# Introduction: Scientific Opportunities and Challenges in Single-Cell Analysis

Steven A. McCarroll, PhD

Harvard Medical School
Boston, Massachusetts

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

SOCIETY *for* NEUROSCIENCE

## Background

Individual cells are the basic units with which larger biological systems—circuits, tissues, and entire organisms—are built. Cells in the same tissue or circuit have various biological missions; a cell's missions are reflected in its size, morphology, physiology, and use of its genome. Adjacent cells often use the same genome in dramatically different ways.

Historically, insights about cell types and their specialization were obtained one at a time, as a result of varying combinations of serendipity and painstaking work. The discovery of a cell population with unusual physiological properties might be followed later by the identification of a molecular marker for those cells, and then eventually by insights into these cells' interactions with and connectivity to other cells. Several technological innovations promise to transform the pace of discovery about cell types and their properties—first, by allowing the collection of genome-scale information (e.g., about gene expression or DNA sequence) from individual cells (Tang et al., 2009), and more recently, by allowing genome-scale analyses to be conducted on vast numbers of individual cells at once (Klein et al., 2015; Macosko et al., 2015). The pace of data generation has increased dramatically; the pace of biological insights will, one hopes, begin to increase as well.

## Moving From Proofs of Concept to Useful Data Resources to Insights

Emerging fields in genomics often follow a similar trajectory. Early "proof of concept" studies serve to illustrate that new kinds of analysis can be executed. Although the data and analysis methods are often quickly replaced by better approaches, such early results help many readers to expand their sense of the possible.

As experimental approaches begin to stabilize and mature (such that the shelf life of a dataset is longer and its quality more assured), it becomes possible to build data resources that have cumulative value. In human genetics, for example, datasets on human genome variation (alleles and allele frequencies at each site in the genome) are used in thousands of genetic inquiries every day, supporting both genome-scale studies and analyses of individual genes (International HapMap Consortium, 2015; Lek et al., 2016). For single-cell transcriptomics, such resources may increasingly take the form of digital atlases in which the expression profiles of individual cell types can be looked up (Tasic, et al., 2016). Such resources may come to have great value because they allow routine lookups of genes' expression patterns across cell types. Their immediate results may be more facile, quantitative, and reliable than images collected by laborious slogs involving antibodies of varying qualities, tissues and fixatives with varying properties, and hours of microscopy.

The most rewarding phase can occur as new tools and data resources begin to support scientific insights into molecular and cellular mechanisms, and as broader experimental programs and plans reshape themselves to utilize the opportunities inherent in new kinds of data and new ways of monitoring biological systems.

## Approaching Integrated Analysis

For single-cell analysis, a growing scientific opportunity will come from beginning to draw connections among the different ways of characterizing individual cells—to appreciate how morphology, physiology, connectivity, and gene expression are codistributed and interconnected mechanistically. Ideally, the cell atlases of the future will report not only what genes each type of cell expresses but also what shape(s) it assumes, what neurons it connects with, what transmitters it responds to, and what voltage and ionic dynamics it has. Armed with this kind of characterization, we will be able to begin to understand how gene expression, morphology, physiology, and connectivity influence and arise from one another. In an early step in this direction, a recent study related the electrophysiological properties of individual cells to their molecular profiles (Cadwell et al., 2016).

A practical challenge of integrated analysis involves the fact that many kinds of analyses of individual cells (e.g., transcriptomics, fixation for immunohistochemistry) destroy these cells' other properties, leaving little room for subsequent analyses of the same cells. In this Short Course, we will discuss the opportunities that arise from integrating multimodal data types at single-cell resolution and the practical challenges of accomplishing this.

