TITLE:

Implementation and validation of single cell genomics experiments

Co-authors (in alphabetical order):

Marco Colonna[1]*, Genevieve Konopka[2]*, Shane Liddelow[3-6]*, Tomasz Nowakowski[7-10]*, Rajeshwar Awatramani[11], Helen Bateup[12-14], Cathryn R Cadwell[7,9,15], Emre Caglayan[2], Jerry L Chen[16-18], Jesse Gillis[19,20], Martin Kampmann[21,22], Fenna Krienen[23], Samuel Marsh[24-26], Michelle Monje[27,28], Michael R O'Dea[3], Rickie Patani[29,30], Alex Pollen[31,32], Francisco J. Quintana[33,34], Marissa Scavuzzo[35], Matthew Schmitz[30,31], Steven Sloan[36], Paul Tesar[35], Jessica Tollkuhn[20], Maria Antonietta Tosches[37], Madeline E Urbanek[38], Jonathan Werner[9,20], Omer Bayraktar[39]*, Ozgun Gokce[40,41]*, Naomi Habib[42]*

AFFILIATIONS:

1. Department of Pathology and Immunology, Washington University School of Medicine in St. Louis, St. Louis, MO, USA
2. Department of Neuroscience, Peter O'Donnell Jr. Brain Institute, UT Southwestern Medical Center, Dallas, TX., 75390, USA
3. Neuroscience Institute, NYU Grossman School of Medicine, New York, NY., 10025, USA
4. Department of Neuroscience & Physiology, NYU Grossman School of Medicine, New York, NY., 10025, USA
5. Department of Ophthalmology, NYU Grossman School of Medicine, New York, NY., 10025, USA
6. Parekh Center for Interdisciplinary Neurology, NYU Grossman School of Medicine, New York, NY., 10025, USA
7. Department of Neurological Surgery, University of California, San Francisco, CA, USA
8. Eli and Edythe Broad Center for Regeneration Medicine and Stem Cell Research, University of California, San Francisco, CA, USA.
9. Weill Institute for Neurosciences, University of California, San Francisco, CA, USA
10. Department of Anatomy, University of California, San Francisco, CA, USA
11. Department of Microbiology and Immunology, Northwestern University, Chicago, IL, USA.
12. Department of Molecular and Cellular Biology, University of California Berkeley, Berkeley, CA 94720, USA
13. Helen Wills Neuroscience Institute, University of California Berkeley, Berkeley, CA 94720, USA
14. Chan Zuckerberg Biohub, San Francisco, CA, USA
15. Department of Pathology, University of California, San Francisco, San Francisco, CA, United States.
16. Department of Biomedical Engineering, Boston University, Boston, MA 02215, USA.
17. Center for Neurophotonics, Boston University, Boston, MA 02215, USA.
18. Department of Biology, Boston University, Boston, MA 02215, USA.
19. Department of Physiology and Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Canada
20. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA

21. Institute for Neurodegenerative Diseases, University of California, San Francisco, CA 94158, USA
22. Department of Biochemistry and Biophysics, University of California, San Francisco, CA 94158, USA
23. Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08544, USA
24. F.M. Kirby Neurobiology Center, Boston Children's Hospital, Boston, Massachusetts, USA
25. Harvard Medical School, Boston, MA, USA
26. Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA
27. Department of Neurology and Neurological Sciences, Stanford University, Stanford, CA, USA
28. Howard Hughes Medical Institute, Stanford University, Stanford, CA, USA
29. Department of Neuromuscular Disease, UCL Queen Square Institute of Neurology, London, UK.
30. The Francis Crick Institute, Human Stem Cells and Neurodegeneration Laboratory, London, UK.
31. Eli and Edythe Broad Center of Regeneration Medicine and Stem Cell Research, University of California, San Francisco, San Francisco, CA., USA
32. Department of Neurology, University of California, San Francisco, San Francisco, CA., USA
33. Ann Romney Center for Neurologic Diseases, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA.
34. Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA.
35. Department of Genetics and Genome Sciences, Case Western Reserve University School of Medicine, Cleveland, Ohio 44106, USA
36. Department of Human Genetics, Emory University School of Medicine, Atlanta, GA, USA
37. Department of Biological Sciences, Columbia University, New York, NY., 10027, USA
38. Biomedical Sciences Graduate Program, University of California, San Francisco, CA, USA
39. Wellcome Sanger Institute, Cambridge, UK
40. Institute for Stroke and Dementia Research, LMU University Hospital, Ludwig-Maximilians University Munich, Munich, Germany
41. Munich Cluster for Systems Neurology (SyNergy), Munich, Germany.
42. Edmond & Lily Safra Center for Brain Sciences, The Hebrew University of Jerusalem, Jerusalem, Israel

CORRESPONDANCE:

* mcolonna@wustl.edu; genevieve.konopka@utsouthwestern.edu; shane.liddelow@nyulangone.org; tomasz.nowakowski@ucsf.edu; ob5@sanger.ac.uk; ozguengoekce@gmail.com; naomi.habib@mail.huji.ac.il

ABSTRACT:

Single-cell/-nuclei transcriptomics (sc/snRNAseq) is a powerful tool for identifying cell types and states. However, the gene expression information must be validated, and its functional relevance needs to be established. The choice of validation depends on numerous factors. Here, we present types of orthogonal and functional validation experiments to strengthen

preliminary findings obtained using sc/snRNA-seq as well as the challenges and limitations of these approaches.

MAIN TEXT:

Single-cell/-nuclei transcriptomics (sc/snRNAseq) is a powerful tool for studying various aspects of biology, evolution, and diseases, such as identifying and characterizing cell types and cell states, detecting gene expression and epigenetic changes in disease and dysfunction, inferring developmental trajectories and cell state transitions, predicting gene regulatory mechanisms, and comparing evolutionary modifications in the specific tissues across species. However, the high-throughput nature of these experiments and the ever-growing array of computational tools for analyzing data they produce also increases the risk of false discoveries. Therefore, to confirm the preliminary findings of single-cell genomics experiments, validation experiments ought to be performed using orthogonal and functional methods (**Figure 1, Table 1**). The choice of validation experiments depends on the specific single-cell finding being tested and the evolutionary hypothesis being addressed. In this review, we discuss several general use cases and examples of single-cell findings that can be validated using different approaches (see **BOX 1**, **Table 1**). In **BOX 3**, we discuss why evolutionary comparisons are important for the neuroscience field from both a basic science and a translational perspective. We also discuss the challenges and limitations of single-cell genomics for evolutionary comparisons, such as disentangling homology versus convergent evolution, relating in vitro models to in vivo biology, and accounting for technical and biological variability.

We present some examples of single-cell findings that may require validation experiments, covering a variety of common scenarios and methods. These are not exhaustive, but illustrate how different types of validation can complement and confirm single-cell results. The simplest confirmatory experiments are those which seek to validate the expression of a small number of genes *in situ*. In the first two examples above, such an approach using *in situ* hybridization with probes targeting a cell type-specific gene and the gene of interest would provide confirmation of the single-cell finding. Protein-level validation could also be examined using immunostaining. However, these methods of validation are not suitable for validation of large numbers of gene expression changes, or in cases when several genes are necessary for the identification of a cellular state or multiple states (**BOX 1**, use case #3). To validate such claims, multiplex *in situ* methods like MERFISH or genome wide spatial sequencing methods like 10X Visium may be necessary. With a growing array of spatial transcriptomics and proteomics methods, careful consideration of the pros and cons of each method is warranted. Finally, many single-cell analysis findings reach beyond description of gene expression and infer putative function and gene regulatory mechanisms, as in the cases of cell-cell communication analysis and gene regulatory network inference[1]. While spatial visualization of gene expression or downstream protein abundance may be useful for demonstrating co-localization of a ligand and receptor, for example, visualization alone is not sufficient to demonstrate biologically meaningful functions, bona fide cell-cell communication, or transcription factor-enhancer/promoter interactions. Validating these mechanistic and functional inferences may require perturbation experiments (see **Figure 2**) and further functional studies, which we discuss in depth below.

Primary validation approaches

Validation of clusters versus validation of individual differentially expressed genes

Single cell genomics is a powerful approach to identify molecular phenotypes that contribute to cellular diversity[2]. By facilitating the simultaneous comparison of profiles from a collection of cells, single cell genomics can define clusters based on identified differentially expressed genes (DEGs)[3-5]. However, algorithms used to process and analyze single cell datasets are designed to detect transcriptional differences, and depending on the resolution specified, they will continue to subset cells even if the differences are so minute that they constitute noise[6]. It is essential to set several resolutions in clustering followed by iterations of computational post-hoc validation prior to biological validation (see companion piece on sequencing technologies). The inherent features of the clustering algorithms that provide modularity of the system also reinforce the importance of post-hoc validation methods to ensure cluster numbers reflect real biology.

An important first step is loyal cell type calling by integrating novel experimental datasets into large-scale atlases or databases. The tools used can be either classical integration methods with the external dataset of interest (to understand if the cell type at question has an exact correspondence to the published cell type), or transcriptomic similarity approaches (to understand if the new cell type is similar to an existing cell type). The latter can be accomplished by using tools like AddModuleScore() in Suerat[7], or tools like Celltypist[8] and CellHint[9]. Further computational methods to validate clustering numbers include visualization of most highly enriched transcripts per cluster (top 10 or top 100) and projection across all cells by cluster[10]. This method allows to see the level of transcriptional discernment between cells in each group: the best clustering illustrates a diagonal line where each transcript is enriched relatively uniform across all cells in a cluster, but is lowly expressed in all other cells across other clusters (dot plots, violin plots, and heatmaps are routinely used). If there is high expression of cluster-enriched genes outside of the candidate cluster, this indicates that there is overclustering and too high of a resolution indicated, and that clusters likely represent the same cells or subtle changes between cell states. It is best to initially set the resolution high to overcluster before modifying the resolution, slightly lowering it until robust cluster separation is able to be visualized.

An additional computational method to ensure clustering isn't an artifact of similar cells being inputted into analysis and the algorithm forcing differences that aren't there is to pulse in a known disparate cell types to see if the prior clusters collapse. For example, many groups subcluster cells from larger datasets to allow identification of more subtle, cellular subtypes or states that are otherwise hidden in larger datasets due to divergent cell types in the larger dataset. While this method is standard and acceptable, adding in a different cell type is a stringent method to determine if these subtypes or states are in fact truly different.

Finally, and most importantly, all clustering by single cell data necessitates post-hoc validation in biological samples[11]. Single cell data should be viewed as a prediction of biology that allows for guided hypothesis generation. Single cell data should not be considered fact or provide conclusions in the absence of confirmation in tissue or cells[12]. Different visualization methods can be employed to validate clusters by assessing cluster-enriched or cluster-specific transcripts or proteins in cells or tissue; these methods are discussed below. Targets should be

visualized overlapping with other cluster-enriched or cluster-specific markers, but separate from other clusters. These methods also allow for detection of DEGs between clusters.

