Cerebral cortical neuron diversity and development at single-cell resolution
Matthew B Johnson¹,²,³ and Christopher A Walsh¹,²,³,⁴,⁵,⁶

Over a century of efforts to categorize the astonishing diversity of cortical neurons has relied on criteria of morphology, electrophysiology, ontology, and the expression of a few transcripts and proteins. The rapid development of single-cell RNA sequencing (scRNA-seq) adds genome-wide gene expression patterns to this list of criteria, and promises to reveal new insights into the transitions that establish neuronal identity during development, differentiation, activity, and disease. Comparing single neuron data to reference atlases constructed from hundreds of thousands of single-cell transcriptomes will be critical to understanding these transitions and the molecular mechanisms that drive them. We review early efforts, and discuss future challenges and opportunities, in applying scRNA-seq to the elucidation of neuronal subtypes and their development.

Addresses
¹ Division of Genetics and Genomics, Boston Children’s Hospital, Boston, MA, USA
² Manton Center for Orphan Disease Research, Boston Children’s Hospital, Boston, MA, USA
³ Howard Hughes Medical Institute, Boston Children’s Hospital, Boston, MA, USA
⁴ Department of Pediatrics, Harvard Medical School, Boston, MA, USA
⁵ Department of Neurology, Harvard Medical School, Boston, MA, USA
⁶ Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA

Corresponding author: Walsh, Christopher A (christopher.walsh@childrens.harvard.edu)

Introduction
The classification of cell types in the cerebral cortex has challenged the greatest minds in the history of neuroscience, and so perhaps it is no surprise that we do not quite have it figured out yet. Ramon y Cajal and other early histologists described the two major cortical neuronal types — those with long, distantly projecting axons and those with short, locally projecting axons — and documented their many morphological variations [1]. Brodmann, Campbell, Vogt and others used the distribution of morphological types to subdivide the cortex into cytoarchitectonic areas which we now understand have important functional correlates [2–4]. Yet, whereas classical neuroscientists reached consensus around the neuronal types in brain structures like the cerebellum over a century ago, the effort to develop a comprehensive neuronal ‘parts list’ for the cortex has lagged. Electrophysiological and circuit analyses arrived in the mid-twentieth century with new tools and the idea that morphological and functional classes of neurons might somehow correspond, though the labor-intensive nature of combining electrophysiology and morphology has limited the ability to integrate form and function. The revolution in molecular biology of the late 20th century allowed an integration of developmental lineage, inferred from the expression of a few marker genes [5], yet still it is not clear whether these criteria can define a clean, non-overlapping ‘periodic table’ of cortical neuronal types, or whether instead the classification of cortical neurons is inherently less precise than in other brain areas, with a mix of some sharply defined classes and other, fuzzier categories [6,7]. This review will focus on the relevance of single-cell transcriptomics to the classification of cortical neuron subtypes by genome-wide gene expression, and explore the unique perspective afforded by scRNA-seq on the dynamic processes of cortical neurogenesis and differentiation.

Transcriptomic classification of neuronal cell types
Pioneering single-cell microarray and qRT-PCR studies elucidated progenitor and neuronal subtypes in the mouse brainstem [8], olfactory system [9], retina [10,11], inner ear [12], and embryonic cortex [13,14], as well as developing human and ferret cortex [15]. Now, single-cell RNA-seq has opened the floodgates for deep transcriptomic analysis of CNS cell types [16,17**,18–26] (see also recent review by [27]). Although some early scRNA-seq studies have tested specific hypotheses — for example, confirming the ‘one neuron-one receptor rule,’ that each individual primary olfactory neuron expresses one and only one olfactory receptor gene [28–31] — most have aimed to elaborate on the molecular identities of classically defined neuronal types, discover new types, and begin to establish definitive brain cell type taxonomies (Table 1). These studies employ a generalizable two-stage approach to scRNA-seq-based cell type classification. In the first stage of analysis, single-cell transcriptomes are grouped through
### Table 1

