How deep is enough in single-cell RNA-seq?

Aaron M Streets & Yanyi Huang

Guidelines for determining sequencing depth facilitate transcriptome profiling of single cells in heterogeneous populations.

In recent years, single-cell RNA-seq has emerged as a powerful, new approach for characterizing the cell types present in a mixed population. These studies usually involve a trade-off between the number of samples analyzed and the number of RNA transcripts sequenced per cell, or sequencing depth, that can be achieved. In this issue, Pollen *et al.*[1] present quantitative guidelines for determining the sequencing depth necessary to distinguish the cell types in a complex sample. Using a commercial microfluidic platform to capture hundreds of cells from a variety of human tissues and performing RNA-seq to different depths, they demonstrate accurate and reliable classification of cell types at a sequencing depth of only 50,000 reads per cell (**Fig. 1a**)—about two orders of magnitude fewer than what has been typically reported.

Identification of cell types in mixed populations has long been done using known biomarkers analyzed by fluorescence-activated cell sorting or multiplexed, quantitative PCR. In contrast, the complete transcriptional profiles generated by single-cell RNA-seq allow cells to be identified objectively without a priori knowledge of biomarkers. This approach also enables the discovery of novel biomarkers.

In a typical single-cell RNA-seq experiment, tens to hundreds or even thousands of single cells are isolated from a tissue or culture, and the transcriptome of each cell is reverse transcribed into cDNA. The cDNA is then amplified and further processed for next-generation sequencing. The output of this pipeline is a list of sequence fragments called reads. Mapping of the reads to the reference genome produces estimates of normalized gene expression levels in an $N \times L$ matrix, where N is the number of cells and L is the total number of genes identified among all the cells (**Fig. 1b**).

Statistical analysis is then used to find trends in gene expression across many single cells. A common technique is unsupervised hierarchical clustering, which takes the $N \times L$ gene expression matrix and re-orders the row and column indices to minimize the difference between expression levels of adjacent elements along both dimensions. This process yields groups of cells and genes with similar expression levels and a dendrogram that identifies the distance between cells or genes. The result is often presented as a heat map of gene expression levels (**Fig. 1b**) to illustrate the subgroups.

Another technique, principal component analysis, allows the high-dimensional gene expression data set to be projected onto a lower-dimensional space in which the variation between samples is represented with fewer variables, called principal components. Groups of cells with similar transcriptional profiles can be visualized by plotting the first two or three principal components of each sample, revealing clusters in a two- or three-dimensional space (**Fig. 1c**).

Both of these methods have been applied to characterize changes in transcriptional profiles during development, for example, in human preimplantation embryos and embryonic stem cells[2] and in the mouse lung during alveolar differentiation[3]. These studies analyzed between 100 and 200 cells.