DEGs can thus be extrapolated as rudimentary markers of cell types and/or states. However, these transcriptomic profiles only provide a snapshot of cellular behaviors. Given that identification of DEGs is an inherently comparative process, additional genes that may be playing a pivotal role in driving cell identity may escape the radar of DEG analysis. For example, a pivotal interaction between two genes may be necessary for supporting a cell's function. In the absence of one of those genes, a different population of cells is actually losing the combined effect of both; while this first gene may appear in a DEG analysis between these two clusters, the second gene will go largely unnoticed[12-14]. Herein lies the need for orthogonal and functional validation of individual DEGs, as the complex mechanisms and interplay of DEGs remain largely hidden within transcriptomic data.

Standard approaches for orthogonal validation aim to recapitulate the DEGs identified by sequencing data. As a first-pass check, orthogonal validation is a crucial intermediate in determining whether a DEG derived from a cluster is recapitulated within a cellular context relevant to the biological question at-hand. Depending on the parameters used for DEG discovery, as well as the condition of samples prior to sequencing, DEGs that define a cluster in a given dataset may not translate back to a tissue setting. The expression level of DEGs may also fluctuate even within a cell's normal state, thus cell types or identities cannot be declared in lieu of lineage tracing and developmental analysis. It then becomes important to demonstrate that any genes of interest are truly and consistently being expressed at different levels between cell states of interest before moving towards mechanistic interrogation, this orthogonal evidence includes but is not limited to the description of the epigenomic, morphological, spatial, and biophysical properties or function of cells[11].

After orthogonal validation has been completed, functional validation approaches can be utilized to investigate how DEGs may be driving cell state. Typically, functional validation is reliant on the conclusions of both genomic analyses and orthogonal validations, as these studies can provide insight into possible gene interactions and correspondent cell states that correlate to a DEG of interest. Functional studies continue to serve as the gold standard, as the physiological relevance of different clusters cannot be inferred from the transcriptional state, it can only serve as a guide towards cell type assessment.

Understandably, prior cell type classification schemes have been heavily weighted towards molecular assays due to the unprecedented scale and throughput of single-cell genomics[15]. However, as noted above, there are a number of issues that can arise when single-cell genomics assays are interpreted in a vacuum: 1) it is difficult to distinguish molecular features that define stable cell types from transient cell states, 2) the resulting cell type atlases may vary depending on the sample size and analytical parameters used for clustering, leading to lack of reproducibility with no clear ground truth and 3) the functional relevance of molecularly-defined cell types is unclear. Recent multimodal single cell analyses call into question the very notion of discrete cell types, suggesting that continuous and correlated variation in cellular morphology,

biophysical properties and molecular features contributes substantially to cellular diversity within broad transcriptomic classes[16]. True validation of transcriptomic clusters, i.e. confirming that they are meaningful and worthwhile to study, requires orthogonal and functional validation. One of the major reasons for requiring such orthogonal validation of in silico clusters is that many large cell atlas studies are grossly underpowered for lowly abundant or rare cells or cell states, often leading to artifacts of clustering. Some methods have been developed to overcome this, such as FINDseq (focused interrogation of cells by nucleic acid detection and sequencing)[17] – developed for the study of rare astrocyte populations isolated on the basis of the expression of a few mRNA markers. Additional validation steps, including alternative sequencing efforts, multi-dataset integration and meta-analyses, visualization (e.g. in situ, MERSCOPE, etc.) are integral to validate the biological truth behind computational modeling and clustering.

Orthogonal validation of transcriptomic clusters rests on the hypothesis that bona fide cell types should form discrete entities regardless of the particular assay used. In other words, if a group of cells segregates as a distinct population using multiple assays, this would support its designation as a valid cell type or cell state. The number and type of assays needed to validate a new cell type remain unclear, but at a minimum it is recommended that findings from sc/snRNAseq be validated using at least one independent assay (e.g. visualization using in situ or spatial transcriptomics, scATAC to highlight chromatin accessibility for DEGs, or functional assay). However, we would encourage investigators and the field as whole to move beyond this minimum requirement to validate new cell types not only at the level of individual genes and transcripts, but at the level of proteins, cellular physiology, developmental lineage, morphology, connectivity and in vivo function. These additional axes of phenotypic variability will establish the robustness of the cell type under investigation and provide a more mechanistic understanding of its functional role in the nervous system in vivo.

Methods of validation by visualization

Spatial transcriptomics

Spatial transcriptomics, in conjunction with in situ hybridization (ISH), including single molecule ISH or RNAscope), immunofluorescence (IF), and immunohistochemistry (IHC), represents a powerful combination of techniques for comprehensive characterization of gene expression and protein localization within tissues. While sc/snRNAseq provides information on transcriptomes at the single-cell /-nuclei level, ISH enables the visualization of specific RNA molecules directly within intact tissues, confirming their spatial distribution and validating sequencing results. Furthermore, IF and IHC techniques allow the detection and localization of proteins within tissue sections, providing additional information on cell types, protein-protein interactions, and cellular functions. Integrating these complementary techniques not only verifies sc/snRNAseq findings but also allows researchers to study the co-expression of genes and proteins in the context of tissue architecture, providing a more comprehensive understanding of cellular behavior and molecular interactions within complex biological systems. By leveraging the strengths of spatial transcriptomics, ISH, IF, and IHC, researchers can unravel the intricate spatial dynamics of gene expression and protein localization, advancing our understanding of tissue development, disease pathogenesis, and potential therapeutic targets.

While each of these methods requires the a priori knowledge of DEGs from sc/snRNAseq experiments, the advent of spatial transcriptomics has revolutionized the validation of sequencing data by enabling the integration of spatial information with transcriptomics analysis – often without having a starting dataset. There are two general categories of spatial transcriptomics technologies: sequencing based, and *in situ* based. The genome wide, nature of Visium enables the researcher to localize groups of DEGs (often called 'gene modules') in tissue sections. Visium captures 3' ends of transcripts via poly-A tail. Therefore, although pre-selection of targets is not required, non-poly-A transcripts are missed. By combining spatially resolved gene expression profiling with high-throughput sequencing, Visium technology allows researchers to validate and complement sequencing data.

Sequencing based methods such as Visium and Slide-seq[18] employ a spatially barcoded array of capture probes, enabling the simultaneous capture of gene expression information from multiple spatially defined regions within a tissue sample. The resulting genome-wide, untargeted spatial transcriptomics data provides a comprehensive view of gene expression patterns in their native tissue context, with 50-100 µm resolution. *In situ*-based methods such as MERFISH[19], STARmap[20] , and *in situ* sequencing[21] provide high-throughput, direct identification of RNA transcripts at sub-cellular resolution of panels of several hundred genes by single molecule fluorescence in situ hybridization (FISH) with sequential imaging and signal amplification techniques. Since *in situ* imaging-based methods readout facilitates the visualization of spatial gene expression patterns in the context of tissue morphology, enabling a more comprehensive understanding of cellular organization and function.

Both technologies provide a validation step for sequencing experiments enhances our understanding of the spatial organization of gene expression, cell-cell interactions, and tissue architecture. They offer the flexibility to study both coding and non-coding RNA species to corroborate sequencing findings, unravel novel biological insights, and identify spatially restricted gene expression gradients. Through the integration of spatial transcriptomics, researchers can attain a deeper understanding of the spatial dynamics of gene expression, ultimately advancing our knowledge of tissue development, disease mechanisms, and potential therapeutic targets.

Gene expression versus protein

Use of spatial transcriptomics or imaging-based RNA visualization to validate the results of sc/snRNAseq results, and/or the elucidate spatial context of results, provides important context and orthogonal validation. However, it is critical when inferring potential functional consequences of results to consider protein-level validation. While transcript and protein levels often exhibit fairly good agreement there are a number of regulatory and other mechanisms that can lead to dichotomy in transcript:protein ratios[22-24]. One area where protein-level validation provides critical information is in the inference of cell-to-cell communication. The power of analysis on the level of single cells have led to a rapid expansion of methods to assess/infer cell-to-cell communication through ligand-receptor interactions from scRNAseq and/or spatial transcriptomics data[25-27]. Results from these tools, however, should be considered hypothesis generating and not hypothesis validating. Beyond assessing whether ligands and receptors truly

interact, a lack of confirmation that ligands and receptors are expressed at the protein level and in appropriate spatial context represents a critical hole in many of these analyses. Expression and/or spatial colocalization can be accomplished by relatively straightforward techniques such as IHC or flow cytometry/cyTOF and examining physical interaction can be performed via co-IP or newer techniques such as “nativeomics”[28]. Finally, functional studies to alter a proposed cell-to-cell communication network provide a critical piece of the puzzle which also must be assessed.

Methods of validation by interrogation

Single-cell transcriptomic approaches uncover differentially expressed genes and transcriptomic signatures between different cell types and cell states. These data are descriptive and do not establish causality or mechanism. However, they can generate mechanistic and functional hypotheses for properties of different transcriptomic states and signatures, and for upstream regulators of transcriptomic signatures. Perturbation-based approaches can test these functional hypotheses, interrogate causality, and link gene expression to cellular function.

The complementary approach to the observation of transcriptomic changes is the targeted manipulation of transcript levels. A number of methods exist for the perturbation of genes in cultured cells and model organisms. RNA interference technology enables knockdown of mRNAs by synthetic short interfering RNAs (siRNAs) or transgenically expressed short hairpin RNAs (shRNAs), but suffers from pervasive off-target effects.

CRISPR-based approaches have fewer off-target effects and provide an expanding toolkit to manipulate genes and their products, including genome and epigenome editing, control of gene expression and post-transcriptional mRNA processing. CRISPR technology, which enables gene knockout to achieve a complete loss of function, has transformed our ability to interrogate gene function in a scalable and precise manner[29]. CRISPR-based approaches, including base editors and prime editors, can also introduce precise genome edits to interrogate the genome with high resolution[30]. The CRISPR interference (CRISPRi) and CRISPR activation (CRISPRa) approaches use a catalytically inactive Cas9 protein to recruit transcriptional regulators to genomic sites of interest, enabling the modulation of expression levels of endogenous genes over many orders of magnitude[31], and thereby providing a strategy to directly model changes in expression levels of specific genes that are observed in single-cell transcriptomic studies. CRISPRi/a also can target distal regulatory elements, such as enhancers, to establish their function in controlling gene expression[32,33]. Other CRISPR-based tools can edit the epigenome[34] or affect post-transcriptional processes such as splicing and mRNA stability[35].

CRISPR-based gene perturbation can target genes of interest in individual experiments, but also in large, massively parallel screens using pooled sgRNA libraries targeting genes of interest. Pooled screens can be conducted for a large range of phenotypes, including cell survival, cellular functions and states read out by fluorescent markers or reporters, and single-cell RNA sequencing (using the Perturb-seq[36-39] or CROP-seq[38] approaches). They can also be used for screening of cis-regulatory regions (e.g. in iPSC-derived microglia[40]) and lincRNAs[41]. Such CRISPR screens can also be implemented with spatial readouts, using pooled optical

screening approaches, or spatial transcriptomics that uncover sgRNA identity by in situ sequencing.

