Briefings in Functional Genomics, 2017, 1–6

doi: 10.1093/bfgp/elx044
Review paper

Clustering single cells: a review of approaches on high-and low-depth single-cell RNA-seq data

Vilas Menon

Corresponding author. Vilas Menon, Howard Hughes Medical Institute, Janelia Research Campus, 19700 Helix Drive, Ashburn, 20147 VA, USA. E-mail: menonv@janelia.hhmi.org

Abstract

Advances in single-cell RNA-sequencing technology have resulted in a wealth of studies aiming to identify transcriptomic cell types in various biological systems. There are multiple experimental approaches to isolate and profile single cells, which provide different levels of cellular and tissue coverage. In addition, multiple computational strategies have been proposed to identify putative cell types from single-cell data. From a data generation perspective, recent single-cell studies can be classified into two groups: those that distribute reads shallowly over large numbers of cells and those that distribute reads more deeply over a smaller cell population. Although there are advantages to both approaches in terms of cellular and tissue coverage, it is unclear whether different computational cell type identification methods are better suited to one or the other experimental paradigm. This study reviews three cell type clustering algorithms, each representing one of three broad approaches, and finds that PCA-based algorithms appear most suited to low read depth data sets, whereas gene clustering-based and biclustering algorithms perform better on high read depth data sets. In addition, highly related cell classes are better distinguished by higher-depth data, given the same total number of reads; however, simultaneous discovery of distinct and similar types is better served by lower-depth, higher cell number data. Overall, this study suggests that the depth of profiling should be determined by initial assumptions about the diversity of cells in the population, and that the selection of clustering algorithm(s) is subsequently based on the depth of profiling will allow for better identification of putative transcriptomic cell types.

Key words: RNA-sequencing; single cells; transcriptomics; clustering methods; cell type identification; read depth

The recent explosion in single-cell RNA-sequencing (RNA-seq) studies has led to the profiling and characterization of multiple organs. Cells have been classified into putative transcriptomic types in the hematopoietic system [1], the central nervous system [2–21], other organs [22–24] and *in vitro* systems studying various lineages [25–28]. These studies use a variety of methods for cell selection and isolation, reverse transcription, complementary DNA (cDNA) amplification and cell type clustering. However, despite these differences, studies examining the same regions and organs have consistently identified similar classes of cells, suggesting that some broad transcriptomic signals are robust to experimental methods and technical variation. For example, three recent studies of the portions of the mouse hypothalamus show significant overlap in cell types and specific marker genes for these types [19–21].

Despite the existence of multiple experimental protocols, the generation of transcriptome-wide single-cell RNA-seq data follows a standard overall procedure (Figure 1). First, cells of interest are isolated using fluorescence-activated cell sorting (FACS), manually, or through microfluidics, resulting in individual cells separated into distinct wells, tubes or droplets in a suspension. After collection and isolation, the cells are lysed and the RNA is reverse transcribed; selective reverse transcription of mRNAs is a common approach in single-cell RNA-seq, achieved with oligo-dT primers to select for polyadenylated transcripts. After reverse transcription, the resulting cDNA is amplified, fragmented and prepared for sequencing. The differences in the most commonly used experimental protocols result from decisions about whether to obtain whole gene-body coverage of transcripts or only the 3′ (or 5′) ends of the transcripts, the use of unique molecular identifiers to correct for amplification bias, the degree to which cDNA is pooled before amplification and the type of amplification itself (Figure 1, [29]). The output data after profiling are a set of sequencing reads, which are then

Vilas Menon is a researcher at the Howard Hughes Medical Institute Janelia Research Campus, studying how gene expression is linked to neuronal identity and phenotype.