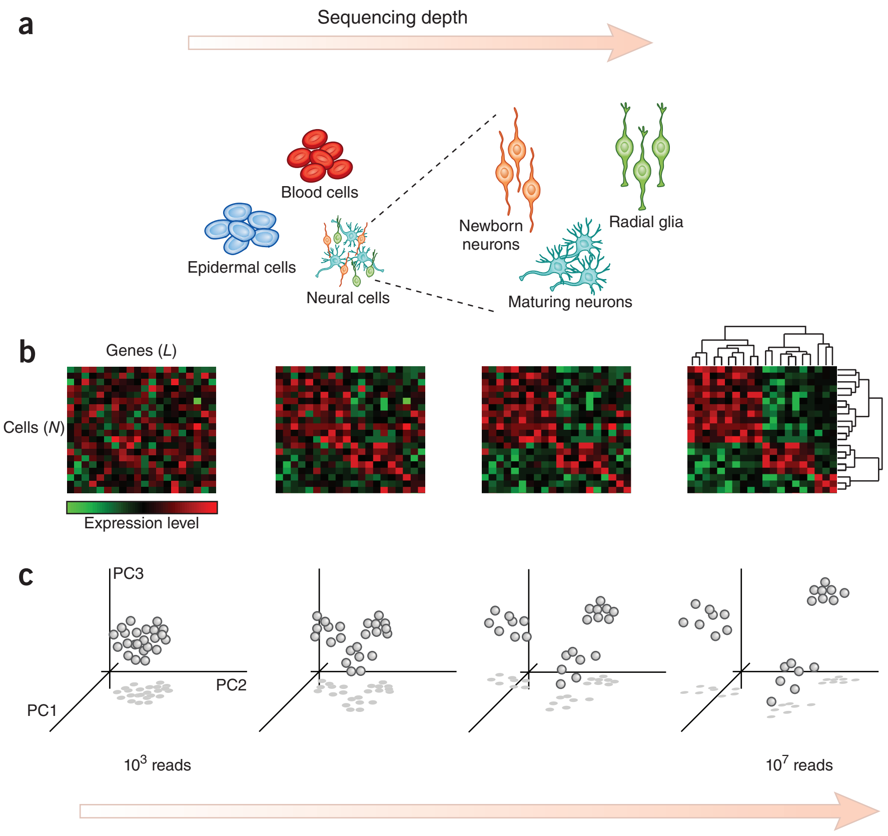
As the phenotypic heterogeneity of a biological sample increases, so does the sample size N needed to accurately describe the population. Recently, microfluidic[4] and robotic[5] platforms were applied to process thousands of single dendritic cells for whole-transcriptome analysis. Large sample sizes impose a practical limit on sequencing depth. So what depth is sufficient? One study showed that estimated expression levels from one million reads per cell strongly correlate with those from 10 million reads per cell[4], suggesting that one million reads per cell may suffice. In an even more ambitious study[5], highly multiplexed barcoding was used to process thousands of cells on a single sequencing run using ultra-shallow sequencing of only 20,000 reads per cell. This drastic undersampling of the transcriptome was still adequate to distinguish different cell types in splenic tissue, albeit with limited resolution.

Until now, however, the dependence of cell type classification on sequencing depth has not been understood as there has been no systematic analysis of how the transcriptional identity of a cell is preserved as sequencing depth is decreased. Pollen *et al.*[1] address this question by tracking the performance of hierarchical clustering and principal component analysis for phenotypic classification as cDNA libraries are sequenced to diminishing depths (**Fig. 1**). Microfluidic automation allows them to interrogate a variety of cell types, including blood cells, neurons and pluripotent cells, in a high-throughput fashion. In addition to increased throughput, the microfluidic approach provides improved sensitivity and reproducibility[6,7].

Pollen *et al.*[1] conclude that a majority of the primary genes that contribute to transcriptional variance among these diverse groups are identified by both high- and low-coverage sequencing analysis. The values of the first two principal components for individual cells are largely preserved down to as low as 10,000 reads per cell. They also show the capabilities of shallow sequencing in scenarios where subgroups differ by subtle transcriptional variation. Neural cells at different stages of the developing human neural cortex are distinguished with merely ~50,000 reads per cell, and new genetic markers of neural subclasses are identified.

This study adds to a growing body of literature aimed at assessing the technical performance of RNA-seq as a tool for quantitative biology.

Aaron M. Streets and Yanyi Huang are at the Biodynamic Optical Imaging Center (BIOPIC), and College of Engineering, Peking University, Beijing, China.
e-mail: *yanyi@pku.edu.cn*



Marina Corral Spence/Nature Publishing Group

Figure 1 The effect of sequencing depth on cell identification. Pollen *et al.*[1] use hierarchical clustering and principal component analysis to explore the relationship between sequencing depth and cell type identification. (**a**) Cells with substantially different gene expression profiles can be broadly distinguished with ultra-low sequencing (<10,000 reads). Neural cells at various stages of development can be identified with ~50,000 reads per cell. (**b**) Gene expression heat maps are used to visualize groups of neural cells with similar transcription profiles. The authors identify heterogeneity in various stages of the developing neural cortex using unsupervised hierarchical clustering to sort single cells (rows) and genes (columns). As sequencing depth is increased, cell and gene clusters are more easily distinguished. A dendrogram illustrates the distance between cells and genes (right panel). (**c**) Principal component analysis is also used to visualize groups of similar cells. The first three principal components (PC1, PC2 and PC3) retain a majority of the variation between cells, allowing groups of cells with similar expression profiles to be readily distinguished even with minimal sequencing depth.