Summary of experimental methods and main results for selected scRNA-seq studies classifying cell types in the mammalian brain. UMI, unique molecular identifiers; ERCC, spike-in synthetic control RNAs.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Cell selection and isolation method</th>
<th>cDNA type</th>
<th>Number of cells used (total cells sequenced)</th>
<th>Source of cells assayed</th>
<th>Sequencing depth (avg or median reads per cell)</th>
<th>Numbers of genes detected (average per cell and/or cumulative) and used for classification</th>
<th>Cell types identified/classified</th>
<th>Other notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kodama [8]</td>
<td>Manual dissection and cell picking</td>
<td>qRT-PCR (3-Prime end)</td>
<td>167 (208)</td>
<td>Mouse medial vestibular nucleus</td>
<td>N/A</td>
<td>59 hand-picked genes 6 neuronal types: 3 excitatory, 3 inhibitory</td>
<td>Further subdivisions likely, but classification correlates with known morphological and functional subtypes</td>
<td></td>
</tr>
<tr>
<td>Saraiva [29]</td>
<td>FACS selection followed by Fluidigm C1</td>
<td>Full-length (SMARTer) with ERCC</td>
<td>21 (58)</td>
<td>Mouse olfactory sensory neurons</td>
<td>4.4 Million</td>
<td>4717 detected per cell; 13 582 total; 509 genes found to be differentially expressed between individual neurons</td>
<td>18 known cell types, one confirmed new cell type</td>
<td>Confirmed the 'one-neuron-one-receptor' hypothesis</td>
</tr>
<tr>
<td>Uoskin [17**]</td>
<td>Manual dissection and automated cell picking</td>
<td>5-Prime (STRT) with UMI</td>
<td>622 (799)</td>
<td>Mouse dorsal root ganglion</td>
<td>1.1 Million</td>
<td>3574 ± 2010 detected per cell; 12 750 used to ID initial 4 neuronal subgroups &amp; non-neuronal cells; 11 658 used for final iterative clustering of 11 subtypes</td>
<td>11 neuronal subtypes</td>
<td>68 ‘outliers or unresolved ID’ (6.5% of all cells sequenced)</td>
</tr>
<tr>
<td>Darmanis [35]</td>
<td>Fluidigm C1</td>
<td>Full-length (SMARTer)</td>
<td>466 (482)</td>
<td>Human adult temporal lobe and fetal cortex</td>
<td>2.8 Million</td>
<td>~4000 detected per cell 7 neuronal types: 2 excitatory, 5 inhibitory; one fetal progenitor class</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tasic [33**]</td>
<td>Manual dissection followed by single-cell FACS into microtiter plates</td>
<td>Full-length (SMARTer) with ERCC</td>
<td>1679 (1739)</td>
<td>Mouse (8 wks) primary visual cortex</td>
<td>8.7 Million</td>
<td>7278 detected per cell; 13 878 used for classification 49 ‘core’ cell types: 19 excitatory, 23 inhibitory, 7 non-neuronal; plus ‘intermediate’ cells w/mixed identity between two or more ‘core’ types</td>
<td>255 (15.2%) good cells are of ‘intermediate’ neuronal subtype</td>
<td></td>
</tr>
<tr>
<td>Zeisel [32**]</td>
<td>Manual dissection ± FACS selection followed by Fluidigm C1</td>
<td>5-Prime (STRT) with UMI</td>
<td>3005 (3315)</td>
<td>Mouse (3–5 wks) somatosensory cortex &amp; hippocampus</td>
<td>0.5 Million</td>
<td>~4500 detected per cell; 15k total detected; top 5000 by variance used for classification 47 cell types: 7 excitatory neuron, 16 interneuron, 2 astrocyte, 6 oligodendrocyte, and 2 immune classes</td>
<td>310 outlier/poor quality cells (9.4%)</td>
<td></td>
</tr>
</tbody>
</table>
a combination of dimensionality reduction and hierarchical clustering, with varying degrees of iteration and supervision. In the second stage, the resulting cell groups are contrasted against each other to identify differentially expressed marker genes. This approach has so far met with great success in marrying transcriptomic classifications to known cell types. As sample sizes expand and studies proliferate, new methods to standardize these taxonomies and map other data modalities onto transcriptomic cell types will be critical.

The cerebral cortex: the ultimate cell type diversity challenge

Three of the largest scRNA-seq studies of neuronal identity published to date have, fittingly, tackled the most heterogeneous brain region, the cerebral cortex [32**,33**,34**]. In the first such study, unbiased sampling of mouse primary somatosensory and hippocampal cortex identified many non-neuronal cell types in addition to seven excitatory and 16 inhibitory neuronal types, corresponding well to existing layer-defined and marker-defined classes [32**]. Acknowledging that an unbiased sampling captures few cells from rare populations, the authors also oversampled a subtype of interneurons by FACS isolation, enabling them to find a novel subtype of Pax6+ neurogliaform cells in layer I, nicely validated by immunohistochemistry and electrophysiology [32**].

