PCR buffer [16,33]. Cells are lysed by influx of fluids through their cell membranes due to hypotonic pressure. Then, the RNA is reverse transcribed and PCR-amplified for 14-20 cycles using a mixture of primers for a pre-determined 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.

The resulting data is in the form of a matrix of threshold cycles (Ct's) 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 standard deviation) and partitioned into groups using clustering algorithms such as K-means or hierarchical clustering. Cell sub-populations 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) since each single cell is inherently normalized.

Since microfluidic single cell qPCR is based on reverse transcription and amplification of specific pre-selected 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 multi-gene picture of the cell sub-population 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 due to non-specific 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 (e.g. [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 in order 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 mRNA transcript length** sequencing experiment, single cells are isolated by FACS (Fluorescence-activated cell sorting) or micro-pipetting and inserted into individual tubes. After cell lysis, 1st 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 (Fig. 2). In order 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, prior to 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 2nd 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 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 2nd 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]. Since 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 TnS transposase-mediated library preparation (Nextera technology) [66–69], a technique in which both DNA fragmentation and adapter insertion are performed in a single ~5-minute in-vitro step (“Tagmentation”; Fig. 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 (Fig. 2).
Libraries can be sequenced single-end or paired-end. The resulting data consists of short reads (usually 50-150bp 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, copy number variations (CNV's), and mutations) can be obtained for transcripts with sufficient coverage.

In the last few years single cell RNA sequencing has been used to transcriptionally characterize the sub-population 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 (IFC) were developed. These devices can automatically and simultaneously perform multiple (96) single cell whole transcription amplification workflows on nano-liter scales, thus allowing for reduced reagent consumption, relatively quick implementation in non-expert labs, and increased measurement sensitivity [21,57]. For example, in the Fluidigm C1 system (Fig. 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. Since 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 (for example by FACS) are needed in order 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. In order to overcome these limitations, “molecular tagging” strategies were developed in order 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 (Fig. 3), specialized primers are used in the very 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 cDNA’s are mixed together in a single tube for pre-amplification and library preparation using standard benchtop pipetting in 100ul-2ml volumes. Libraries are sequenced paired-end, where one end is aligned to the reference genome in order to identify the gene of origin, while the other end is used to recover the cell-specific barcode in order to identify the cell of origin. After sequencing, the reads of individual cells are de-multiplexed 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: since 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, micro-wells [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 pre-printing barcoded primers into each micro-well. While it is somewhat faster to create many thousands of droplets, micro-wells 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 3 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 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 less than 1,000 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 (Figs. 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, UMI’s are usually appended by inserting a sequence of random bases after the cell-specific sequence tag within the 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 UMI’s that align to the position of that gene within the genome (Fig. 4). In order to properly count the number of molecules, the number of distinct UMI’s 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=1,024 molecules, which is more than the typical number of
transcripts originating from most genes. UMI’s 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/micro-well 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, single nucleotide variations (SNV’s), or allele specific expression, is lost (Fig. 4).

One difficulty encountered in molecular tagging techniques is the occurrence of cell doublets within a single compartment (chamber/microwell/droplet), which may occur due to cells sticking together or due to 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 employing a wash step after cell trapping. Finally, the number of UMI’s is often over-estimated due to PCR amplification errors. This can be solved by merging together very similar UMI’s 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, modelled, 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 is often quite sparse due to 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 sub-populations 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 (PCA) 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 due to 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], bi-clustering [84,86,87], affinity propagation [84,88], density-based spatial clustering (e.g. DBSCAN) [26,89], modelling as a mixture of multivariate distributions [90,91], and networks-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 2 or 3 dimensions for visualization and further analysis [26]. Once cell sub-populations have been identified, their transcriptomes can be combined to form “pooled” transcriptomes for deeper characterization [30] in order 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 last 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 datasets 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, copy number variations (CNV’s), 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.
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 the Ramalingaswami Fellowship, DBT, from the Government of India.

