Novel Technologies for Single-Cell Resolution Whole-Transcriptome Analysis in CNS Tissue

Evan Macosko, MD, PhD

Stanley Center for Psychiatric Research
Broad Institute of MIT and Harvard
Harvard Medical School
Cambridge, Massachusetts

SOCIETY *for* NEUROSCIENCE

Introduction

Individual cells are the building blocks of tissues, organs, and organisms. Each tissue contains cells of many types, and cells of each type can switch among biological states. Especially in the mammalian brain, our knowledge of cellular diversity is incomplete. In particular, the extent of cell-type complexity in the brain remains unknown and is widely debated (Luo et al., 2008; Petilla Interneuron Nomenclature Group et al., 2008). Many important but rare cell populations likely remain undiscovered, potentially limiting our understanding of physiological function. In addition, the overall landscape of transcriptional variation, even among abundant cell types, is mostly undescribed.

A major determinant of each cell type's function is its transcriptional program. Consequently, ascertainment of sufficient numbers of single-cell gene expression profiles may enable a comprehensive taxonomy of cell populations across the mammalian nervous system. Although two molecular techniques for isolating and amplifying small amounts of mRNA were developed some time ago—T7 amplification (Eberwine et al., 1992) and SMART (switching mechanism at 5' end of RNA template) technology (Matz et al., 1999; Zhu et al., 2001)—it was the advent of high-throughput next-generation sequencing technologies, coupled with these amplification techniques, that has made the analysis of meaningful numbers of single-cell gene expression profiles possible. Together with improved techniques for isolating individual cells, barcoding their transcriptional contents, and miniaturizing amplification volumes, single-cell gene expression profiling has moved rapidly from an era in which only a handful of profiles could be gleaned in a major study, to one in which the routine ascertainment of tens of thousands of profiles in a single experiment is now possible.

This chapter is divided into three sections, describing (1) the various technological innovations that made this recent transformation possible; (2) the important technical parameters for assessing the quality of data produced by these techniques; and (3) a discussion of biological applications of single-cell gene expression analysis and future technological directions.

Single-Cell mRNA-seq: From Handfuls to Thousands of Cell Profiles

Amplifying and interrogating small quantities of mRNA

Gene expression analysis at the level of individual cells began soon after the advent of techniques for amplifying minute quantities of mRNA. In 1992, Eberwine and colleagues used T7 amplification to prepare cDNA libraries from individually hand-picked hippocampal cells (Eberwine et al., 1992). T7 amplification works by reverse transcription of an mRNA pool using an oligo dT primer fused to a T7 RNA polymerase promoter sequence. After second-strand synthesis, the double-stranded cDNA is used as the template for *in vitro* transcription amplification by T7 RNA polymerase. The resulting RNA amplicons are reverse transcribed in bulk to yield an amplified cDNA library. By repeating this process twice, Eberwine's group was able to achieve an amplification factor of ~10^6. Sometime later, an alternative approach was developed that uses the template-switching capability of MMLV (Moloney murine leukemia virus) reverse transcriptase (known as SMART) to amplify small quantities of cDNA by PCR (Matz et al., 1999). This approach is the basis of the suite of RNA amplification products manufactured and sold by Clontech Laboratories (Mountain View, CA). Initially, the single-cell cDNA libraries produced by these amplification schemes were interrogated by hybridization (Northern blot and microarray analysis). Today, however, the improved throughput, precision, and accuracy of next-generation sequencing have made mRNA sequencing (mRNA-seq) the near-universal choice for measuring the concentration of individual RNA species.

The most common single-cell RNA-seq protocols currently in use continue to feature either T7 or SMART amplification to generate cDNA libraries. The two amplification schemes have different advantages: T7 amplification, because it is linear, is generally believed to produce more even amplification of a diverse cDNA library, while SMART is somewhat less technically demanding.

Approaches to isolating individual cells

A major impediment to high-throughput examination of single-cell profiles is the technical difficulty associated with isolating individual cells. Hand-picking cells (the traditional approach) allows for visual confirmation of cell capture and morphological screening for a desired cell population, but is inherently very time-consuming. Flow cytometry sorting of individual cells into microtiter plates (Jaitin et al., 2014; Tasic et al., 2016) provides a significant improvement in scale and can be combined with fluorescent staining to screen for subsets of cells of interest. Microfluidic techniques have also been developed to isolate cells. Traditional valve-based microfluidic devices capture cells within individual chambers and process the isolated mRNA in parallel

NOTES

(White et al., 2011). Two commercially available products from Fluidigm (South San Francisco, CA) and WaferGen Bio-systems (Fremont, CA) enable several hundred cells to be captured and processed at once. In contrast, microfluidic droplet–generation devices can disperse tens of thousands of precisely sized ("monodisperse") picoliter-scale or nanoliter-scale droplets per minute (Umbanhowar, 2000; Thorsen et al., 2001). By critically diluting a cell suspension to a concentration far lower than one cell per droplet, individual cells can be isolated in extremely high throughput in these emulsions (tens of thousands per hour).