In contrast, selection of known neuronal populations by microdissection and FACS purification from reporter mice yielded a more diverse taxonomy of 42 neuron classes [33**], reassuringly concordant with the selected laminar and marker-based populations, but also further subdividing many of these classes into putative novel subpopulations. In addition, these authors’ classification algorithm is relatively lenient with regard to cell type ambiguity, assigning ~15% of cells an ‘intermediate’ identity between two neuronal classes. In the most extreme case, one third of layer IV neurons were classified as intermediate between two of the three proposed layer IV subtypes. What remains to be determined is the extent to which these intermediate cells reflect algorithmic ‘oversplitting,’ or subdivision of cell types based on transcriptional variability that is in fact stochastic or state-related rather than subtype identity-dependent. Further studies will be necessary to differentiate state versus trait transcriptional signals in these and any other proposed novel neuronal types.

Adult human brain presents particular challenges for single-cell studies, given the highly myelinated and dense extracellular milieu of the tissue and the typical storage conditions of postmortem samples, which make clean dissociation of intact whole cells difficult [35]. There are, however, reliable and reproducible protocols for isolating single neuronal nuclei from frozen postmortem human brain [36,37]. Fortunately, the nucleus contains
a significant amount of messenger RNA, and several studies have now demonstrated single-nucleus RNA sequencing [38–40]. Lake et al. have taken this approach for scRNA-seq-based cell type classification in the human cerebral cortex, identifying 16 neuronal subtypes — 8 excitatory and 8 inhibitory [34**]. Most intriguingly, these authors also detected differences between neocortical areas, for example, between the layer IV transcriptomic subtypes present in frontal versus occipital cortex.

**Merging transcriptomic taxonomy with other data types**

Given that traditional methods of cell type classification are much lower throughput than scRNA-seq, they are likely to be increasingly guided by single-cell transcriptomic taxonomies, through the ability to collect the transcriptome of a single cell that has already been characterized by another method and map that transcriptome onto large scRNA-seq reference data sets. Several groups have recently demonstrated the collection and sequencing of RNA from cells following electrophysiological recordings [41*,42,43]. In one study, 45 CCK+ interneurons in cortical layer I were first characterized by patch-clamp recording and classified into five subtypes based purely on their electrophysiological properties [41*]. Each cell’s contents were then aspirated into the patch pipette and the RNA sequenced. The 45 single-cell transcriptomes were then mapped onto 16 interneuron subtypes defined previously [32**]. Interestingly, none of the five electrophysiological subtypes corresponded to a single transcriptomic subtype, with cells from each group mapping to two, three, or even four different scRNA-seq-defined interneuron classes. Similar discord was observed for excitatory neurons.

Discrepancy between electrophysiological and transcriptomic cell type could arise trivially from the technical noise inherent in scRNA-seq, or from differences in experimental conditions. However, it is notable that electrophysiology seems to provide greater unity of CCK+ interneuron subtypes compared to scRNA-seq, again suggesting that the algorithm applied to the larger scRNA-seq data set [32**] may have over-split some cell types. Alternatively, the methods used to map one taxonomy onto the other may improve with inclusion of additional *a priori* knowledge; about a third of the Patch-seq interneurons assayed in somatosensory cortex mapped to subtypes predominantly found in hippocampus in the prior scRNA-seq study. The electrophysiological properties of Patch-seq cells were also highly correlated with their expression of 24 of 167 genes encoding relevant proteins — channels, pumps, receptors, so on. It would be informative to cluster the larger scRNA-seq sample using the same 167-gene panel and map the Patch-seq interneurons onto the resulting taxonomy, which should improve the correspondence between the two data modalities. Indeed, overlapping but distinct sets of genes are likely to determine distinct neuronal properties, including morphology, electrophysiology, and connectivity, as well as developmental processes, like migration, that interact with these features. Although several studies have found a nearly 1:1 correspondence between morphology and electrophysiology for layer I interneurons [42,44], such correspondence is strikingly lacking for interneurons in the rest of the cortical layers [44], supporting the partial independence of these properties. It is likely that classifications based on the summation of all of these transcriptional signals will yield neither unambiguous subtypes nor clear correspondence to individual cellular properties, and the challenge rather will be to deconvolve single-cell transcriptional profiles into the distinct signatures that correspond to each domain of neuronal properties — electrophysiology, location, dendritic arborization, axonal projection, so on.

**Droplets and the advantages of 10× higher throughput**

Although the methods employed by the vast majority of scRNA-seq studies to date — manual or automated cell picking [17**,28,31]; single-cell flow cytometry [33**]; and/or microfluidics (i.e., the Fluidigm C1 system) [19,29,30,32**,34**,35] — may capture as much as 25% of each cell’s mRNA, they are limited in throughput, and by the cost-per-cell for library construction and sequencing (Table 2). An alternative approach involves encapsulating single cells in nanoliter-volume droplets and performing lysis and barcoded reverse transcription within the droplets [26,45,46*]. For comparison, to assay 3300 cortical cells [32**] required weeks of cumulative C1 run time; manual screening of cDNA samples; and dozens of library preparations and sequencing lanes.