ACKNOWLEDGEMENTS

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


42. Anders S, Pyl PT, Huber W. HTSeq-A Python framework to work with high-throughput sequencing data. Bioinformatics 2015; 31:166–169


67. Reznikoff WS. Tn5 as a model for understanding DNA transposition. Mol. Biol. 2003; 47:1199–1206


76. Shiroguchi K, Jia TZ, Sims PA, et al. Digital RNA sequencing minimizes sequence-dependent


88. Frey BJ, Dueck D. Clustering by passing messages between data points. Science 2007; 315:972–976


<table>
<thead>
<tr>
<th>Method</th>
<th>Number of cells</th>
<th>Number of genes</th>
<th>Information provided</th>
<th>Selected references or website</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Measuring gene expression of specific genes</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single cell qPCR</td>
<td>100-1,000 (Dynamic Array: 48 or 96 cells per chip)</td>
<td>Typically 48-96 (Up to 280)</td>
<td>Ct values</td>
<td>Benchtop: [32] Microfluidics (Dynamic Array): [16,18,20,33,35,59,60]</td>
</tr>
<tr>
<td>Targeted single cell RNA sequencing</td>
<td>100-1,200</td>
<td>150-200</td>
<td>Short sequencing reads covering the 3' end of pre-selected transcripts with cell specific barcodes and UMI’s</td>
<td>Benchtop: [73]</td>
</tr>
<tr>
<td>CytoSeq</td>
<td>1,000-50,000 (100,000 micro-wells per chip)</td>
<td>150-200</td>
<td>Short sequencing reads covering the 3' end of pre-selected transcripts with cell specific barcodes and UMI’s</td>
<td>Microwells: [23]</td>
</tr>
<tr>
<td><strong>Full transcript length mRNA sequencing</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tang and Surani single cell RNAseq</td>
<td>100-1,000</td>
<td>Whole transcriptome (~25,000)</td>
<td>Short sequencing reads covering full transcript length</td>
<td>Benchtop: [63,99,100] Microfluidics: [22]</td>
</tr>
<tr>
<td>SMART-seq</td>
<td>100-1,000 (C1: 96 cells per chip)</td>
<td>Whole transcriptome (~25,000)</td>
<td>Short sequencing reads covering full transcript length</td>
<td>Template switching: [64,65] Tagmentation: [66–69,101] Benchtop: [28,49,71] Microfluidics (C1): [15,19,72,102,103]</td>
</tr>
<tr>
<td>SMART-seq2</td>
<td>100-1,000</td>
<td>Whole transcriptome (~25,000)</td>
<td>Short sequencing reads covering full transcript length</td>
<td>Benchtop: [29,54,70,104,105]</td>
</tr>
<tr>
<td>MATQ-seq</td>
<td>100-1000</td>
<td>Whole transcriptome (~25,000)</td>
<td>Short sequencing reads covering full transcript length (with UMI’s)</td>
<td>Benchtop: [106]</td>
</tr>
<tr>
<td><strong>Molecular tagging and counting strategies</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>STRT-Seq</td>
<td>100-1,000</td>
<td>Whole transcriptome (~25,000)</td>
<td>Short sequencing reads covering the 5’ end of each transcript with cell specific barcodes</td>
<td>Benchtop: [75,107]</td>
</tr>
<tr>
<td>STRT/C1-Seq</td>
<td>100-1,000 (C1: 96 cells per chip)</td>
<td>Whole transcriptome (~25,000)</td>
<td>Short sequencing reads covering the 5’ end of each transcript with cell specific barcodes and UMI’s</td>
<td>Microfluidics (C1): [47,84,87]</td>
</tr>
<tr>
<td>CEL-Seq</td>
<td>100-1,000</td>
<td>Whole transcriptome (~25,000)</td>
<td>Short sequencing reads covering the 3' end of each transcript with cell specific barcodes</td>
<td>Benchtop: [108]</td>
</tr>
<tr>
<td>CEL-Seq2</td>
<td>100-1,000 (C1: 96 cells per chip)</td>
<td>Whole transcriptome (~25,000)</td>
<td>Short sequencing reads covering the 3’ end of each transcript with cell specific barcodes</td>
<td>Benchtop / Microfluidics (C1): [31]</td>
</tr>
<tr>
<td>MARS-Seq</td>
<td>1,000-10,000</td>
<td>Whole transcriptome (~25,000)</td>
<td>Short sequencing reads covering the 3’ end of each transcript with cell specific barcodes and UMI’s</td>
<td>Benchtop: [30,85]</td>
</tr>
<tr>
<td>SCR-Seq</td>
<td>1,000-10,000</td>
<td>Whole transcriptome (~25,000)</td>
<td>Short sequencing reads covering the 3’ end of each transcript with cell specific barcodes and UMI’s</td>
<td>Benchtop: [109]</td>
</tr>
<tr>
<td>DropSeq</td>
<td>1,000-50,000</td>
<td>Whole transcriptome (~25,000)</td>