CRISPR-based screens have recently been implemented in cell types relevant to neuroscience, including human iPSC-derived neurons[42], microglia[43], astrocytes[44], and brain organoids[45], and in mouse brains in vivo[46,47]. CRISPR-based screens in relevant cell types provide a scalable approach to generate a genome-wide “look-up-table” for the impact of expression levels of a given gene on relevant cellular functions. When possible, these CRISPR-based validation studies should be performed in vivo, to minimize artifacts linked to some in vitro, cell isolated, systems[48]. This can then be used to predict functional consequences of changes in gene expression that are observed in single-cell transcriptomic datasets

Functional validation

Following transcript level validation (and characterizing translational status) as described above, it is crucial to consider whether distinct clusters of cells possess defining functional properties. Such inference is not trivial and, as alluded to above, transcriptionally distinct clusters should not be assumed to always associate with functional differences within a cell type or cell type differences. Indeed, due to the heuristic nature of clustering tools, some separation of single cell RNAseq data is almost inevitable and so must be interpreted with appropriate caution in terms of its biological relevance. Cluster fidelity has been successfully studied through evolutionary conservation[49]. It is noteworthy that fleeting changes in cell state such as stage within the cell cycle may have greater transcriptomic impact than that cell type or substate[50]. The degree of functional plasticity in this context should also not be underestimated[51]. These issues notwithstanding, single cell sequencing has revealed relatively consistent signatures across adult differentiated tissues, predictably correlating with changes in cellular morphology for example[52].

One reductionist approach to begin to interrogate function of a cluster of cells involves use of (induced or embryonic) pluripotent stem cell (PSC) models, either using two or three dimensional culture systems[53-55]. Through ontogeny-recapitulating directed differentiation or forward programming approaches, these can be used to generate enriched populations of c differentiated cells for further study[56,57]. Astrocyte reactive state can be used as an exemplar, with the comparison of untreated PSC-derived astrocytes versus an induced reactive state following treatment with TNF, IL1α, and C1q[58,59]. Clusters of interest can be isolated through live sorting (e.g. FACS or MACS) and after revalidation of transcriptional state, consensus homeostatic functional attributes can be interrogated (e.g. sodium dependent glutamate uptake, synaptogenesis assays, cytokine secretome, non-cell autonomous neuronal survival etc.) and separately the gain of entirely new functions (e.g. neurotoxicity for astrocytes). These could represent ‘workhorse assays’ in trying to infer a functional correlate for transcriptomically defined heterogeneous states, perhaps comparing to other cluster(s) that lack the transcriptional profile of interest. Identification of transcripts within a functionally defined cell can alternatively / also be performed retroactively through post-hoc RNA FISH or analogous approaches decreased above.

Reducing this approach to practice in general terms would mean generating and isolating cell cultures with particular transcriptomic signatures that are viable for such functional characterization assays, which may present different challenges for different cell types and/or sub-states. Validation by genetic manipulation approaches listed above is critical before 'final' sub-state assignment. Where single genes or gene sets implicated by statistical prioritization (e.g. by significance and effect size), necessity and sufficiency for a particular function (or set of functions) can be genetically interrogated using the approaches described above. The detection of functional differences would have good positive predictive value for the role of a particular gene or gene set. However, if a functional difference is not detected the negative predictive value is somewhat limited because there may be specific functions that have not been considered or are not easy to assay. Additionally, the requisite functional validation may vary by cell type, cellular sub-state and must be defined in relation to the research question / scope of analysis. Orthogonal validation of key findings is crucial e.g. using ex vivo or in vivo approaches and indeed in human postmortem tissue where applicable.

Additional functional validation can be seeded by results from new methods for the study of cell-cell interactions between cells. Examples include recent investigations involving astrocytes and other glial cells, such as RABID-seq (rabies barcode interaction detection followed by sequencing[60]), SPEAC-seq (systematic perturbation of encapsulated associated cells followed by sequencing) which enables droplet-based culture of putative interactive cells[39], and LIPSTIC (Labelling Immune Partnerships by SorTagging Intercellular Contacts[61]) for cell-cell interactions first used to uncover interactions between T cells and dendritic cells.

Notwithstanding all these excellent tools, both new and old, some confusion still exists about what the definition of 'functional validation' means. Broadly the field uses this term to link the "genotype" as measured through RNAseq data with a "phenotype" as a way to confirm that the gene modules identified in sc/snRNAseq are biologically meaningful. Recently, in vivo imaging of neuronal activity has been combined with post-hoc with spatial transcriptomics[62-64] to provide functional validation of putative molecularly defined cell types. These experiments demonstrate a general workflow allows for many relevant phenotypes (neuronal activity, structural plasticity, neuron migration, etc) measured using fluorescence-based live cell imaging technologies (two-photon microscopy, confocal microscopy, lightsheet microscopy etc) to followed up with gene expression measurements to assess their validity.

There are a number of defining features of CNS cells (i.e. neurons and glia) that allow them to be grouped into different types. The unique morphology of different neuron and glia types was documented by Ramon y Cajal as one of the first examples of cataloging and classifying distinct cell types in the nervous system. With the advent of electrophysiological recording methods, different types of neurons could also be characterized by their intrinsic membrane properties and synaptic properties. Developments in neural circuit tracing have allowed the dissection of neighboring neuron types that participate in distinct circuits (e.g. direct and indirect pathway neurons of the striatum). More recently, optogenetic and chemogenetic tools have allowed selective manipulation of specific cell populations to probe their behavioral functions. These classically defined morphological and functional properties of neural cells can now be integrated

with transcriptomic and molecular features as discussed above to provide a more complete cellular taxonomy of the nervous system[65].

For cell types that have been described by scRNAseq, it is important to perform functional validation to 1) ensure that these transcriptionally-defined clusters represent a true cell type and 2) understand their properties in order to place them into the context of the larger circuit or brain region. To do this in animal models, it may be necessary to generate genetic reporter animals that label a given cell type based on its expression of a marker gene[66,67]. This strategy has been valuable to catalog the properties of many different types of neurons in the brain[68]. In many cases, however, cell types cannot be selectively labeled by expression of one marker gene alone, as other cell types may also express that gene to some extent. In that case, it can be useful to employ an intersectional strategy whereby cells expressing two genes (or expressing one gene but not another) can be selectively targeted[69]. As an example, this strategy has been used to study different types of neuromodulatory neurons, including dopamine and serotonin neurons, which have been shown by single cell profiling to be highly heterogeneous[70-73]. These studies have demonstrated that specific subpopulations of these cells, initially defined transcriptionally, indeed have different anatomical localization, cellular morphology, connectivity, and/or electrophysiological properties. These genetic reporter animals can also be useful for mapping the *in vivo* properties of cell types in the nervous system including their activity patterns and contributions to behavior.

Functional validation of neuron types has been performed in several large-scale cellular taxonomy papers[65]. A more general strategy moving forward could be to start with cell types defined by sc/snRNAseq, then perform spatial transcriptomics to demonstrate anatomical localization (discussed above), then record from and fill cells with a fluorescent dye to examine intrinsic membrane (or synaptic) properties, and perform morphological reconstruction (measure the properties of dendrites/spines/axons, etc), or perform Patch-seq to pair electrophysiological recordings with gene expression data from the same cell[74,75]. A further step could include generating a reporter animal that labels that cell population, which may require an intersectional approach with two or more gene drivers defining the cell type. The in vivo circuit connectivity and behavioral relevance of that specific cell population can then be tested. This workflow need not only be relevant to animal models without the need to have a different strategy for functional validation of cell types identified in brain organoids. Recent work in human iPSC-derived neuron/astrocyte organoid provided functional validation of astrocytes from these organoids at several ‘ages’, performing a variety of assays that probe known astrocyte functions (e.g. phagocytosis ability, calcium signaling, synaptogenic properties, developmental outgrowth of branches[76]). In organoids, cell morphology, protein expression (IHC), differentiation potential (for progenitors) and intrinsic physiological properties (for neurons) can still be measured as functional validation. Similarly, functional validation of astrocytes in different ‘reactive’ and disease states has also seen success. Starting with transcriptomic data as a roadmap for ‘what’ a reactive sub-state of astrocytes look like in vivo, researchers can isolate primary rodent or iPSC-derived human astrocytes, recapitulate the original gene expression signature, then continue with functional validation to determine if any loss-of- or gain-of-function changes are present in each sub-state. Several examples exist for this method producing high throughput

and controllable platforms for validating astrocyte disease biology[58,59,77-79]. And while many fields struggle with an internal dialogue about the biological relevance of in vitro cell-based systems, it is important to note that nearly all cell functions have been identified using these systems – which then require validation back in vivo to ensure discovered functions are not artifacts of the culture system. This includes neuron trophic support[80,81], synaptogenesis[82,83], synapse pruning/phagocytosis[84,85], neurotoxicity[58,59,78], among many more. In the case of astrocytes, it is hard to state a single astrocyte function that was discovered in vivo.

How to distinguish "cell types" from "cell states" in snRNA-seq data

A complicated question raised when interpreting sc/snRNAseq data is whether clusters defined by DEGs are in fact terminally differentiated cell types/states, or if they are transient clusters that are some way along a trajectory of differentiation to a terminal state. While both options are likely important and of interest to understanding much complex underlying biology, validation (in particular functional validation) of transient cell states is difficult.

Perhaps the most fundamental question is 'what is this cell?' – a focus on which cell types have been collected and sequenced. By analyzing the gene expression profiles obtained from sc/snRNAseq data, researchers can identify distinct transcriptomic signatures associated with different cell types. Cell types are characterized by a set of specific marker genes that are consistently expressed across cells of the same type. These marker genes define the core identity of a cell type and help distinguish it from other cell types. While some cells can be defined by individual genes (e.g. *Rbfox3* for neurons, or *Aldh1l1* for astrocytes in the CNS), others may require multiple markers to disentangle closely-related cell types (e.g. specific combination of genes encoding multiple neurotransmitters for neuron). On the other hand, cell states refer to transient or context-dependent changes in gene expression within a given cell type – often in response to external stimuli like infection, disease, or trauma. By examining the expression levels of marker genes and assessing their consistency across cells, researchers can differentiate between stable cell types and variable cell states within a population, however determining how reliable these states are has caused some confusion in particular fields.

This can be somewhat mitigated by sequencing large numbers of the cell of interest, either by pre-enrichment, isolation using genetically encoded fluorophores or surface antigens, or by brute force and sequencing millions of cells. Cluster analysis can be applied to these well-powered datasets to identify distinct cell populations based on gene expression patterns. Clustering algorithms group cells with similar gene expression profiles together, enabling the separation of different cell types. However, it is important to consider that clustering alone may not completely distinguish cell types from cell states, as it may group cells with similar transcriptional profiles into subpopulations representing different cell states within the same cell type. Further analysis, such as trajectory inference or pseudotime analysis, can help identify cell states along developmental or activation trajectories within a particular cell type, though these methods may also have problems (see below).