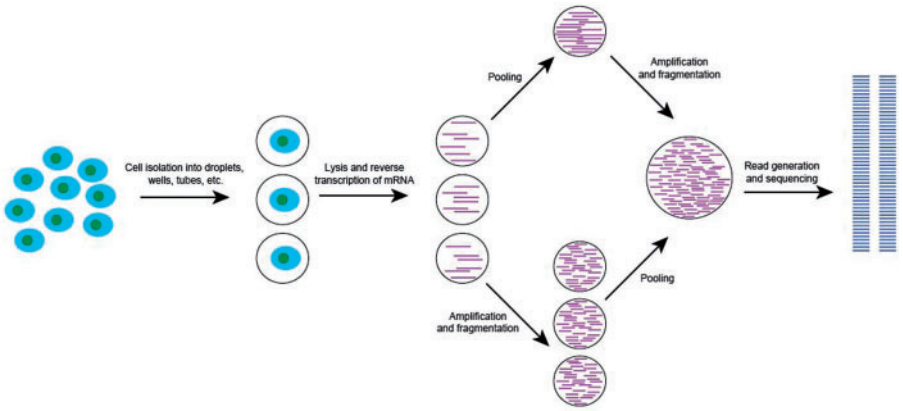

Figure 1. A simplified schematic of the overall strategy for single-cell RNA-seq. Cells are first isolated from a population into tubes, wells or droplets using FACS, manual selection or microfluidics devices. The cells are then lysed within their isolated environment, and their mRNA is reverse transcribed. At this stage, individual tubes/wells/droplets can be pooled if the reverse transcription step incorporates a cell barcode, and then the cDNA can be amplified and fragmented. Alternatively, the cDNA from each cell can be amplified and fragmented, adding on a sample-specific sequence, and then pooled. After pooling, the library of fragments is sequenced to generate the set of reads that is aligned to a reference transcriptome and genome.

mapped to the reference genome and transcriptome of the species of interest, and finally quantified to obtain estimated abundances for each mRNA species in each cell.

An important decision for any single-cell RNA-seq experiment is how to distribute sequencing reads: the options are to detect many transcripts in a fewer number of cells (i.e. conducting deeper sequencing per cell at the expense of cell number), or to perform shallow sequencing on a larger number of cells. Figure 2 shows the trade-off between cell number and read depth per cell, given different total read sequencing capacities, as well as cell number/read depth combinations explored by some recent studies. In general, the main constraint on the total number of reads for a given profiling study is budgetary—sequencing costs are decreasing, but remain a significant portion of the total budget of single-cell profiling experiments. Large-scale studies using droplet-based sequencing [7, 18, 19, 21] have surveyed >20 000 cells at <10 000 reads per cell (Figure 2), whereas targeted studies have surveyed many fewer cells at depths up to 50 million reads per cell [13]. This wide variation in the distribution of reads raises the question of whether certain computational approaches are better suited than others to identify putative cell types in various sampling strategies.

Identifying putative transcriptomic types from single-cell data requires clustering cells' gene expression profiles by their similarity. Whereas early studies used established clustering techniques [2, 22], more sophisticated computational methods have been developed using variants of principal component analysis (PCA) [7, 9, 18, 30], gene clustering [9, 11, 25, 26, 31] and matrix reordering methods [5, 14–16, 20]. The former two approaches rely on similar techniques (PCA versus gene clustering) to generate modules or components comprising multiple genes based on expression similarity, and then use this low-dimensional gene space to cluster cells. In contrast, matrix reordering methods simultaneously cluster genes and cells to obtain coherent 'blocks' in the expression matrix. In general, once putative cell types have been identified computationally, the existence of a subset of these types is probed through nontranscriptomic methods, including looking at spatial localization, functional properties or developmental trajectories. Studies using multiple flavors of computational methods [9, 18] have shown that different techniques largely agree, but this correspondence remains underexplored. Whether different classes of methods perform better on shallow read depth data versus deeper sequenced data has not been explored systematically, and is the purpose of this study.

In general, cell typing studies have advocated for both strategies of distributing reads, but rarely explore multiple classes of computational methods on the same data set. Deeper sequencing of fewer cells results in the detection of more transcripts per cell, and thus provides a more complete picture of the transcriptome of each cell (Figure 2); this allows for more confidence in assessing gene–gene correlations, as dropout of genes detected in a cell could artificially narrow the distribution of correlation values over the entire data set. In addition, deeper sequencing allows for cell types distinguishable primarily by moderate- to low-expressing genes to be well-separated. However, sequencing fewer cells may exclude or undersample rare cell types. The second strategy—sampling more cells at fewer reads per cell—aims to circumvent some of these issues. For clustering, the presence of many representatives of the same putative cell type provides a sufficient signal to link types together, even if any given pair of cells of the same type may have somewhat divergent observed transcriptomes because of shallow sampling [1, 7, 18] (Figure 2). Large-scale sampling is also more likely to detect rare cell types and subdivisions of existing types, as shown in several studies [1, 18]. Whereas each of the flavors of computational methods has been applied to data sets of varying read depth and cell number, a direct comparison of their ability to resolve cell types on the same data set is missing.