Massive molecular barcoding

Following technical improvements in the ease and throughput of cell isolation, particularly by droplet microfluidics, the major obstacle to routine, massively multiplexed single-cell mRNA-seq became the cost and time required to prepare individual libraries from so many cells in individual microtiter reactions. If the mRNA content of individual cells could be barcoded at the start of processing, then all subsequent molecular amplification and library preparation steps could be performed in a single bulk reaction, dramatically simplifying the process. Recently, two barcoding approaches were developed that address this problem (Klein et al., 2015; Macosko et al., 2015). In each, a collection of microparticles (beads) is generated, each of which harbors a large number of barcoded oligo dT primers on its surface; the barcode is the same across all the primers on the surface of any one bead but differs from the barcodes on all other beads. In the first method, Drop-seq, barcode diversity is generated through a modified form of chemical oligonucleotide synthesis, in which beads are repeatedly split and pooled to achieve millions of unique sequences (Fig. 1). The second method, inDrop, uses an enzymatic approach to combinatorially stitch together two sets of barcoded oligos, resulting in a pool of beads with hundreds of thousands of individual barcodes. Both methods are able to collectively barcode and process thousands of cells in a single experiment.

Technical assessments of single-cell RNA-seq data

To glean meaningful biological signals from any technology, it is vital to have technical measurements that assess the strengths and limitations of the data. Single-cell RNA-seq (scRNA-seq) technologies should be evaluated by several criteria: (1) the amount of RNA that is captured; (2) the specificity of the signal (how truly "single-cell" the profile is); and (3) how consistent the resulting profile is across individual technical replicates.

RNA capture efficiency

The most common method for estimating the proportion of sampled transcripts is to process a synthetic library of RNAs (known as the External RNA Controls Consortium [ERCC] "spike-in" controls) and compute the fraction of these RNAs that are reported by sequencing. In general, these analyses have produced estimates of between 2% and 12% capture efficiency across different technological platforms (Grun et al., 2014; Klein et al., 2015; Macosko et al., 2015). One study explained the majority of the loss by inefficiency in the mRNA hybridization step (Macosko et al., 2015); it remains unknown whether this step is also the bottleneck for other methods.

A typical mammalian cell contains 5–10 pg of total RNA (Tang et al., 2011), of which 1%–10% is polyadenylated, mature mRNA. This corresponds to ~100,000–500,000 unique mRNA molecules, distributed across thousands of individual genes. This means that, at a capture efficiency of 10%, many minimally expressed genes will go undetected in a given cell. High-throughput single-cell technologies like Drop-seq and inDrop can address this problem by repeatedly sampling cells of the same type to accrue observations of these low-copy genes.

Doublet rates and purity

One mode of failure in any single-cell method involves cells that stick together or happen to otherwise be co-isolated for library preparation. To measure doublet rates, two groups recently sequenced mixtures of cells derived from two species and calculated organism purity rates of individual cell barcodes. For droplet-based approaches (i.e., inDrop and Drop-seq), the doublet rate could be adjusted to arbitrarily low levels by reducing the cell concentration. Although doublet rates can be higher in other systems (e.g., Fluidigm C1), many of these doublets can be identified up front by fluorescence microscopy of the capture chambers (Fluidigm, 2016). Species-mixing experiments enable a careful quantification of single-cell purity across libraries. In Drop-seq, impurity was strongly related to the concentration at which cell suspensions were loaded: organism purity ranged from 98.8% at 12.5 cells/µl to 90.4% at 100 cells/µl.