Some differentiation trajectories can be determined using traditional fate mapping of cells[86-89], or they can be predicted in silico by sequencing multiple timepoints across a developmental time

period or reactive/disease response by an individual cell type. Such trajectory inference methods allow researchers to explore the continuous transitions and cell fate trajectories within a homogeneous cell type, but heterogeneous cell state population. By ordering cells along a trajectory based on their transcriptomic profiles, it becomes possible to capture both cell types and cell states. Trajectory inference methods reveal the underlying biological processes, such as cell differentiation, lineage commitment, or cellular response to stimuli. By visualizing the trajectory and examining gene expression changes along it, researchers can distinguish stable cell types that represent distinct branches and transient cell states that occur along the trajectory.

Distinguishing cell types from cell states in sc/snRNAseq data, while seeming simple, remains one of the most hotly contested spaces in modern biology. Some computational methods exist, for instance one can take the mRNA splicing information of a cell and perform an automatic computation of whether the cell state is terminal or not (i.e. the "CellRank" package[90]). Moreover, recent tools have been developed to address whether perturbations in transcription factor states correspond to cell state emergence as opposed to cell type[91]. Researchers use an array of methods including analyzing transcriptomic signatures and marker gene expression consistency, clustering to separate cell populations, and employing trajectory inference methods to capture dynamic cell states along developmental or activation trajectories. Validating and characterizing these cell states through functional experiments is crucial for a comprehensive understanding of cellular heterogeneity and the functional implications of different states within a given cell type, and it is often only at this validation stage that a failure to recapitulate a particular gene expression signature is realized as the discovery of a transient state (see also companion piece on planning and executing sequencing experiments).

When and how does validation fail?

Often validation fails due to poorly designed initial sequencing experiments. This can be due to contamination of cell types of interest in bulk sequencing efforts, or more common with the uptick in sc/snRNAseq studies – a failure to properly power the cell type of interest (an example is the gross underpowering of astrocytes in whole CNS cell atlases – often less than 5% of all cells sequenced[92]). Similar artifacts arise when insufficient biological replicates are sequenced – often due to high costs of single cell/nuclei sequencing experiments[93-95]. Even when these initial pitfalls are overcome, computational errors can cause havoc for the validating biologist. While great computation power exists, and new tools are produced continuously, "grey" areas still exist in the field, which could explain why validation fails in some instances. **BOX 2** provides a brief overview of common validation difficulties from a computation standpoint.

The process of validating single-cell/nuclei sequencing experiments is complex and fraught with potential pitfalls. Poorly designed initial sequencing experiments, such as contamination or underpowering of cell types of interest, can undermine subsequent validation efforts. Computational errors and limitations in existing technologies further contribute to the challenges faced by validating biologists. The directionality of RNA velocity analyses can be perturbed by overlaying it on existing two-dimensional spaces, and tools for cell tracing and stage prediction have limitations that require further investigation. Assumptions about mRNA-protein correlation

in cell-cell communication analyses and the reliance on databases for known interactions introduce potential inaccuracies. Data integration tools are valuable but may artificially smooth analyses and obscure real biological differences. Additionally, the functional validation of aligned cell types across organisms remains essential, as conservation of circuitry and functional properties cannot be assumed. Although challenges exist, emerging technologies such as RNA editing and cell-based assays offer promise for improved validation across species. Careful consideration and exploration of validation options are necessary to ensure the reliability and robustness of single-cell/nuclei sequencing findings.

Cross-species comparisons

Single cell sequencing provides a molecular 'common language' definition of cell types across any species with a quality genome. While the definition of a cell type typically invokes other features (e.g., connectivity, function, and morphology), these features can be difficult or impossible to acquire in many species or in high throughput[15,96]. Moreover, cell-autonomous gene expression programs are the foundation on which many (but not all) structural and functional features of a cell are built[97]. This foundation of shared gene expression programs and functional properties across species enables an inference process termed 'homology mapping.' Properties such as connectivity and physiology are far easier to study in genetically tractable and experimentally accessible animal models (e.g., *Mus musculus* or *Drosophila*), and then can be 'transferred' through anchoring to homologous cell types in other species, even humans[98].

Identifying homologous cell types across species ideally involves identifying sets of cells in each species that access similar regulatory programs for their differentiation[99]. Single cell sequencing combined with lineage tracing or fate mapping is powerful for reconstructing the developmental history of cell types[100], and hence their relationships across closely related species. We still have a fragmented understanding of lineages of transcriptionally-defined cell types in any one species[101-104] and few comparative studies have attempted to match progenitor classes across species (but see[98,105]). The challenge with matching homologous types from adult data alone is distinguishing shared evolutionary history from phenotypic convergence[101,106], although with large enough sets of species this challenge can be alleviated[107]. Transcription factors (TFs) potently specify cell type identity, suggesting that prioritizing TFs in cross-species cell type mapping may improve homology assignments. A single TF or small set of factors can be sufficient to switch the fate and ultimate cellular identity. However, TFs are developmentally regulated and may not be conserved between the early stages of specification and adulthood.

Across species, cell types change their abundance, gene expression profiles, and spatial context. Each of these carries its own challenges for cross-species comparisons. In general, cell type similarity decreases – and homology mapping becomes less accurate – with increasing evolutionary distance[98,107-111]. snRNA-seq has been used to reveal conservation and novelty of brain cell types across 500 million years of evolution[106,107,112,113]. Integrating transcriptional profiles from single cells across species can be difficult because approaches may rely on assumptions of 1:1 orthologous genes[114], and gene duplications, losses, and sequence level divergence increase with evolutionary distance. Evolutionary modifications of cell types may result from neutral drift, physical constraints associated with brain reorganization, or new

functional requirements. Overall patterns of transcriptional divergence have been linked to neutral drift among primates coupled with stabilizing selection over longer time scales at the level of tissues and cell types[107,115].

Understanding the neuroanatomical and/or physical constraints that drive evolutionary features such as proportional shifts in cell types across species remains challenging. In some cases, increases in the abundance of a cell type across species can be intuitively linked to species-specific adaptations, such as the proportional increase of a retinal ganglion cell subtype in primates that may relate to neocortically-driven adaptations to high visual acuity[107]. However, sometimes the mechanism driving the differential modification of each cell type is difficult to resolve. For example, the observed reduced proportions of subcortically-projecting cortical neurons in larger mammalian brains[116] may relate to functional requirements to maintain scaling relationships between upper and lower motor neurons despite disproportionate cortical expansion, or to shifts in the migration of homologous types during development, with either mechanism leading to different anatomical distributions. Another example is the recently observed relative increase in the proportion of oligodendrocyte progenitor cells compared to mature oligodendrocytes in human compared to non-human primate brain[117]. This difference might support enhanced neuronal or myelin plasticity in the adult human brain or perhaps some other phenotype that has not yet been linked to oligodendrocyte function. Increased sampling through single cell genomics surveys both across species and across individuals within a species will help to distinguish between processes of drift and selection, while further analysis of scaling relationships and functional changes will be required to resolve the contribution of physical constraints and new cellular specializations. Arbitrating between such possibilities would enable a better understanding of the targets of evolutionary selection. Finally, and as discussed in other sections, issues like overfitting to a single species “reference” dataset, and differences in genome quality across species both add technical complexity to cross-species comparisons.

Conceptual limitations of transcriptomic-based homology inference

Importantly, homology describes phylogenetic relationships and is not a synonym of “similarity”. Cell types may have similar transcriptomes because they descend from a common ancestor, or because they acquired these properties by convergence after evolving under similar selective pressures (see **Figure 3**). The ideal way to discriminate between these two possibilities is to sample many species and reconstruct ancestral states using the principle of parsimony (convergent characters tend to lack “phylogenetic continuity”). Because this is not always feasible for detailed single cell genomic characterization, here we highlight some observations that may help with comparing transcriptomic data across species.

Data from neuronal cell types where homologies were established by independent criteria (e.g., morphology, input-output connectivity, etc.), offer two key insights. First, although the transcriptomic divergence of homologous neurons is generally a function of their phylogenetic distance[107,109,116], the rate of transcriptomic divergence is cell-type specific: in the primate cerebral cortex, for example, non-neuronal cells diverged more rapidly than neurons[109]. Second, transcriptomic divergence is not even across gene families. TFs known for specifying cell

identity have conserved expression in homologous neuron types, whereas the expression of terminal markers or “effector genes” may switch more rapidly[109]. This indicates that homologous neurons may acquire species-specific functions, such as new electrophysiological properties[116], without changes to their core genetic identity.

These observations are in line with an evolutionary definition of cell type, whereby homologous cells share the expression of TFs that establish and maintain their genetic identity[118]. Comparing the expression of TFs can help disambiguate homology and convergent evolution in cell type comparisons of distantly-related vertebrate species[106,110,112,119-121]. For example, distinct classes of cortical GABAergic interneurons in amphibians, reptiles, birds, and mammals express the same TFs defining class identity; however, the expression of certain effector genes such as the calcium-binding protein parvalbumin, which marks a class of mammalian GABAergic interneurons, is not conserved across species[106,110,112,119].

Are all genes equally informative?

Whether and how to give different weights to genes for homology inference remains an open question both conceptually and algorithmically. While TFs seem to carry a higher weight for homology inference, TF combinatorial codes may themselves drift, for example by paralog switching[122]. Paralog switching is particularly relevant for comparing distantly-related species. Standard approaches use one-to-one orthologs, because of the assumption that these genes carry the same functions. This assumption cannot be made for paralogs because gene duplication may be followed by sub- or neofunctionalization. However, limiting the analysis to one-to-one orthologs filters out a considerable fraction of the transcriptome, when the species compared are separated by large phylogenetic distances. Computational solutions to solve this problem have been proposed recently[114,123].

Finally, limiting cross-species comparisons to TFs comes with the risk of providing an oversimplified representation of cellular diversity. As the field defines subtler distinctions of subtypes within given classes, TF level gradients and or post-translational modifications may be identified in eliciting distinct transcriptional programs, making cross-species comparisons more complicated. Moreover, the potential developmental regulation of TF expression makes their use to define identity challenging from a temporal perspective, even within the same species. For example, the comparison of TFs may not be powered to identify cell types that have diverged recently (“sister cell types”[118]), which by virtue of their recent diversification share a significant fraction of their transcriptomes.

Are cellular transcriptomes enough to infer neuronal homologies?

As described above, homology inference becomes harder with increasing phylogenetic distance, especially when there are large branch lengths between clades with no extant species, e.g., comparing mammals to reptiles. Natural selection “cares” about the output of brain activity, i.e., the ability of the brain to support adaptive behaviors in the environment of an organism. The substrate of selection is the frequency of allelic variants in the population. However, the mapping between genotype and the phenotypes under selection in the brain is non-trivial: genes do not control behavior directly (with a few exceptions), rather, they affect

behavior by instructing cell type identity, neuronal wiring, activity, spatial allocation, etc. Transcriptomes alone might be insufficient to infer homology when we do not know the traits under selection or how genes are related to those traits. Other comparisons that can help with homology inference are defining the developmental origin and neuronal connectivity of a given cell, although these criteria have their own caveats. The concordance of developmental origin, transcriptomic similarity, and input-output connectivity is ideal for a solid homology inference.