<table>
<thead>
<tr>
<th>Table 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pros and cons of current single-cell capture and processing methods.</strong></td>
</tr>
<tr>
<td>Cell capture method</td>
</tr>
<tr>
<td>Manual/automated cell picking</td>
</tr>
<tr>
<td>Flow cytometry</td>
</tr>
<tr>
<td>Microfluidics</td>
</tr>
<tr>
<td>Droplets</td>
</tr>
</tbody>
</table>

* Fluidigm’s high-throughput chip is expected to capture up to 800 cells per run, with some level of on-chip barcoding.
whereas a droplet-based study of the mouse retina [46] generated cDNA from 49,300 cells, pooled into seven libraries, over the course of four days — that is, 15 times the number of cells assayed by one fifth the amount of sequencing, and probably less than one tenth the hands-on processing time. The two main drawbacks of current droplet-based methods are (1) only about 10% of a cell’s mRNA molecules are captured; and (2) the lack of full-length cDNA generation limits alternative splicing analyses (Table 2), although technological improvements are likely to ameliorate both of these disadvantages in the near future.

Increasing throughput by three orders of magnitude alleviates the pressure to choose between a hypothesis-driven, targeted design that may miss unknown cell types [33**] and an unbiased approach that undersamples minority populations [32**]. In fact, analyzing 49,300 retinal cells first required excluding a large proportion of the rod photoreceptors that, making up two thirds of the retina, masked the transcriptional signatures differentiating other cell types [46]. Following this in silico selection step, 39 cell types were defined based on a ‘training set’ of ~13,000 cells, and the remaining ~36,300 cells were classified by correlation to the training set. Notably, the large sample size enabled identification of known and novel neuronal subtypes as rare as 0.1%. Remarkably, however, these subtypes are still far fewer than the 60–100 retinal cell types defined by morphology and physiology [47]. In particular, retinal ganglion cells are known to be highly diverse and yet, because they compose only about 0.5% to 1% of cells in the retina, the current sample of 49,300 cells was insufficient to detect heterogeneity within the ~500 RGCs assayed. Thus, even with the high throughput of droplet-based methods, careful experimental design will be paramount, and a logical expectation is that a complete cell type taxonomy from any tissue that contains rare subpopulations will require a tiered approach of initial unbiased sampling followed by targeted subpopulation studies using known or novel markers to isolate the rarest cells.

Progenitor heterogeneity of the human fetal cortex
Remarkably, while only a single scRNA-seq analysis of mouse embryonic cortex has so far been published [48], a large number of studies have applied single-cell transcriptomics to fetal human cortex [15,35,49*,50*,51], motivated by the intriguing morphological heterogeneity of primate cortical neural progenitor cells (NPC) [52–55]. A major contributor to the diversity of primate NPC is the relative abundance of basal or outer radial glia (ORG), which morphologically, functionally, and transcriptionally resemble apical radial glia of the ventricular zone (VZ) but are located in the subventricular zone (SVZ). Two studies employing scRNA-seq to uncover the distinct transcriptional program of ORG employed different selection methods to enrich for their NPC parent population, and notably arrived at similar results [49*,50*]. By manually microdissecting the VZ/inner SVZ from the outer SVZ before scRNA-seq, Pollen et al. were able to subsequently correlate gene expression with germinal zone location and thus identify genes specifically enriched in ORG [49*]. At the same time, Thomsen et al. developed a protocol for light fixation, permeabilization, and fluorescent immunolabeling compatible with FACS purification of NPC followed by scRNA-seq [50*]. Remarkably, both efforts identified several of the same ORG-enriched genes, including HOPX, FAM107A, and TNC. Nevertheless, the sample sizes of these studies are small compared to those on the adult cortex, and much remains to be done. It is important to note that single progenitor transcriptomes are heavily influenced by cell cycle phase, and probably also reflect dramatic changes in fate potential over the course of cortical neurogenesis. Indeed, several ORG-enriched genes were found to be expressed by VZ NPC slightly earlier in development [49*]. Altogether, there is a strong case for further developmental studies not only including greater cell numbers but also sampling a wide range of time-points, and applying new analysis methods that can probe the dynamic development and complex lineage relationships of the developing cortex.