A direct comparison of three classes of methods—PCA-based, gene clustering-based and biclustering-based—has not been undertaken to date, especially with respect to their performance on high-depth versus low-depth sequencing data. Given that each class of methods contains variations on parameter selection, implementation and so on, this study selects a representative from each: the Seurat package [18, 32] as a PCA-based approach, the iterative WGCNA algorithm [9, 26] as a gene clustering approach and BackSPIN [5] as a biclustering-based approach (Figure 3A). Seurat first reduces dimensionality using PCA, and then clusters cells in PCA space using the Jaccard overlap to compute a cell–cell distance and the Louvain algorithm to identify clusters of cells. The iterative Weighted Gene Co-Expression Network Analysis (iWGCNA) approach first clusters genes into modules using WGCNA [33], and then clusters cells into groups in eigengene coordinate space; this is conceptually similar to PCA, as gene modules and principal components are

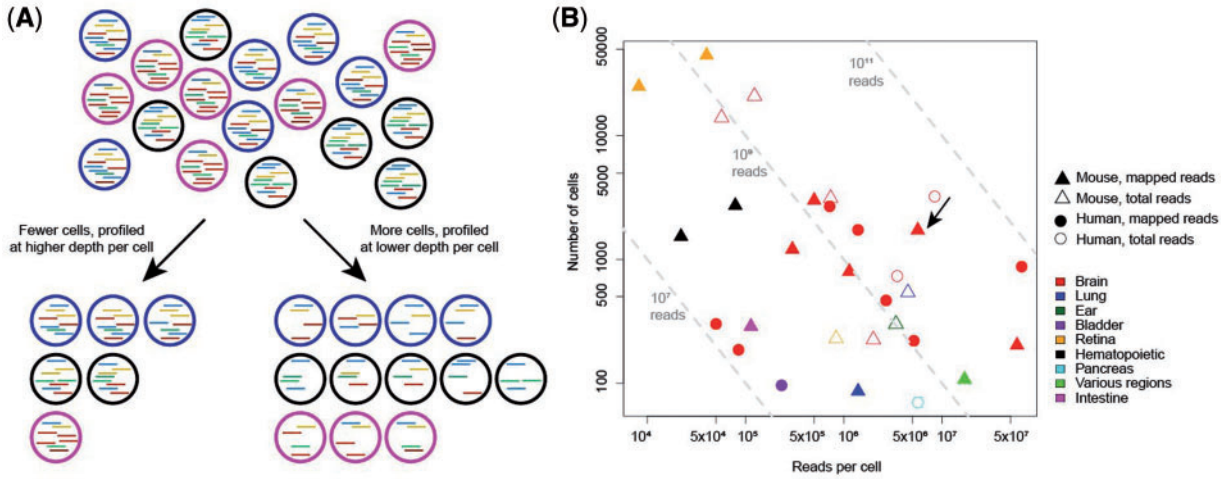

Figure 2. Distributing reads over cells. (**A**) Given a population of cells and a total number of reads available, reads can either be used to sequence fewer cells more deeply (right) or to sequence more cells at a shallower depth (left). Here, cell type identity and transcript species are indicated by different colors. Using the strategy on the left, there may not be enough cells sampled of a given type (pink) to identify a cluster. Using the strategy on the right, cells of a given type may not share enough transcriptional similarity to be identified as belonging to the same cluster. (**B**) Cell numbers and read depths for single-cell RNA-seq studies with the goal of identifying transcriptomic cell types. For certain studies, mapped read counts were not clearly stated, so overall read counts are reported—the number of reads with useful information in these studies is less than shown on the graph. The arrow indicates the study used for Figure 3.

both linear combinations of genes. However, gene modules are, by definition, significantly sparser in their membership than principal components. Finally, the BackSPIN algorithm rearranges the rows and columns of the expression matrix using the SPIN algorithm [34] and then iteratively divides cells and genes into subgroups so as to maximize their separation. Further details on the implementation and parameters for each of these algorithms can be found in their respective references.