In vitro models

The inaccessibility of neural tissues in humans and other species can limit experimental approaches, making *in vitro* models particularly valuable tools for neuroscientists[124,125], especially for evolutionary neurobiology. Neural organoids[125], 3D *in vitro* cell cultures, are experimentally tractable systems that can model the cell-type heterogeneity[125-131] and spatial organization[132,133] of *in vivo* neural tissues. A limitation of organoids is that do not faithfully reproduce in vivo spatial organization, only rudimentary cell type heterogeneity and developmental trajectories. However, for cross-species comparisons, neural organoids offer investigative avenues not typically available in non-model organisms, representing a potential medium for expansive comparative analyses. As a striking example, human neural organoids have been generated with or without a single amino acid change found in Neanderthals, enabling the study of the neurobiological consequences of genetic variation found in an extinct species[134,135]. Single-cell dissection of neural organoids can provide cell-type resolution of developmental trajectories[136,137], enable perturbation of dynamic gene regulatory networks[138], model neuro-disease mechanisms[139-141] and support neurodevelopmental cross-species comparisons[130,136,142]. However, generalizing observations in organoids to natural biology requires evaluation of the resemblance of *in vitro* derived cell types, and their composition and organization to *in vivo* development.

Single cell and spatial genomics approaches enable assessing these metrics of organoid fidelity but require accounting for technical variability and limitations of measurements in single cells. For example, single cell comparisons demonstrate the capacity of neural organoids to model broad *in vivo* neural cell-types across numerous genomic modalities[126-128,137,143]. However, these single-cell data comparisons typically use a singular *in vivo* dataset as a reference, which ignores potential variability within the reference data that organoids may not recapitulate. Especially with sparse and noisy single-cell data, any individual dataset carries error in the biological signal. A neural organoid model that recapitulates signal from a single *in vivo* dataset may in fact be a poor general model if the reference signal is of low quality and fails to replicate. Therefore, to avoid overfitting it is useful to incorporate cross-validation of *in vivo* signal among *in vivo* datasets[144]. This approach establishes a measure of expected error for reference signal, which provides a data-driven threshold for identifying neural organoid datasets that produce signal comparable to *in vivo* data. *In vivo* cross-validation can consist of quantifying signal replicability across the increasing amount of publicly available single-cell data of developing neural tissues or, within a single dataset, quantifying signal replicability from a withheld portion of the data. For example, quantifying differential gene expression statistics across cell-types and collating the p-values and fold-changes of genes derived from individual *in vivo* datasets

establishes a benchmark of reference signal for interpreting organoid differential expression statistics.

While data generation and benchmarking of *in vivo* data and neural organoids is an ongoing endeavor, much can concurrently be learned using organoid tissue. Environmental[145] and genetic perturbations[138,139] of neural organoids coupled with single-cell technologies provide informative tests for the mechanisms underlying how a genome operates in a neural tissue setting. For cross-species comparisons, single-cell dissection of neural organoids can resolve key developmental differences across species, such as molecular mechanisms underlying neural progenitor variation across human and primate organoids[130,136]. However, observations are only robust to the class of variability sampled and assessments should be applied to diverse genetic backgrounds (cell lines) or differentiation protocols to identify signals that are not specific to an individual cell line or protocol. As examples, different organoid protocols aiming to derive similar neural lineages (cortical organoids) have reported biases in differentiation patterns[143] and cell-line specific effects[146,147] can obscure disease variability in organoid models. Sampling over increased genetic and/or technical variability increases the likelihood of replicable signal and acts as a buffer against overfitting. Future studies employing new spatial transcriptomic and cellular barcoding strategies (see companion pieces) to examine organoid fidelity to normal development can harness these principles and evaluate additional properties beyond cellular transcriptomes, including spatial organization, developmental lineage relationships, and connectivity.

Technical limitations of transcriptomic-based homology inference

While we have attempted to provide conceptual criteria above, there are computational challenges with defining cell type homologies because there are no formal or uniformly accepted criteria. Multiple methods predict homologous cell types, including utilizing shared nearest neighbors and axes of transcriptional variation[114,148-153]. However, integrating across species can be difficult because identification of homologous cell types often relies on heuristics such as shared nearest neighbors and non-linear data transformations, rather than formal models of gene expression divergence and cell type evolution[154]. As such, the inclusion or exclusion of cell types within a given dataset can alter which cell types appear homologous. For example, a putative primate-specific cell type thought to be most similar to other striatal interneurons[108] was determined to actually be more similar to diencephalic neurons when such cell types were further included in the analysis[155]. This issue is a potential caveat for any type of comparison whether it is between species, regions, or developmental time periods.

Compositional concerns are especially pressing in the context of in vitro studies, in which different iPSC lines respond divergently to patterning factors and generate cultures with variable compositions. It is also important to consider that conserved populations can be repurposed to different brain structures over development. For example, recent work showed that classes of inhibitory neurons that migrate to rodent olfactory bulb have been redirected to the expanded primate white matter[105], and a mammalian conserved interneuron type is most numerous in the mouse hippocampus but more abundant in the primate neocortex[108].

Thus, the challenges in using single cell approaches to study cell types across species are multiplicative. Even with reliable in vivo data from multiple sources, spatiotemporal context and

biological variation must be considered when modeling homology. In vitro studies have the same challenges amplified: cell type distributions are untethered to the anatomy that is reproducibly generated in vivo, with the added concern that the observed cell states approximate those seen in vivo, heavily layered with various sources of technical variation. Despite these challenges, existing data and tools wielded with perspicacious judgement have enabled the discovery of new cell types and shared features and principles of vertebrate brain development and function.

Technical and biological artifacts

Since evolutionary findings may be challenging to experimentally validate, it is important to consider experimental factors that could lead to erroneous interpretations. Some of these are pertinent to evolutionary comparisons, but most are generalizable to other types of comparisons (and can be mitigated by careful orthogonal validation – see above). Recent studies have shown that technical artifacts such as doublets and ambient RNA contamination can lead to misinterpretations[156]. This is exacerbated when datasets are compared without properly adjusting for single-cell sequencing artifacts. For example, if datasets for one species contained more artifacts (e.g., higher doublet rate, greater ambient RNA contamination) than the other species, the result could be misinterpreted as a species-specific effect (see **Figure 4**). Biological information can also lead to misinterpretations. It is crucial to obtain demographically and spatiotemporally similar brain tissues from all species for a proper evolutionary comparison. For example, if regional boundaries are not rigorously considered during dissection, it is possible to compare improperly matched brain regions between two species which can also lead to misclassification of region-specific cellular and molecular features as species-specific effect. We note that spatial transcriptomics may alleviate this problem for many species with small brain sizes. In addition to brain regions, developmental time points should be matched across species to prevent misinterpreting age-specific effects as species-specific effects (see **Figure 4**). However, matching developmental timepoints in distantly related species might be impossible and heterochrony should also be considered as a mechanism for evolutionary change. Finally, it is important to consider that age-matching often depends on an estimate based on life history traits and some cell types may be more sensitive to age-effect than others (e.g., glia change more than neurons in very old age[157]). Thus, interpretation of any species-specific result should consider the age bracket of the samples.

Conclusions

In this review, we have discussed the applications and challenges of sc/snRNAseq for studying cellular heterogeneity and evolution across different biological systems, disease states, and cross-species comparisons. We have highlighted the importance of validating sc/snRNAseq data through various steps, such as in-depth data analysis, functional characterization, cross-validation, multi-omics integration, and follow-up validation experiments. Furthermore, sharing validated scRNAseq data with the scientific community is essential for fostering collaboration and scientific progress. These steps can help researchers to maximize the utility and impact of sc/snRNAseq findings, uncover novel insights into cellular dynamics and function, and identify potential therapeutic targets for complex biological processes or disease contexts. We also emphasize the need for specific practices to handle confounds in cross-system analyses, such

as sampling broadly within each system, measuring variance, assessing similarity without merging, and reporting robustness with effect sizes (see also companion pieces). These practices can help researchers to avoid overfitting and bias, and to provide meaningful cross-system assessments that can reveal the molecular mechanisms of brain evolution and adaptive behavioral phenotypes. By applying sc/snRNAseq approaches with careful consideration of the inherent challenges and limitations, researchers can advance our understanding of cellular heterogeneity and evolution across different biological systems.

BOXES:

Text Box 1. Questions that can be addressed using single cell/nuclei sequencing.

- Single-cell RNA-seq data indicates gene X is expressed by cell type A.
- Single-cell RNA data indicates gene Y is upregulated in cell type B during disease/pathology/etc.
- Single-cell data indicates cell type C is composed of three sub-states characterized by expression of gene X, gene Y, and gene Z, respectively.
- Compositional analysis indicates cell type/state D increases/decreases in abundance in disease/pathology/etc.
- Trajectory inference/RNA velocity/etc suggests that gene X is [upregulated/downregulated] as cells [differentiate/respond to insult or pathology/etc].
- Gene regulatory network inference, peak-gene linkage analysis, etc. suggests transcription factor TF1 (or enhancer/repressor E1) modulates expression of gene A.
- Cell-cell communication analysis suggests cell type A modulates cell type B through the interactions of ligand L1 and receptor R1.

Text Box 2. Current challenges and opportunities in sc/snRNAseq and spatial transcriptomic analysis

- **Cell lineage tracing**: the current technologies that exist for studying RNA velocity are limited because the directionality that they output can be perturbed by overlaying it on already-existing two-dimensional spaces. Moreover, RNA velocity packages were not validated in in vitro systems, which is a limitation when performing scRNAseq from in vitro cultures. Also, pseudotime and the pros/cons of these types of "lineage" approaches are well-known and have been covered elsewhere in great depth[87,158]. For further reading, see also (Cite accompanying piece).
- **Cell stage prediction**: there are ways of predicting the cell cycle stage by cross-referencing the genes expressed by the cell to a database of genes annotated for each stage, though the reliability of this has not been studied in depth. For further reading, see also (Cite accompanying piece).
- **Cell-cell communication** tools exist, though they are limited, again, as most rely on databases on known protein-protein interactions, and, furthermore, there is a general assumption (as in all of scRNAseq) that mRNA levels correlate with protein levels. For further reading, see also (Cite accompanying piece).
- **Data integration**. See also (Cite accompanying piece) Appropriate and reliable tools that exist for this in both humans and mice. Cross-species comparison analyses are definitely possible to do, and should be attempted, though with the exception that there are genes that do not have a homolog or ortholog in the other species, and this reduces the scope of the comparison. New tools are being developed with great speed, and these concerns will be likely overcome soon. One attempt employed by many is to worry less about exact matching of individual genes, but to instead focus on anchoring gene expression modules within individual clusters – mitigating the lack of orthologs in some instances. Methodologies to perform data integration, across datasets, species, modalities, and more, are invaluable tools in the analysis of single cell and spatial

genomics. However, these technologies present risk as well, in that they can potentially eliminate or mitigate real biological differences and therefore artificially smooth analyses. Importantly, and often overlooked, is that once cell types are aligned between organisms (e.g. rodent and primate) using scRNAseq, they still need to be functionally validated across organisms, as the circuitry and functional properties are not necessarily fully conserved. This becomes challenging as cell type specific tools are lacking outside rodents, but new RNA editing technologies could hold promise for use in primates, while the use of cell-based assays and organoids provides some validation options for human cells/functions.