With advances that enable study of larger sample sizes and facilitate analysis of big data, experimental design becomes more dependent on a thorough understanding of the capabilities and limitations of the techniques involved. For example, recent work[6–9] has characterized the sensitivity, accuracy and precision of microfluidic single-cell transcriptome analysis. Precision, or reproducibility, is a particularly important parameter when characterizing heterogeneity in large cell populations because it can be difficult to distinguish biological variation from experimental noise. To assess the noise in single-cell RNA-seq, researchers have measured transcript abundance in identical samples prepared by pooling RNA extracted from many cells[6,8,10]. The application of unique molecular barcodes for quantification of the pre-amplified transcript abundance[9,11] might enable modeling of noise and suppression of noise effects through statistical filtering[8,11]. Such technical investigations are critical to the interpretation of biological variation among single cells when using RNA-seq.

Another question in validating the sensitivity and accuracy of single-cell RNA-seq is whether data pooled from multiple single-cell experiments accurately represent the transcriptome measured by RNA-seq on bulk populations. This is a reasonable question because significant amplification is required to detect the genetic material in a single cell, and mRNA capture efficiency is limited. Thus, single-cell analysis is an undersampled and skewed approximation of the true transcript distribution. Despite these obstacles, Pollen *et al.*[1] verify previous findings[6,7,10] that single-cell transcriptomes can be merged to accurately represent a majority of the ensemble transcriptome. In fact, with as few as ten low-coverage, single-cell transcriptomes, over 80% of the bulk transcriptome can be detected with strongly correlated expression levels[1,6].

The work of Pollen *et al.*[1] defines the effectiveness of RNA-seq as sequencing depth decreases and establishes quantitative guidelines for experimental design. It also demonstrates that microfluidic technology facilitates reproducible, high-throughput, single-cell analysis. The results offer a roadmap for how to consider sequencing depth when performing cell type classification in any RNA-seq investigation with a large sample size. This will be particularly valuable in the study of dynamic responses of cell populations to stimuli or of complex tissues, such as brain[12] or tumor samples, likely to contain abundant phenotypic diversity and previously unidentified transcriptional states.

As researchers take on larger and larger sample sizes, a comprehensive understanding of the capabilities and limits of RNA-seq technology is essential. Studies that quantitatively assess the performance of genomic tools are a necessary foundation of ambitious biological investigation.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Pollen, A.A. *et al. Nat. Biotechnol.* **32**, 1053–1058 (2014).
2. Yan, L. *et al. Nat. Struct. Mol. Biol.* **20**, 1131–1139 (2013).
3. Treutlein, B. *et al. Nature* **509**, 371–375 (2014).
4. Shalek, A.K. *et al. Nature* **510**, 363–369 (2014).
5. Jaitin, D.A. *et al. Science* **343**, 776–779 (2014).
6. Streets, A.M. *et al. Proc. Natl. Acad. Sci. USA* **111**, 7048–7053 (2014).
7. Wu, A.R. *et al. Nat. Methods* **11**, 41–46 (2014).
8. Brennecke, P. *et al. Nat. Methods* **10**, 1093–1095 (2013).
9. Islam, S. *et al. Nat. Methods* **11**, 163–166 (2014).
10. Marinov, G.K. *et al. Genome Res.* **24**, 496–510 (2014).
11. Grün, D., Kester, L. & van Oudenaarden, A. *Nat. Methods* **11**, 637–640 (2014).
12. Zhang, Y. *et al. J. Neurosci.* **34**, 11929–11947 (2014).