## Developing Clearer, More Useful Standards and Metrics

New fields often struggle to clarify their thinking about how to quantify and compare findings and how to distinguish real signals from artifacts. Single-cell analysis of somatic DNA variation, for example, now indicates that rates of somatic retrotransposition are far lower than was reported in some earlier studies (Evrony et al., 2016). Perhaps nowhere has such confusion been more abundant than in single-cell transcriptomics. Today, the thoroughness of single-cell experiments is still often evaluated in terms of "reads per cell": the ratio of the number of sequencing reads generated to the number of cells analyzed. However, this metric may offer little information about what was learned, because any number of DNA or RNA molecules can be amplified into an arbitrarily large number of copies and then resequenced using an arbitrarily large number of sequencing reads without generating any new information. (Put another way, if a tree falls in the woods, it matters little whether that event is documented by one, 10, or 1000 observers, so long as the fall is recorded reliably and distinguished from that of other trees.)

A similar confusion involves the use of the metric "genes detected per cell." The number of genes expressed in a cell depends strongly on cell type, and more important, this number is inflated when an analysis is not truly single-cell (e.g., when a cell doublet is assumed to be a single cell). This problem appears to have inflated estimates in early single-cell studies. Significant advances, such as the use of unique molecular indicators (UMIs) (Kivioja et al., 2011), which affix a particular molecular barcode to each cDNA and allow digital counting with correction for amplification effects, are increasingly enabling true estimates of transcript ascertainment. To return to our "tree falling" analogy, UMIs make it possible to recognize when many observers reporting a "tree falling" are in fact all talking about the same tree. Not surprisingly, the figures yielded by UMI-informed analyses—typically quantified as transcripts per cell (trees) rather than reads per cell (observers)—are also far more modest. Still, UMIs have offered a significant step forward in clarity, even if the resulting estimates have less "bling." A goal of this Short Course will be try to clarify such terms and help scientists to design, evaluate, and think about experiments in quantitative ways.

## Scaling Up Computational Approaches

Most single-cell experimental approaches in use today produce novel kinds of datasets for which computational methods are still in their infancy. For example, methods for collecting gene-expression information from tens of thousands of individual cells have created a new scientific opportunity to infer cell types and cell states (including rare ones) in "unsupervised" ways that are not constrained by earlier theories, categories, or lists of markers. This opportunity needs to be met increasingly by new analytical approaches. Many computational approaches that were developed for early, small single-cell RNA-seq datasets do not scale up successfully to, or do not realize the opportunities inherent in, the far-larger datasets that are being generated. Thus, an important direction will be to develop algorithms that can recognize patterns and structure in vast multidimensional datasets and then present these patterns in ways that lead to biological insights. This exciting emerging area will benefit from close multidisciplinary collaborations among scientists who have expertise in computer science, biology, statistics, and mathematics.

## Seizing the Opportunity Ahead

The functions of tissues and organs derive from interactions and collaborations among specialized individual cells. Elucidating how tissue and circuit functions encompass the actions of specialized cells expressing distinct genes and molecular complexes, with varying proximity and connectivity, is one of the great scientific challenges of our time. Aspiring to such understanding is also increasingly within our grasp.

## References

Cadwell CR, Palasantza A, Jiang X, Berens P, Deng Q, Yilmaz M, Reimer J, Shen S, Bethge M, Tolias KF, Sandberg R, Tolias AS (2016) Electrophysiological, transcriptomic and morphologic profiling of single neurons using Patch-seq. Nat Biotechnol 34:199–203.

Evrony GD, Lee E, Park PJ, Walsh CA (2016) Resolving rates of mutation in the brain using single-neuron genomics. Elife 5. pii: e12966.

International HapMap Consortium (2015) A haplotype map of the human genome. Nature 437:1299–1320.

Kivioja T, Vähärautio A, Karlsson K, Bonke M, Enge M, Linnarsson S, Taipale J. (2011) Counting absolute numbers of molecules using unique molecular identifiers. Nat Methods 9:72–74.

Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA, Kirschner MW (2015) Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell 161:1187–1201.

NOTES

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, Tukiainen T, Birnbaum DP, Kosmicki JA, Duncan LE, Estrada K, Zhao F, Zou J, Pierce-Hoffman E, Berghout J, Cooper DN, et al. (2016) Analysis of protein-coding genetic variation in 60,706 humans. Nature 536:285–291.

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.

Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A, Lao K, Surani MA (2009) mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods 6:377–382.

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.