Text Box 3. Why do we care about evolution?

- Evolution is a general biological principle. Thus, understanding the contribution of evolution to nervous system function provides important foundational basic science knowledge. In addition, understanding the evolutionary constraints and opportunities that have occurred in many organisms informs our understanding of the relevance of these changes in humans.
- Understanding the similarity or differences between cell types helps us better interpret our findings in one organism to another (e.g., from other mammals to humans).
- Convergent evolution informs us about the constraints that shape brain evolution in terms of plasticity and functional organization of the tissue. In this manner, we can focus on the potential cellular and molecular mechanisms that correlate with convergent behaviors (e.g., direct corticospinal connections onto lower motor neurons and fine motor control).
- The implementation of evolutionary approaches can result in adaption of new model systems that may offer some technical advantages for studying a general problem (e.g., the evolution of sleep[159]).
- Evolved nervous system function may be directly linked to the emergence of many types of nervous system disorders in humans that are not observable in other species.

FIGURES:

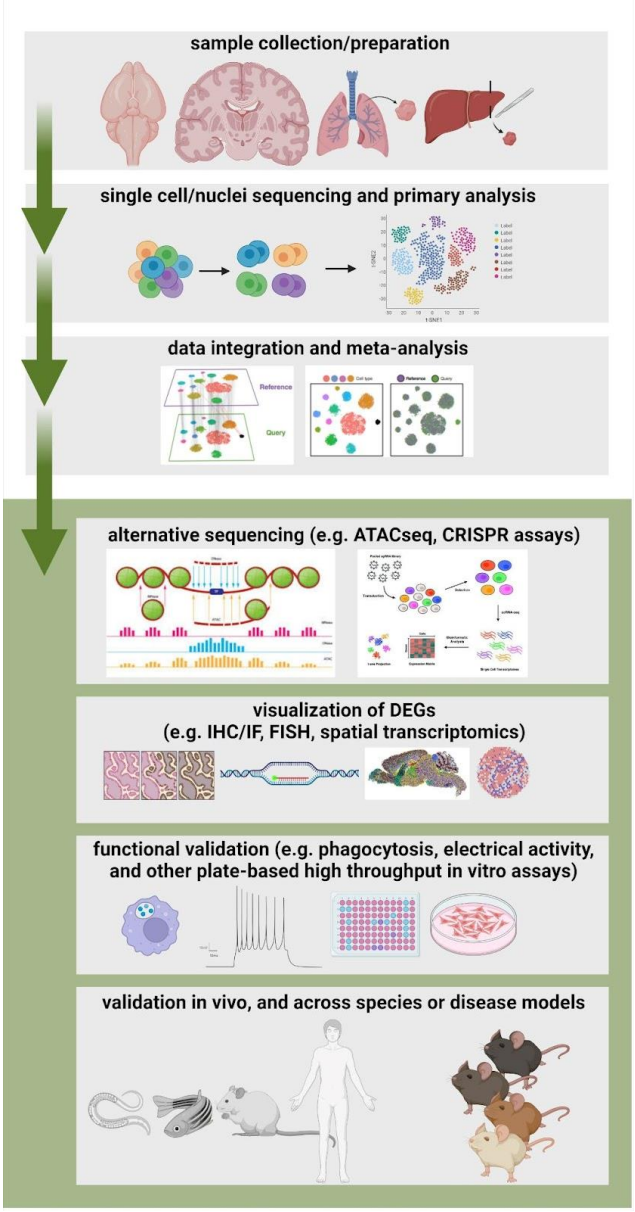

Figure 1. Considerations for orthogonal and functional validation of sequencing data. Single cell/nuclei RNA sequencing data, regardless of the tissue originally collected form, requires multiple validation steps to ensure biological validity. In addition to ensuring proper powering of cell-types-of-interest, additional steps should be applied for best practices. This includes integration with other datasets (either across disease models, species, or labs), alternative sequencing methods (e.g. to access chromatin accessibility, or to perturb individual transcripts), visualization (using immunohistochemistry, FISH, or spatial transcriptomics), functional validation to ensure sub-types/-states of cells are indeed terminal and no transitory, and finally cross-species validation (of particular importance when using animal models of disease to ensure relevance to human patients). One or all of these methods, among others, may be required to validate a number of differentially expressed genes (DEGs) identified in initial single cell/nuclei sequencing experiments.

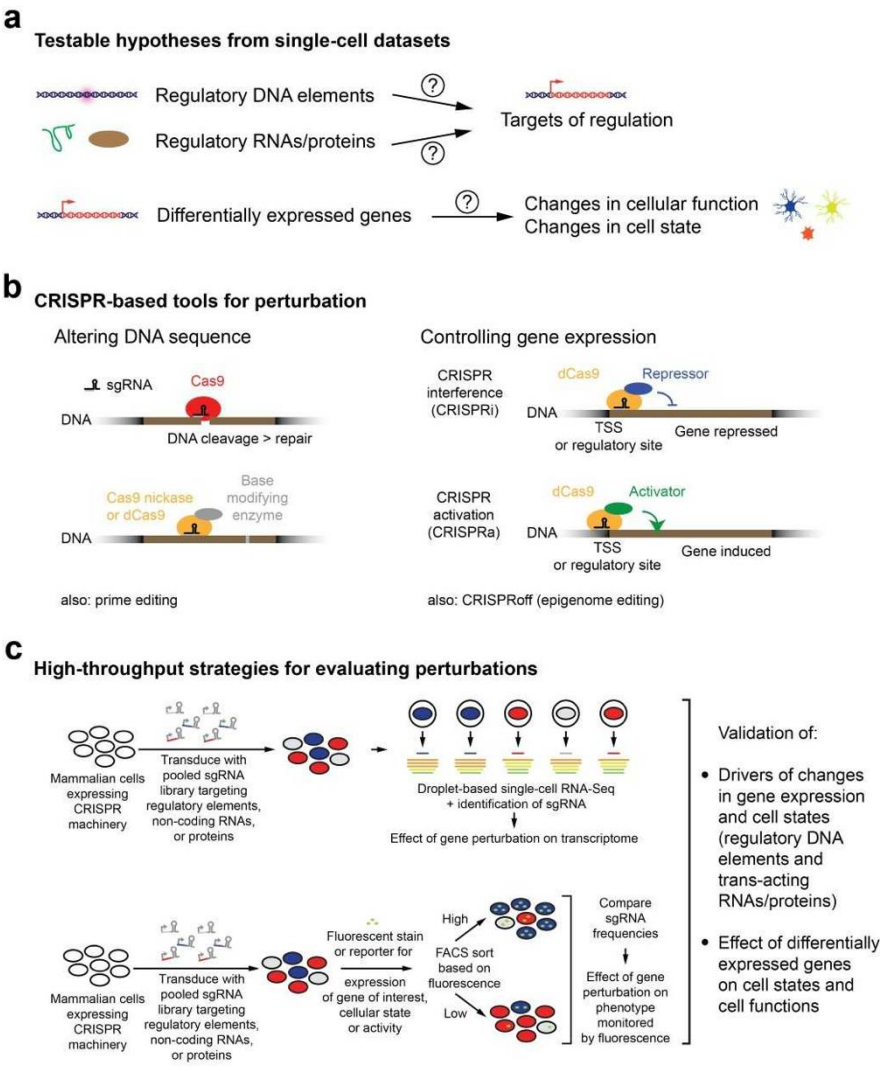

Figure 2. Overview of perturbation-based validation approaches. a, sc/snRNA-seq datasets can generate different types of functional or mechanistic hypotheses. **b**, Examples for CRISPR-based tools to perturb genome sequence and gene expression. **c**, Experimental strategies for high-throughput CRISPR-based perturbation experiments to validate and test functional or mechanistic hypotheses from sc/snRNA-seq datasets.

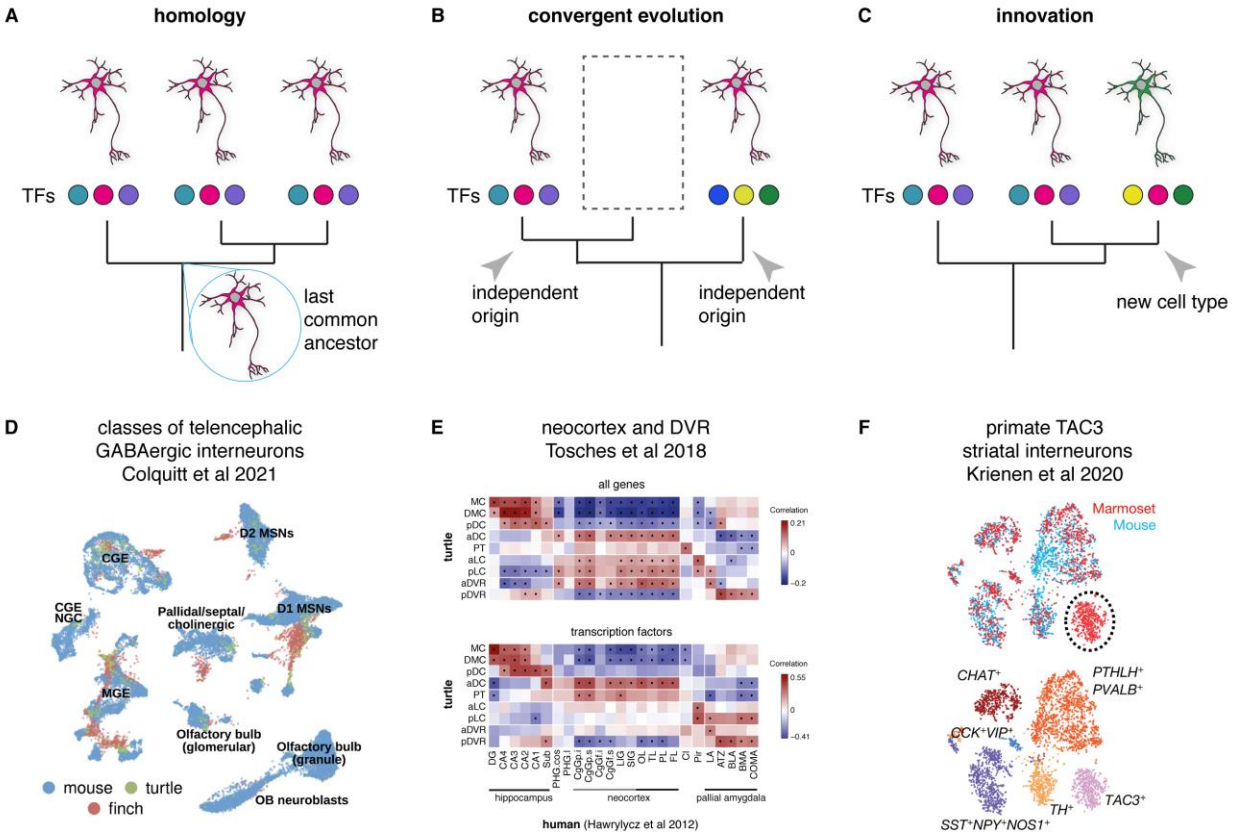

Figure 3. Illustration of cell type homology, convergence, and innovation. A-C) Schematics of cell type evolution. Circles indicate transcription factors (TFs). D-F) Examples from the literature[108,110,119] using single cell genomics to address each type of cell type evolution.