Each of these three approaches can be run on matched low and high read depth data sets, generated successive subsampling of a well-characterized high read depth data set [9]. This data set comprises ~1700 cells isolated from the mouse visual cortex, and includes cells of known neuronal classes. In the original study, 42 putative neuronal cell types and 7 putative nonneuronal cell types were identified using a combination of PCA-based and gene clustering-based methods. Although the function of every individual putative type has not been exhaustively characterized using nontranscriptomic methods, the clusters can be arranged into broader categories of cells corresponding to previously known, functionally relevant biological classes with well-characterized gene markers. Because these established classes differ from each other substantially (as in the case of neurons versus nonneuronal cells) or more subtly (as in the case of L5a and L5b glutamatergic classes), they can be used to test discriminability over varying scales. For testing, six cluster sets were selected, from high to low distinctiveness: (1) neurons versus glia, (2) glutamatergic versus GABAergic neurons, (3) Vip+ versus Sst+ versus Pvalb+ GABAergic neurons, (4) superficial versus deep glutamatergic neurons, (5) Layer 5 versus Layer 6 glutamatergic neurons and (6) Layer 5a versus Layer 5b glutamatergic neurons. All of these comparisons contain at least 130 cells with at least 1 million reads mapped to exonic transcript regions, allowing for subsampling on both the read depth and cell number axes. Each comparison was subsampled along both axes, as shown in Figure 3B, and all three of the methods mentioned above were run on each data set. To address the sensitivity of any given clustering algorithm to parametrization, a semi-exhaustive parameter search was run for every clustering; this included the number of PCs and the Louvain algorithm resolution for Seurat, the gene cluster-adjusted *P*-value and dynamic branch cut parameter for iterative WGCNA and the cluster depth and number of genes used for BackSPIN. For each combination of subsampled data set and clustering algorithm, the clustering output best corresponding to the original separation (corresponding to known classes) was assessed to see if >95% of cells were segregated properly.

The three clustering methods perform differentially for low and high read depth data, as shown in Figure 3B. All methods are able to distinguish highly divergent and similar classes from each other at high read depth and high cell number, reflecting the overall ability of these methods in finding relevant clusters of cells. However, these methods show differing degrees of discriminability on subsampled data. Seurat performs substantially better on low read depth, high cell number data, especially when distinguishing similar classes of cells, such as Layer 5 and Layer 6 cells, or Layer 5a cells from Layer 5b cells. In contrast, iWGCNA and BackSPIN are better at segregating similar cell types when reads are distributed more deeply over a smaller number of cells, as opposed to more shallowly over more cells. In addition, neither iWGCNA nor BackSPIN performs consistently better than the other in this series of tests. Whereas all three methods easily distinguish highly distinct classes (neurons versus glia, or glutamatergic versus GABAergic cells) with low read depths and cell numbers, the finer distinctions are not discoverable by any of the methods beyond a minimum read depth and cell number, as evidenced by the blank rows and columns in the examination of finer distinctions in Figure 3B. This suggests that lower limits on read depth and cell number are an important consideration for cell typing studies: certain cell type distinctions may be unrecoverable if not enough cells or reads are acquired, even with 10-fold compensatory increases in read depth or cell number. Finally, for classes that are similar—such as Layer 5 versus Layer 6 neurons, or Layer 5a versus Layer 5b neurons—it appears that the lowest total read numbers for cluster separation correspond to scenarios where the reads are distributed at higher depth over fewer cells (see diagonal lines, Figure 3B). Although a full examination of why specific methods perform better at certain read depth regimes requires a separate study examining the underlying noise distributions and structure of signal correlations, some intuition can still be gained