<td>Short sequencing reads covering the 3’ end of each transcript with cell specific barcodes and UMI’s</td>
<td>Droplets: [26]</td>
</tr>
<tr>
<td>Technology</td>
<td>Cells per Chip</td>
<td>Whole Transcriptome (~25,000)</td>
<td>Short sequencing reads covering the 3' end of each transcript with cell specific barcodes and UMI’s</td>
<td>Notes</td>
</tr>
<tr>
<td>---------------------</td>
<td>----------------</td>
<td>-------------------------------</td>
<td>-------------------------------------------------------------------------------------------------</td>
<td>-------</td>
</tr>
<tr>
<td>Seq-Well</td>
<td>1,000-50,000</td>
<td>Whole transcriptome (~25,000)</td>
<td>Short sequencing reads covering the 3' end of each transcript with cell specific barcodes and UMI’s</td>
<td></td>
</tr>
<tr>
<td>inDrop</td>
<td>1,000-50,000</td>
<td>Whole transcriptome (~25,000)</td>
<td>Short sequencing reads covering the 3' end of each transcript with cell specific barcodes and UMI’s</td>
<td></td>
</tr>
<tr>
<td>Fluidigm HT IFC</td>
<td>1,000-10,000</td>
<td>Whole transcriptome (~25,000)</td>
<td>Short sequencing reads covering the 3' end of each transcript with cell specific barcodes and UMI’s</td>
<td></td>
</tr>
<tr>
<td>Biorad ddSEQ</td>
<td>1,000-50,000</td>
<td>Whole transcriptome (~25,000)</td>
<td>Short sequencing reads covering the 3' end of each transcript with cell specific barcodes and UMI’s</td>
<td></td>
</tr>
<tr>
<td>Chromium Single Cell 3’ Solution</td>
<td>1,000-50,000</td>
<td>Whole transcriptome (~25,000)</td>
<td>Short sequencing reads covering the 3' end of each transcript with cell specific barcodes and UMI’s</td>
<td></td>
</tr>
<tr>
<td>iCELL8</td>
<td>1,000-50,000</td>
<td>Whole transcriptome (~25,000)</td>
<td>Short sequencing reads covering the 3' end of each transcript with cell specific barcodes and UMI’s</td>
<td></td>
</tr>
<tr>
<td>SPLiT-seq</td>
<td>10,000-100,000</td>
<td>Whole transcriptome (~25,000)</td>
<td>Short sequencing reads covering the 3' end of each transcript with cell specific barcodes and UMI’s</td>
<td></td>
</tr>
<tr>
<td>sci-RNA-seq</td>
<td>10,000-100,000</td>
<td>Whole transcriptome (~25,000)</td>
<td>Short sequencing reads covering the 3' end of each transcript with cell specific barcodes and UMI’s</td>
<td></td>
</tr>
</tbody>
</table>

**Table 1:** Selected single-cell transcriptomic technologies.
**Fig. 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 x 96 = 9,216 qPCR reactions in parallel, thus allowing expression levels of 96 genes from 96 individual cells to be measured in a single experiment.
Fig. 2: An example for a typical “SMART-Seq” protocol for full transcript length mRNA sequencing from 96 individual cells using the Fluidigm 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 is 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 of 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 template-switching oligo (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 template-switching oligo (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).
Fig. 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 12bp which are the same for all primers linked to the same bead), a subsequent UMI (a random sequence of 8bp), 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 re-suspended in reverse transcription buffer. The mRNA is then reverse transcribed and pre-amplified. In order 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 in order 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 in order to choose only those fragments that contain the cell specific barcode and UMI that were inserted in the very first reverse transcription step. This requires using a custom read 1 sequencing primer that contains segments of the reverse transcription primer.
Fig 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 de-multiplexed 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 in order 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 3 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 UMI’s for each gene is counted. This better represents the original number of transcripts.