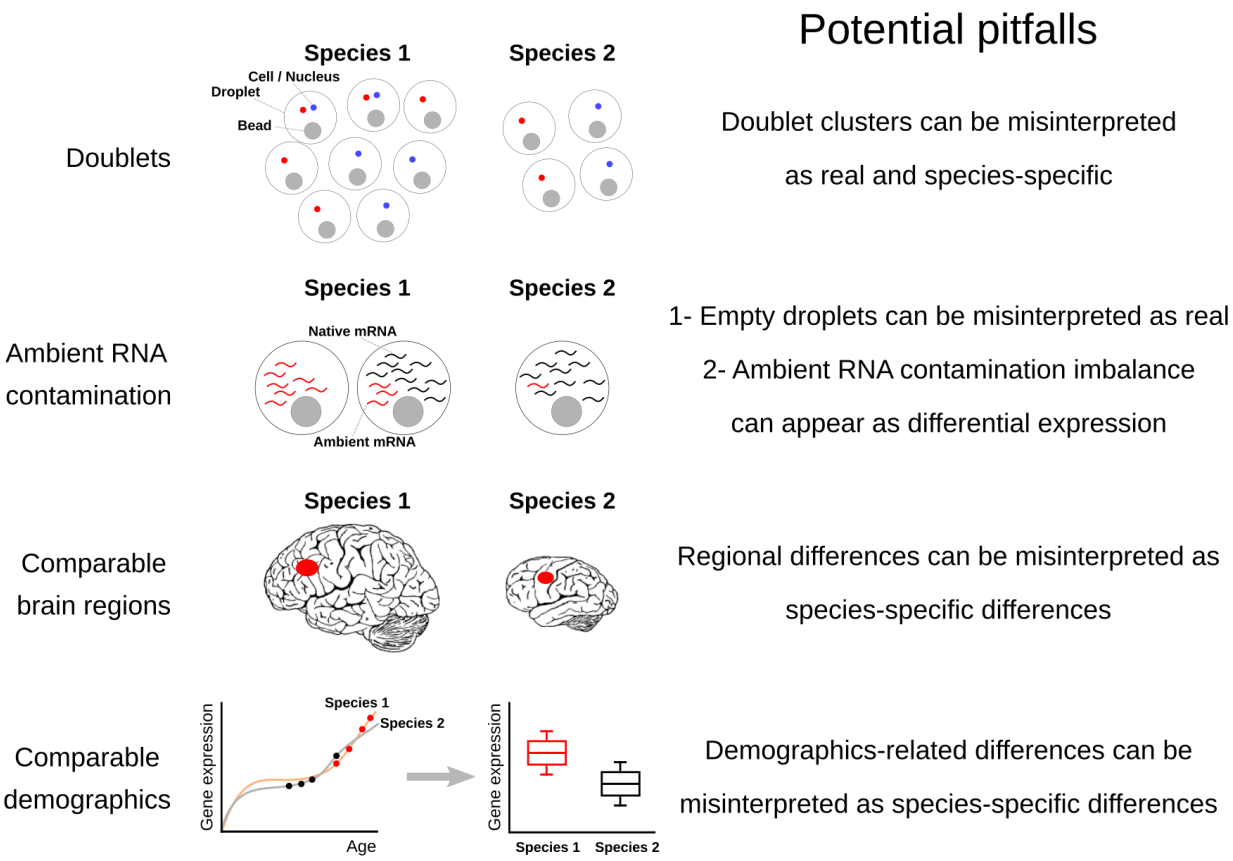

Figure 4. **Illustration of technical and biological artifacts.** Schematics of how evolutionary comparisons using single cell genomics could be vulnerable to misinterpretations due to either biological (e.g., brain region selection/dissection or demographics such as age) and/or technical (e.g., doublet or ambient RNA) artifacts.

TABLES:

Table 1. Overview of common orthogonal validation methods.

Orthogonal validation approach	Readout	What is being validated	Pros	Cons
RNAscope	Molecular	RNA transcripts	Single cell resolution spatial validation, relatively inexpensive	Low throughput
MERFISH (Vizgen)	Molecular	RNA transcripts	Single cell resolution spatial validation	Costly, requires specialized equipment and reagents
Visium (10X)	Molecular	RNA transcripts	High throughput and anatomical validation	Not single cell resolution; requires specialized reagents
Immunohistochemistry	Molecular	Protein	Easily accessible with no specialized reagents required	Need validated antibodies; low throughput
Flow cytometry	Molecular	Protein	Quantitative readouts at protein level	Validating translation potential, which may be discordant from RNA findings; requires validated antibodies
CyTOF	Molecular	Protein	Quantification of multiple cellular components simultaneously (high throughput)	Validating translation potential, which may be discordant from RNA findings
CRISPR knockout	Functional	Gene function	Test necessity of candidate genes; relatively standardized workflows across model systems	Low throughput; can be costly to test multiple genes
CRISPRi/a	Functional	Gene function	Manipulate expression of endogenous genes and monitor phenotypic consequences; can be multiplexed or performed in pooled screens	Variability in degree of interference or activation from gene-to-gene. Susceptible to epigenetic or trans-acting regulatory environment
Perturb-seq	Functional	Gene function	Massively parallel functional readouts of gene perturbation phenotypes by single-cell transcriptomics and individual cell resolution; can be used with traditional Cas9 or CRISPRa/i	Not trivial to design, execute, and interpret; costly; require robust selective challenge
CROP-seq	Functional	Gene function	Guide RNAs read directly; simplified workflow for large screens	Not trivial to design, execute, and interpret; costly; require robust selective challenge
ECCITE-seq	Functional	Gene function	An extension of Perturb-seq/CROP-seq to multimodal readouts	Challenging to implement for intracellular antigens

RABID-seq	Functional connections	Cell-cell interactions	High throughput approach to validate physical cell-cell interactions	Requires specialized reagents and bioinformatic pipelines
Circuit tracing	Functional connections	Cell-cell interactions	Can be used to identify short and long range neuronal connections	May be difficult to label deep brain regions
SPEAC-seq	Functional connections	Cell-cell interactions	Allows individual gene perturbations in cells cultured in individual droplets	Requires specialized reagents and bioinformatic pipelines
Physiological readouts (calcium imaging, electrophysiology, transporter activity)	Functional	Physical properties of cells	Can match biophysical properties of cells to their transcriptional identities; powerful tools available	Requires specialized skillsets (electrophysiology); may require either live intact tissue sections, cell type-specific genetic labeling, or robust purification strategies to target cell types of interest
Live imaging (migration, proliferation)	Functional	Physical properties of cells	Can be performed in high throughput (multiple cells per image); provides input on cellular behavior	Requires specialized microscopes (2-photon, light-sheet) and live cell labeling tools
Dye-filing for morphological readouts	Morphological	Cell morphology	Can provide morphological information that is far more detailed than immunohistochemistry	Low throughput and requires specialized equipment
Viral targeting	Morphological	Cell morphology	High fidelity morphological information, can provide sparse labeling for ease of reconstruction	May be difficult to label deep brain regions
fluorescent protein expression (driver line)	Morphological	Cell morphology	Can label all cells of one type/subtype across the entire CNS	Depending on driver, labeled cell density could be too high to identify individual complex cells

Abbreviations: CRISPR, clustered regularly interspaced short palindromic repeats; CRISPRa, CRISPR activation; CRISPRi, CRISPR interference; CROP-seq, CRISPR droplet sequencing; CyTOF, Cytometry by time of flight; ECCITE-seq, expanded CRISPR-compatible cellular indexing of transcriptomes and epitopes by sequencing; MERFISH, Multiplexed Error-Robust Fluorescence in situ Hybridization; RABID-seq, rabies barcode interaction detection followed by sequencing; SPEAC-seq, systematic perturbation of encapsulated associated cells followed by sequencing.

REFERENCES:

1 Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol Syst Biol* **15**, e8746 (2019). https://doi.org:10.15252/msb.20188746

2 Aldridge, S. & Teichmann, S. A. Single cell transcriptomics comes of age. *Nat Commun* **11**, 4307 (2020). https://doi.org:10.1038/s41467-020-18158-5

3 Trapnell, C. Defining cell types and states with single-cell genomics. *Genome Res* **25**, 1491-1498 (2015). https://doi.org:10.1101/gr.190595.115

4 Hao, Y. *et al.* Dictionary learning for integrative, multimodal, and scalable single-cell analysis. *bioRxiv*, 2022.2002.2024.481684 (2022). https://doi.org:10.1101/2022.02.24.481684

5 Kiselev, V. Y. *et al.* SC3: consensus clustering of single-cell RNA-seq data. *Nat Methods* **14**, 483-486 (2017). https://doi.org:10.1038/nmeth.4236

6 Nayak, R. & Hasija, Y. A hitchhiker's guide to single-cell transcriptomics and data analysis pipelines. *Genomics* **113**, 606-619 (2021). https://doi.org:10.1016/j.ygeno.2021.01.007

7 Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* **33**, 495-502 (2015). https://doi.org:10.1038/nbt.3192

8 Dominguez Conde, C. *et al.* Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science* **376**, eabl5197 (2022). https://doi.org:10.1126/science.abl5197

9 Xu, C. *et al.* Automatic cell type harmonization and integration across Human Cell Atlas datasets. *bioRxiv*, 2023.2005.2001.538994 (2023). https://doi.org:10.1101/2023.05.01.538994

10 Menon, V. Clustering single cells: a review of approaches on high-and low-depth single-cell RNA-seq data. *Brief Funct Genomics* **17**, 240-245 (2018). https://doi.org:10.1093/bfgp/elx044

11 Sokol, L. *et al.* Prioritization and functional validation of target genes from single-cell transcriptomics studies. *Commun Biol* **6**, 648 (2023). https://doi.org:10.1038/s42003-023-05006-7

12 Lahnemann, D. *et al.* Eleven grand challenges in single-cell data science. *Genome Biol* **21**, 31 (2020). https://doi.org:10.1186/s13059-020-1926-6

13 Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587 e3529 (2021). https://doi.org:10.1016/j.cell.2021.04.048

14 Ding, J. *et al.* Systematic comparative analysis of single cell RNA-sequencing methods. 632216 (2019). https://doi.org:10.1101/632216 %J bioRxiv

15 Zeng, H. & Sanes, J. R. Neuronal cell-type classification: challenges, opportunities and the path forward. *Nat Rev Neurosci* **18**, 530-546 (2017). https://doi.org:10.1038/nrn.2017.85

16 Scala, F. *et al.* Phenotypic variation of transcriptomic cell types in mouse motor cortex. *Nature* **598**, 144-150 (2021). https://doi.org:10.1038/s41586-020-2907-3