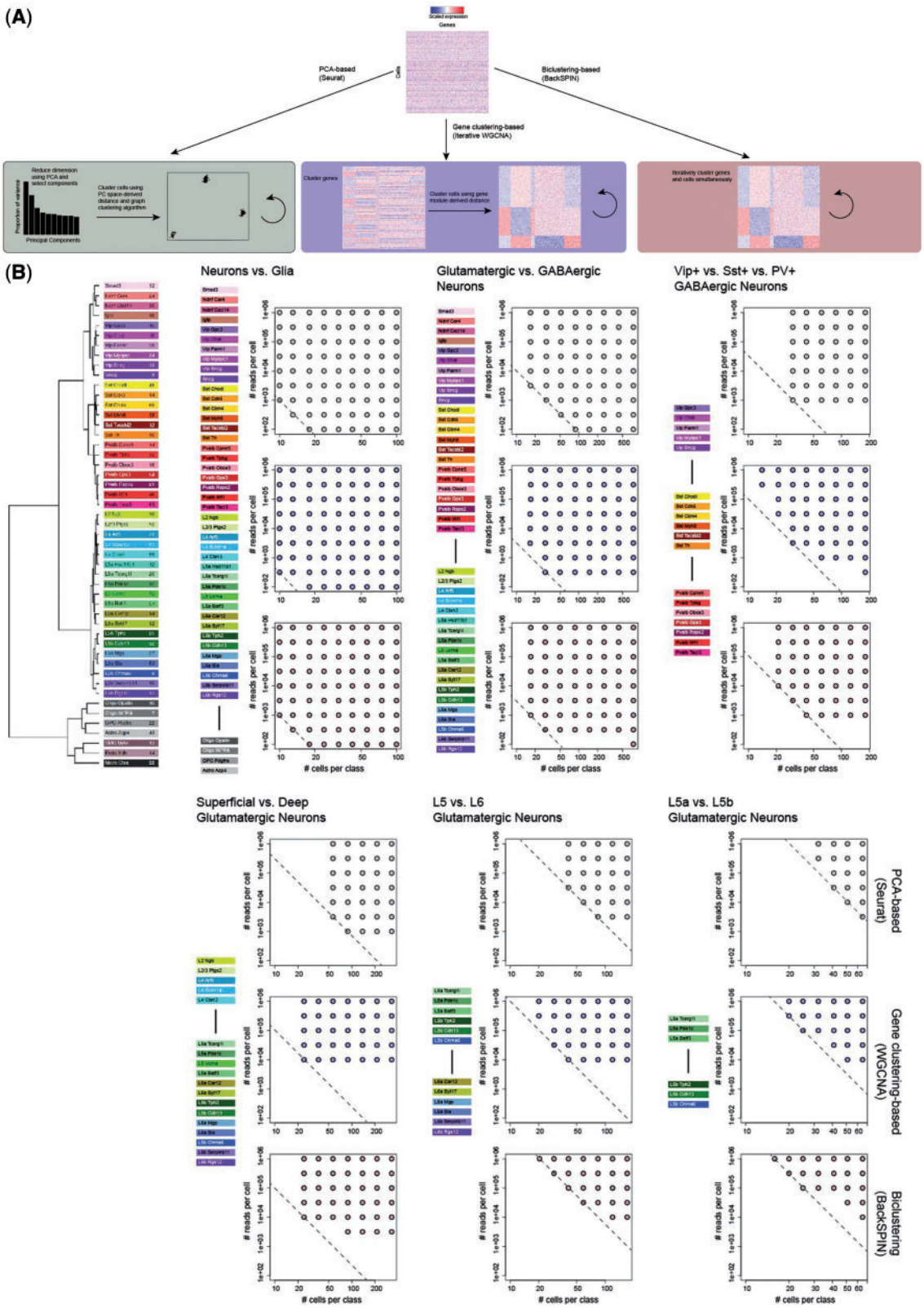

Figure 3. Performance of difference clustering algorithms on data sets with varying degrees of 'distinctness' among cell types, subsampled to different combinations of cell number and read depth per cell. (**A**) Schematic summary of the three clustering algorithms explored. In general, both PCA-based and WGCNA-based clustering can be applied iteratively on a data set until no further statistically significant subdivisions could be found. BackSPIN incorporates this iterative approach directly into the original algorithm. For the analysis in (**B**), the methods were not applied iteratively, as the subsets of clusters used for testing are themselves selected iteratively. (**B**) Clustering methods applied to successive subsets of clusters from a single-cell RNA-seq study of neurons [9]. Each column represents the ability of the three clustering approaches (from top to bottom—Seurat, WGCNA-based and BackSPIN) to distinguish classes of cells (represented by groups of colored bars to the left of each

here. With the caveat that any individual data set can have its own peculiarities, this analysis suggests that low-depth, high cell number data are more amenable to PCA-based methods, whereas gene clustering and biclustering methods can identify cell types better from low cell number, higher-depth data. For most studies, however, a safe recommendation is to apply methods from multiple classes (as in [2, 9]) to ensure that potential regions of transcriptomic space are not left underexplored.