Technical reliability

Replication across experimental sessions enables the construction of cumulatively more powerful datasets for detecting subtle biological signals. Technical variation can arise from day-to-day differences in cell preparation, molecular processing and sequencing,

NOTES

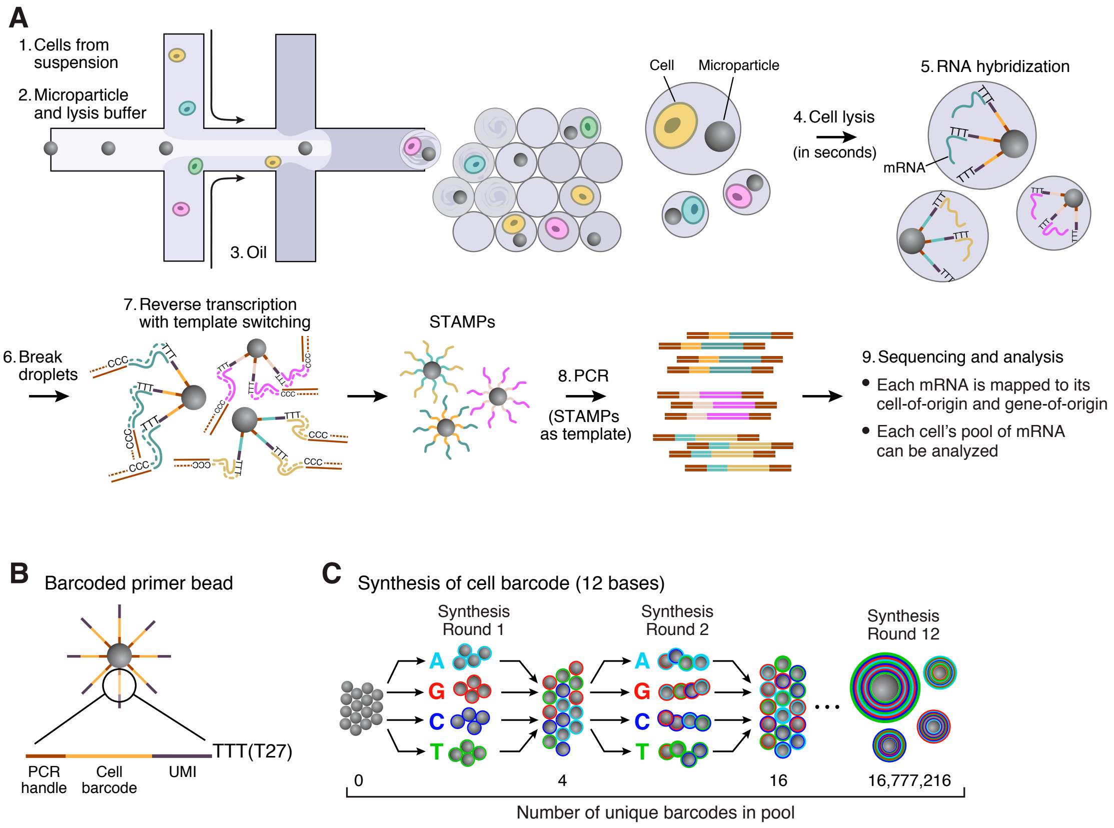
A
1. Cells from suspension
2. Microparticle and lysis buffer
3. Oil
Cell
Microparticle
4. Cell lysis
(in seconds)
5. RNA hybridization
mRNA
TTT
6. Break droplets
7. Reverse transcription with template switching
CCC
STAMPs
8. PCR
(STAMPs as template)
9. Sequencing and analysis
• Each mRNA is mapped to its cell-of-origin and gene-of-origin
• Each cell's pool of mRNA can be analyzed
B
Barcoded primer bead
TTT(T27)
PCR handle
Cell barcode
UMI
C
Synthesis of cell barcode (12 bases)
Synthesis Round 1
Synthesis Round 2
Synthesis Round 12
A
G
C
T
0
4
16
16,777,216
Number of unique barcodes in pool

Figure 1. Drop-seq: molecular barcoding of cellular transcriptomes using droplet microfluidics. ***A,*** Schematic of single-cell mRNA-seq library preparation using Drop-seq. A custom-designed microfluidic device joins two aqueous flows before their compartmentalization into discrete droplets. One flow contains cells, and the other flow contains barcoded primer beads suspended in a lysis buffer. Immediately after droplet formation, the cell is lysed and releases its mRNAs, which then hybridize to the primers on the microparticle surface. The droplets are broken up by adding a reagent to destabilize the oil–water interface, and the microparticles are collected and washed. The mRNAs are then reverse transcribed in bulk, forming STAMPs (single-cell transcriptomes attached to microparticles), and template switching is used to introduce a PCR handle downstream of the synthesized cDNA (Zhu et al., 2001). ***B,*** Sequence of primers on the microparticle. The primers on all beads contain a common sequence ("PCR handle") to enable PCR amplification after STAMP formation. Each microparticle contains >10^8 individual primers that share the same "cell barcode" (***C***) but have different unique molecular identifiers (UMIs), enabling mRNA transcripts to be digitally counted. A 30 bp oligo dT sequence is present at the end of all primer sequences for capture of mRNAs. ***C,*** Split-and-pool synthesis of the cell barcode. To generate the cell barcode, the pool of microparticles is repeatedly split into four equally sized oligonucleotide synthesis reactions, to which one of the four DNA bases is added, and then pooled together after each cycle, in a total of 12 split-pool cycles. The barcode synthesized on any individual bead reflects that bead's unique path through the series of synthesis reactions. The result is a pool of microparticles, each possessing one of 4^{12} (16,777,216) possible sequences on its entire complement of primers. Reprinted with permission from Macosko EZ et al. (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell 161:1203, 1205; their Figs. 1*B*, *C*, and 2*A*. Copyright 2015, Elsevier.