Assessing validity and utility of in vitro models by scRNA-seq
A key use of scRNA-seq will be to validate in vitro models of human brain development and disease by comparing the cell types and developmental dynamics of these models to primary human tissues [51,56,57]. Cerebral organoids are fast becoming a popular model for early human brain development, but neither the variability across iPSC lines or individual organoids nor the correspondence of organoids to in vivo brain development is yet fully understood. An scRNA-seq analysis of cerebral organoids at 33–65 days post-differentiation identified both dorsal and ventral telencephalic NPC and neurons, as well as a few cells resembling those of the cortical hem signaling center, an important source of patterning morphogens [51]. Importantly, parallel analyses of organoids and primary human fetal cortex found a remarkable similarity in transcriptional programs of neurogenesis and differentiation, the biggest difference being a paucity of basal NPC in organoids (6% of cells in the organoids vs 34% in 12–13 weeks of gestation human cortex). Future studies will be needed to determine whether this discrepancy reflects the absence or underdevelopment of an important progenitor niche in the organoids, or simply a mismatch in the developmental stages of the organoids and primary tissues compared in this study.

Challenges and opportunities for future studies
Single-cell technologies and data analysis methods continue to improve rapidly, and will be invaluable in creating
a complete census of cell types and lineage relationships in the brain. We foresee future improvements leading to great opportunities in four general areas. Methods to isolate neuronal cell populations from non-genetic model species will be of particular importance. The FRISCR method [50] has great potential, but relies on suitable antibodies and known cell type markers. Promising alternatives include using fluorescent in situ hybridization or RT-PCR reactions to sort cells on the abundance of mRNA transcripts [58,59], and merging these methods with droplet-based scRNA-seq will be hugely advantageous. In the model of Patch-seq, protocols are sorely needed for sequencing RNA from single cells previously or concurrently characterized by other methods, for example, physiology, connectivity, developmental lineage, or live imaging. Methods to either maintain [60] or reconstruct [61,62] spatial information in conjunction with scRNA-seq need further development for application to mammalian brain studies. Similarly, emerging methods to sequence DNA and RNA from the same single cell [63,64] will provide critical insights into the lineage relationships between cell types, which are otherwise extremely difficult to assess in human brain. New statistical models have been developed to improve gene expression level estimates and quantify heterogeneity in noisy single-cell data [65,66]. Further development should be aimed at integrating these models with advanced clustering and pseudotime methods. Large-scale developmental studies will require new algorithms for inferring cell type lineages from scRNA-seq data collected at multiple real time points, while normalizing for cell cycle phase. Validation of putative novel subtypes will be paramount, and will require innovative approaches to visualize transcriptional dynamics in single cells over time. The true test of cell type identity is stability, and thus time-lapse live imaging of single-cell transcription will be needed to definitively differentiate transcriptional cell type from cell state.

Conflicts of interest
Nothing declared.

Acknowledgements
M.B.J. is supported by the Paul G. Allen Family Foundation. C.A.W. is an Investigator of the Howard Hughes Medical Institute.

References and recommended reading
Papers of particular interest, published within the period of review, have been highlighted as:

* of special interest
** of outstanding interest

signals to activate dormant neural stem cells. Cell 2015, 161:1175-1186.


The first study to apply scRNA-seq to neuron classification in the cerebral cortex highlights the advantages and drawbacks of unbiased sampling of a complex tissue, and uses targeted subsampling to identify a novel interneuron subtype.


This group’s unique classification strategy allows for ‘intermediate’ cell types that exhibit mixed transcriptional signatures between two cortical neuron subtypes. The need for flexible categorization may reflect the complexity of the cerebral cortex and the degree of interplay between ‘state’ (e.g., highly plastic synaptic connectivity and activity) and ‘trait’ (cell type).


This study highlights the potential of single-nucleus RNA-sequencing for analyzing difficult-to-dissociate tissues such as postmortem adult human cortex. Analysis of several Brodmann areas also provides an exciting first look at how transcriptional cell type composition varies across functional domains of the cortex.


This study exemplifies the opportunities and challenges in merging neuronal phenotypes derived from different data modalities, and provides strong evidence that in the cerebral cortex, neuronal types defined by scRNA-seq may not map uniquely onto the types derived from electrophysiology and/or morphology.


Droplet-based scRNA-seq developed by this group and by Klein et al. [45] expands throughput by orders of magnitude, enabling higher sample sizes needed for heterogeneous tissues like cortex. Here, a novel analysis in retina also provides a template for unbiased sampling of tissues that contain both highly abundant and very rare cell types of interest.


This study and Thomsen et al. [50] used distinct experimental designs to arrive at similar results searching for genes enriched in known subtypes of human fetal cortex progenitors. These authors performed extensive validations, and find important developmental differences that highlight the need for larger studies across more ages.