The results presented here are based on a current need to balance cell number and read depth per cell, arising from biological and budgetary constraints. As sequencing costs continue to decrease, it will be possible to obtain higher depth data as a matter of course, up to the limit of the reverse transcription efficiency of the amplification method. For methods with low reverse transcription efficiency, increasing read depth per cell beyond the saturation limit will simply profile the same set of transcripts multiple times. If this saturation point can be determined, then the strategy to distribute reads is simple: profile as many cells as possible at the saturation read depth per cell, and select PCA-based methods if the saturation depth is low, or gene clustering and biclustering methods if the depth allows for the detection of many thousands of genes per cell. For single-cell profiling methods with higher reverse transcription efficiency, the saturation read depth is likely to be high, and thus would be better suited to analysis with gene clustering- or biclustering-based methods. Ultimately, the expectation is that sequencing costs will decrease to a level where large numbers of cells can be profiled at high depth. In parallel, new classes of single-cell clustering methods are in development, which suggests that a consensus approach using multiple methods is likely to become increasingly reliable and standard. Together, these advances will lead to a comprehensive picture of RNA-seq-derived cell types and their transcriptomic profiles, generating a definitive transcriptomic parts list for the biological system of interest.

Key Points

- Single-cell RNA-seq studies aim to classify cells into transcriptomic types and comprise a variety of techniques.
- Multiple computational approaches to classify cells exist, but most of them fall into one of three categories: PCA/ICA-based, gene clustering based or biclustering.
- Highly similar cell types are better distinguished by distributing reads deeply over fewer cells; however, populations with diverse cell types are better investigated by high-cell number data sets.
- PCA-based methods perform better on low read depth data, whereas gene clustering-based and biclustering methods perform better on intermediate and high read depth data.

Funding

Funding for this work was provided by the Howard Hughes Medical Institute.

References

1. Jaitin DA, Kenigsberg E, Keren-Shaul H, *et al*. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* 2014;**343**(6172):776–9.
2. Pollen AA, Nowakowski TJ, Shuga J, *et al*. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat Biotechnol* 2014;**32**(10):1053–8.
3. Usoskin D, Furlan A, Islam S, *et al*. Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. *Nat Neurosci* 2014;**18**(1):145–53.
4. Pollen AA, Nowakowski TJ, Chen J, *et al*. Molecular identity of human outer radial glia during cortical development. *Cell* 2015;**163**(1):55–67.
5. Zeisel A, Munoz-Manchado AB, Codeluppi S, *et al*. Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* 2015;**347**(6226):1138–42.
6. Darmanis S, Sloan SA, Zhang Y, *et al*. A survey of human brain transcriptome diversity at the single cell level. *Proc Natl Acad Sci USA* 2015;**112**(23):7285–90.
7. Macosko EZ, Basu A, Satija R, *et al*. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 2015;**161**(5):1202–14.
8. Okaty BW, Freret ME, Rood BD, *et al*. Multi-scale molecular deconstruction of the serotonin neuron system. *Neuron* 2015;**88**(4):774–91.
9. Tasic B, Menon V, Nguyen TN, *et al*. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat Neurosci* 2016;**19**(2):335–46.
10. Cadwell CR, Palasantza A, Jiang X, *et al*. Electrophysiological, transcriptomic and morphologic profiling of single neurons using Patch-seq. *Nat Biotechnol* 2015;**34**(2):199–203.
11. Thomsen ER, Mich JK, Yao Z, *et al*. Fixed single-cell transcriptomic characterization of human radial glial diversity. *Nat Methods* 2016;**13**(1):87–93.
12. Gokce O, Stanley GM, Treutlein B, *et al*. Cellular taxonomy of the mouse striatum as revealed by single-cell RNA-seq. *Cell Rep* 2016;**16**(4):1126–37.
13. Li CL, Li KC, Wu D, *et al*. Somatosensory neuron types identified by high-coverage single-cell RNA-sequencing and functional heterogeneity. *Cell Res* 2016;**26**(1):83–102.
14. La Manno G, Gyllborg D, Codeluppi S, *et al*. Molecular diversity of midbrain development in mouse, human, and stem cells. *Cell* 2016;**167**(2):566–80.e19.
15. Marques S, Zeisel A, Codeluppi S, *et al*. Oligodendrocyte heterogeneity in the mouse juvenile and adult central nervous system. *Science* 2016;**352**(6291):1326–9.