or peculiarities specific to particular systems. We clustered ~45,000 Drop-seq–derived single-cell profiles from dissociated mouse retinas over the course of seven experimental sessions. The resulting 39 clusters contained cells from each session, suggesting that the technical variation in gene expression was relatively small even compared with the differences between highly similar cell subtypes. New high-throughput technologies should provide large enough datasets to permit more-rigorous computational analyses in which portions of the data are withheld (e.g., *k*-fold cross-validation).

Biological applications and technological improvements

Already, studies using scRNA-seq have transformed our understanding of cellular diversity in many mammalian CNS tissues, including the spinal cord (Usoskin et al., 2015), cortex (Zeisel et al., 2015;

NOTES

Tasic et al., 2016), and retina (Macosko et al., 2015). In addition, biologists are quickly recognizing the plethora of scientific opportunities enabled by ascertaining transcriptional variation in individual cells, beyond performing initial taxonomic analyses of tissues. For example, genome-scale genetic studies are identifying large numbers of genes in which genetic variation contributes to disease risk. Finding the cellular sites and biological activities of so many genes is an important but challenging goal. High-throughput single-cell transcriptomics could localize the expression of all risk genes to specific cell types, and in conjunction with genetic perturbations, help to systematically relate each gene to (1) the cell types most affected by loss or perturbation of those genes and (2) the alterations in cell state elicited by such perturbations. Such approaches could help cross the daunting gap from gene discoveries to insights about pathophysiology.

ScRNA-seq (possibly coupled to additional manipulations) could be used to generate an information-rich, multidimensional readout of the influence of many kinds of perturbations—such as small molecules, genetic mutations (natural or engineered), pathogens, or other stimuli—on many kinds of cells. When studying the effects of a mutation, for example, scRNA-seq could illuminate pleiotropies by revealing the ways in which the same mutation differentially impacts distinct cell types. Single-cell expression analysis could also be used to characterize the heterogeneous responses of diverse cell populations to a drug or metabolite, or combinations thereof.

Enormous opportunities exist to improve approaches to single-cell gene expression analysis. First, the extension of existing methods to the analysis of frozen and/or fixed tissue could help relate functional genomic variation to transcriptional variation in specific cell types and provide novel hypotheses for how specific cell types are altered in disease states whose pathogeneses remain mysterious. Second, tissue dissociation before cell processing introduces artifactual signals (as the dissociated cells begin to die) and does not maintain spatial relationships among analyzed cells. Thus, multiple new technologies, including highly multiplexed *in situ* hybridization techniques (Chen et al., 2015; Coskun and Cai, 2016) and approaches to sequencing mRNA directly from tissue slices (Lee et al., 2014; Ståhl et al., 2016) could ultimately make it possible to perform single-cell profiling without tissue dissociation. Finally, the coupling of scRNA-seq with other cellular readouts, including single-cell epigenetic measurements and DNA sequencing, could someday provide fundamental insights into transcriptional regulation in specialized cell populations.

The functional implications of a gene's expression are a product not just of a gene's intrinsic properties but also of the entire cell-level context in which a gene is expressed. The routine facile, large-scale measurement of single-cell gene expression profiles with new technologies should enable the abundant and routine discovery of such relationships across biology.

Acknowledgments

Parts of this chapter were excerpted with permission from Macosko EZ et al. (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell 161:1202–1214. Copyright 2015, Elsevier.

References

Chen KH, Boettiger AN, Moffitt JR, Wang S, Zhuang X (2015) RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. Science 348:aaa6090.

Coskun AF, Cai L (2016) Dense transcript profiling in single cells by image correlation decoding. Nat Methods 13:657–660.

Eberwine J, Yeh H, Miyashiro K, Cao Y, Nair S, Finnell R, Zettel M, Coleman P (1992) Analysis of gene expression in single live neurons. Proc Natl Acad Sci USA 89:3010–3014.

Fluidigm (2016) Double rate and detection on the C1 IFCs. White Paper PN 101-2711 A1. https://mailbox.zimbra.upenn.edu/home/ngsc@zimbra.upenn.edu/Briefcase/PublicFiles/Equipment/Fluidigm/C1/C1_Doublet_wp_101-2711A1_20160106.pdf

Grun D, Kester L, van Oudenaarden A (2014) Validation of noise models for single-cell transcriptomics. Nat Methods 11:637–640.

Jaitin DA, Kenigsberg E, Keren-Shaul H, Elefant N, Paul F, Zaretsky I, Mildner A, Cohen N, Jung S, Tanay A, Amit I (2014) Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science 343:776–779.

Klein AM, Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz., D.A., Kirschner, M.W. (2015) Droplet barcoding for single cell transcriptomics and its application to embryonic stem cells. Cell 161:1187–1201.

Lee JH, Daugharthy ER, Scheiman J, Kalhor R, Yang JL, Ferrante TC, Terry R, Jeanty SS, Li C, Amamoto R, Peters DT, Turczyk BM, Marblestone AH, Inverso SA, Bernard A, Mali P, Rios X, Aach J, Church GM (2014) Highly multiplexed subcellular RNA sequencing *in situ*. Science 343:1360–1363.

Luo L, Callaway EM, Svoboda K (2008) Genetic dissection of neural circuits. Neuron 57:634–660.

Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, Trombetta JJ, Weitz DA, Sanes JR, Shalek AK, Regev A, McCarroll SA (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell 161:1202–1214.

Matz M, Shagin D, Bogdanova E, Britanova O, Lukyanov S, Diatchenko L, Chenchik A (1999) Amplification of cDNA ends based on template-switching effect and step-out PCR. Nucleic Acids Res 27:1558–1560.

Petilla Interneuron Nomenclature Group (PING), Ascoli GA, Alonso-Nanclares L, Anderson SA, Barrionuevo G, Benavides-Piccione R, Burkhalter A, Buzsáki G, Cauli B, Defelipe J, Fairén A, Feldmeyer D, Fishell G, Fregnac Y, Freund TF, Gardner D, Gardner EP, Goldberg JH, Helmstaedter M, Hestrin S, et al. (2008) Petilla terminology: nomenclature of features of GABAergic interneurons of the cerebral cortex. Nat Rev Neurosci 9:557–568.

Ståhl PL, Salmén F, Vickovic S, Lundmark A, Navarro JF, Magnusson J, Giacomello S, Asp M, Westholm JO, Huss M, Mollbrink A, Linnarsson S, Codeluppi S, Borg Å, Pontén F, Costea PI, Sahlén P, Mulder J, Bergmann O, Lundeberg J, Frisén J (2016) Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science 353:78–82.

Tang F, Lao K, Surani MA (2011) Development and applications of single-cell transcriptome analysis. Nat Methods 8:S6–S11.

Tasic B, Menon V, Nguyen TN, Kim TK, Jarsky T, Yao Z, Levi B, Gray LT, Sorensen SA, Dolbeare T, Bertagnolli D, Goldy J, Shapovalova N, Parry S, Lee C, Smith K, Bernard A, Madisen L, Sunkin SM, Hawrylycz M, et al. (2016) Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. Nat Neurosci 19:335–346.

Thorsen T, Roberts RW, Arnold FH, Quake SR (2001) Dynamic pattern formation in a vesicle-generating microfluidic device. Phys Rev Lett 86:4163–4166.

Umbanhowar PB, Prasad V, Weitz DA (2000) Monodisperse emulsion generation via drop break off in a coflowing stream. Langmuir 16:347–351.

Usoskin D, Furlan A, Islam S, Abdo H, Lonnerberg P, Lou D, Hjerling-Leffler J, Haeggstrom J, Kharchenko O, Kharchenko PV, Linnarsson S, Ernfors P (2015) Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. Nat Neurosci 18:145–153.

White AK, VanInsberghe M, Petriv OI, Hamidi M, Sikorski D, Marra MA, Piret J, Aparicio S, Hansen CL (2011) High-throughput microfluidic single-cell RT-qPCR. Proc Natl Acad Sci USA 108:13999–14004.

Zeisel A, Munoz-Manchado AB, Codeluppi S, Lonnerberg P, La Manno G, Jureus A, Marques S, Munguba H, He L, Betsholtz C, Rolny C, Castelo-Branco G, Hjerling-Leffler J, Linnarsson S (2015) Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science 347:1138–1142.

Zhu YY, Machleder EM, Chenchik A, Li R, Siebert PD (2001) Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. Biotechniques 30:892–897.