column of plots, with broader distinctions to the left and finer distinctions to the right) at various combinations of read depth and cell number. In each plot, the presence of a colored circle indicates that the method was successfully able to distinguish cell classes (with 95% of cells classified properly). Both axes are plotted on log-scale, such that any diagonal line with slope $= -1$ represents scenarios with the same total number of reads. Dashed diagonal lines indicate the lowest total read number required for a given method to distinguish the cell classes. For different classes of cells, there is no strong bias toward higher read depth, as shown by the relatively symmetric distribution of circles in the plots in the leftmost columns. For more similar classes of cells (toward the right), the clustering methods show different degrees and directions of skewing—Seurat works better on higher cell number, lower-depth data, whereas WGCNA-based and BackSPIN work perform better on higher-depth lower cell number data. For every clustering run, parameters were optimized to obtain the best possible agreement with the broad classes.

16. Habib N, Li Y, Heidenreich M, *et al*. Div-seq: single-nucleus RNA-seq reveals dynamics of rare adult newborn neurons. *Science* 2016;**353**(6302):925–8.
17. Lake BB, Ai R, Kaeser GE, *et al*. Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science* 2016;**352**(6293):1586–90.
18. Shekhar K, Lapan SW, Whitney IE, *et al*. Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. *Cell* 2016;**166**(5):1308–23.e30.
19. Chen R, Wu X, Jiang L, *et al*. Single-cell RNA-seq reveals hypothalamic cell diversity. *Cell Rep* 2017;**18**(13):3227–41.
20. Romanov RA, Zeisel A, Bakker J, *et al*. Molecular interrogation of hypothalamic organization reveals distinct dopamine neuronal subtypes. *Nat Neurosci* 2017;**20**(2):176–88.
21. Campbell JN, Macosko EZ, Fenselau H, *et al*. A molecular census of arcuate hypothalamus and median eminence cell types. *Nat Neurosci* 2017;**20**(3):484–96.
22. Treutlein B, Brownfield DG, Wu AR, *et al*. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* 2014;**509**(7500):371–5.
23. Durruthy-Durruthy R, Gottlieb A, Hartman BH, *et al*. Reconstruction of the mouse otocyst and early neuroblast lineage at single-cell resolution. *Cell* 2014;**157**(4):964–78.
24. Muraro MJ, Dharmadhikari G, Grun D, *et al*. A single-cell transcriptome atlas of the human pancreas. *Cell Syst* 2016;**3**(4):385–94.e3.
25. Close JL, Yao Z, Levi BP, *et al*. Single-cell profiling of an *in vitro* model of human interneuron development reveals temporal dynamics of cell type production and maturation. *Neuron* 2017;**93**(5):1035–48.e5.
26. Yao Z, Mich JK, Ku S, *et al*. A single-cell roadmap of lineage bifurcation in human ESC models of embryonic brain development. *Cell Stem Cell* 2017;**20**(1):120–34.
27. Furchtgott LA, Melton S, Menon V, *et al*. Discovering sparse transcription factor codes for cell states and state transitions during development. *Elife* 2017;**6**:e20488.
28. Camp JG, Badsha F, Florio M, *et al*. Human cerebral organoids recapitulate gene expression programs of fetal neocortex development. *Proc Natl Acad Sci USA* 2015;**112**(51):15672–7.
29. Poulin JF, Tasic B, Hjerling-Leffler J, *et al*. Disentangling neural cell diversity using single-cell transcriptomics. *Nat Neurosci* 2016;**19**(9):1131–41.
30. Kiselev VY, Kirschner K, Schaub MT, *et al*. SC3: consensus clustering of single-cell RNA-seq data. *Nat Methods* 2017;**14**(5):483–6.
31. Fan J, Salathia N, Liu R, *et al*. Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nat Methods* 2016;**13**(3):241–4.
32. Satija R, Farrell JA, Gennert D, *et al*. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* 2015;**33**(5):495–502.
33. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 2008;**9**:559.
34. Tsafrir D, Tsafrir I, Ein-Dor L, *et al*. Sorting points into neighborhoods (SPIN): data analysis and visualization by ordering distance matrices. *Bioinformatics* 2005;**21**(10):2301–9.