17 Clark, I. C. *et al.* Identification of astrocyte regulators by nucleic acid cytometry. *Nature* **614**, 326-333 (2023). https://doi.org:10.1038/s41586-022-05613-0

18 Rodriques, S. G. *et al.* Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science* **363**, 1463-1467 (2019). https://doi.org:10.1126/science.aaw1219

19 Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348**, aaa6090 (2015). https://doi.org:10.1126/science.aaa6090

20 Wang, X. *et al.* Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* **361** (2018). https://doi.org:10.1126/science.aat5691

21 Ke, R. *et al.* In situ sequencing for RNA analysis in preserved tissue and cells. *Nat Methods* **10**, 857-860 (2013). https://doi.org:10.1038/nmeth.2563

22 Liu, Y., Beyer, A. & Aebersold, R. On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell* **165**, 535-550 (2016). https://doi.org:10.1016/j.cell.2016.03.014

23 Wegler, C. *et al.* Global variability analysis of mRNA and protein concentrations across and within human tissues. *NAR Genom Bioinform* **2**, lqz010 (2020). https://doi.org:10.1093/nargab/lqz010

24 Buccitelli, C. & Selbach, M. mRNAs, proteins and the emerging principles of gene expression control. *Nat Rev Genet* **21**, 630-644 (2020). https://doi.org:10.1038/s41576-020-0258-4

25 Almet, A. A., Cang, Z., Jin, S. & Nie, Q. The landscape of cell-cell communication through single-cell transcriptomics. *Curr Opin Syst Biol* **26**, 12-23 (2021). https://doi.org:10.1016/j.coisb.2021.03.007

26 Dimitrov, D. *et al.* Comparison of methods and resources for cell-cell communication inference from single-cell RNA-Seq data. *Nat Commun* **13**, 3224 (2022). https://doi.org:10.1038/s41467-022-30755-0

27 Armingol, E., Officer, A., Harismendy, O. & Lewis, N. E. Deciphering cell-cell interactions and communication from gene expression. *Nat Rev Genet* **22**, 71-88 (2021). https://doi.org:10.1038/s41576-020-00292-x

28 Gault, J. *et al.* Combining native and 'omics' mass spectrometry to identify endogenous ligands bound to membrane proteins. *Nat Methods* **17**, 505-508 (2020). https://doi.org:10.1038/s41592-020-0821-0

29 Wang, J. Y. & Doudna, J. A. CRISPR technology: A decade of genome editing is only the beginning. *Science* **379**, eadd8643 (2023). https://doi.org:10.1126/science.add8643

30 Anzalone, A. V., Koblan, L. W. & Liu, D. R. Genome editing with CRISPR-Cas nucleases, base editors, transposases and prime editors. *Nat Biotechnol* **38**, 824-844 (2020). https://doi.org:10.1038/s41587-020-0561-9

31 Gilbert, L. A. *et al.* Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *Cell* **159**, 647-661 (2014). https://doi.org:10.1016/j.cell.2014.09.029

32 Kearns, N. A. *et al.* Functional annotation of native enhancers with a Cas9-histone demethylase fusion. *Nat Methods* **12**, 401-403 (2015). https://doi.org:10.1038/nmeth.3325

33 Thakore, P. I. *et al.* Highly specific epigenome editing by CRISPR-Cas9 repressors for silencing of distal regulatory elements. *Nat Methods* **12**, 1143-1149 (2015). https://doi.org:10.1038/nmeth.3630

34 Nunez, J. K. *et al.* Genome-wide programmable transcriptional memory by CRISPR-based epigenome editing. *Cell* **184**, 2503-2519 e2517 (2021). https://doi.org:10.1016/j.cell.2021.03.025

35 Konermann, S. *et al.* Transcriptome Engineering with RNA-Targeting Type VI-D CRISPR Effectors. *Cell* **173**, 665-676 e614 (2018). https://doi.org:10.1016/j.cell.2018.02.033

36 Adamson, B. *et al.* A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell* **167**, 1867-1882 e1821 (2016). https://doi.org:10.1016/j.cell.2016.11.048

37 Dixit, A. *et al.* Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* **167**, 1853-1866 e1817 (2016). https://doi.org:10.1016/j.cell.2016.11.038

38 Datlinger, P. *et al.* Pooled CRISPR screening with single-cell transcriptome readout. *Nat Methods* **14**, 297-301 (2017). https://doi.org:10.1038/nmeth.4177

39 Wheeler, M. A. *et al.* Droplet-based forward genetic screening of astrocyte-microglia cross-talk. *Science* **379**, 1023-1030 (2023). https://doi.org:10.1126/science.abq4822

40 Yang, X. *et al.* Functional characterization of Alzheimer's disease genetic variants in microglia. *Nat Genet* **55**, 1735-1744 (2023). https://doi.org:10.1038/s41588-023-01506-8
41 Wu, D. *et al.* Dual genome-wide coding and lncRNA screens in neural induction of induced pluripotent stem cells. *Cell Genom* **2** (2022). https://doi.org:10.1016/j.xgen.2022.100177
42 Tian, R. *et al.* CRISPR Interference-Based Platform for Multimodal Genetic Screens in Human iPSC-Derived Neurons. *Neuron* **104**, 239-255 e212 (2019). https://doi.org:10.1016/j.neuron.2019.07.014
43 Drager, N. M. *et al.* A CRISPRi/a platform in human iPSC-derived microglia uncovers regulators of disease states. *Nat Neurosci* **25**, 1149-1162 (2022). https://doi.org:10.1038/s41593-022-01131-4
44 Leng, K. *et al.* CRISPRi screens in human iPSC-derived astrocytes elucidate regulators of distinct inflammatory reactive states. *Nat Neurosci* **25**, 1528-1542 (2022). https://doi.org:10.1038/s41593-022-01180-9
45 Esk, C. *et al.* A human tissue screen identifies a regulator of ER secretion as a brain-size determinant. *Science* **370**, 935-941 (2020). https://doi.org:10.1126/science.abb5390
46 Jin, X. *et al.* In vivo Perturb-Seq reveals neuronal and glial abnormalities associated with autism risk genes. *Science* **370** (2020). https://doi.org:10.1126/science.aaz6063
47 Wertz, M. H. *et al.* Genome-wide In Vivo CNS Screening Identifies Genes that Modify CNS Neuronal Survival and mHTT Toxicity. *Neuron* **106**, 76-89 e78 (2020). https://doi.org:10.1016/j.neuron.2020.01.004
48 Wheeler, M. A. *et al.* Environmental Control of Astrocyte Pathogenic Activities in CNS Inflammation. *Cell* **176**, 581-596 e518 (2019). https://doi.org:10.1016/j.cell.2018.12.012
49 Johnson, M. B. *et al.* Single-cell analysis reveals transcriptional heterogeneity of neural progenitors in human cortex. *Nat Neurosci* **18**, 637-646 (2015). https://doi.org:10.1038/nn.3980
50 Buettner, F. *et al.* Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat Biotechnol* **33**, 155-160 (2015). https://doi.org:10.1038/nbt.3102
51 Velten, L. *et al.* Human haematopoietic stem cell lineage commitment is a continuous process. *Nat Cell Biol* **19**, 271-281 (2017). https://doi.org:10.1038/ncb3493
52 Crow, M., Paul, A., Ballouz, S., Huang, Z. J. & Gillis, J. Characterizing the replicability of cell types defined by single cell RNA-sequencing data using MetaNeighbor. *Nat Commun* **9**, 884 (2018). https://doi.org:10.1038/s41467-018-03282-0
53 Cuomo, A. S. E. *et al.* Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression. *Nat Commun* **11**, 810 (2020). https://doi.org:10.1038/s41467-020-14457-z
54 Virdi, G. S. *et al.* Protein aggregation and calcium dysregulation are hallmarks of familial Parkinson's disease in midbrain dopaminergic neurons. *NPJ Parkinsons Dis* **8**, 162 (2022). https://doi.org:10.1038/s41531-022-00423-7
55 Szebenyi, K. *et al.* Human ALS/FTD brain organoid slice cultures display distinct early astrocyte and targetable neuronal pathology. *Nat Neurosci* **24**, 1542-1554 (2021). https://doi.org:10.1038/s41593-021-00923-4
56 Hall, C. E. *et al.* Progressive Motor Neuron Pathology and the Role of Astrocytes in a Human Stem Cell Model of VCP-Related ALS. *Cell Rep* **19**, 1739-1749 (2017). https://doi.org:10.1016/j.celrep.2017.05.024
57 de Majo, M. *et al.* Granulin loss of function in human mature brain organoids implicates astrocytes in TDP-43 pathology. *Stem Cell Reports* **18**, 706-719 (2023). https://doi.org:10.1016/j.stemcr.2023.01.012
58 Liddelow, S. A. *et al.* Neurotoxic reactive astrocytes are induced by activated microglia. *Nature* **541**, 481-487 (2017). https://doi.org:10.1038/nature21029

59 Barbar, L. *et al.* CD49f Is a Novel Marker of Functional and Reactive Human iPSC-Derived Astrocytes. *Neuron* **107**, 436-453 (2020). https://doi.org:10.1016/j.neuron.2020.05.014

60 Clark, I. C. *et al.* Barcoded viral tracing of single-cell interactions in central nervous system inflammation. *Science* **372** (2021). https://doi.org:10.1126/science.abf1230

61 Pasqual, G. *et al.* Monitoring T cell-dendritic cell interactions in vivo by intercellular enzymatic labelling. *Nature* **553**, 496-500 (2018). https://doi.org:10.1038/nature25442

62 Xu, S. *et al.* Behavioral state coding by molecularly defined paraventricular hypothalamic cell type ensembles. *Science* **370** (2020). https://doi.org:10.1126/science.abb2494

63 Condylis, C. *et al.* Dense functional and molecular readout of a circuit hub in sensory cortex. *Science* **375**, eabl5981 (2022). https://doi.org:10.1126/science.abl5981

64 Bugeon, S. *et al.* A transcriptomic axis predicts state modulation of cortical interneurons. *Nature* **607**, 330-338 (2022). https://doi.org:10.1038/s41586-022-04915-7

65 Network, B. I. C. C. A multimodal cell census and atlas of the mammalian primary motor cortex. *Nature* **598**, 86-102 (2021). https://doi.org:10.1038/s41586-021-03950-0

66 Gerfen, C. R., Paletzki, R. & Heintz, N. GENSAT BAC cre-recombinase driver lines to study the functional organization of cerebral cortical and basal ganglia circuits. *Neuron* **80**, 1368-1383 (2013). https://doi.org:10.1016/j.neuron.2013.10.016

67 Gong, S. *et al.* A gene expression atlas of the central nervous system based on bacterial artificial chromosomes. *Nature* **425**, 917-925 (2003). https://doi.org:10.1038/nature02033

68 Gertler, T. S., Chan, C. S. & Surmeier, D. J. Dichotomous anatomical properties of adult striatal medium spiny neurons. *J Neurosci* **28**, 10814-10824 (2008). https://doi.org:10.1523/JNEUROSCI.2660-08.2008

69 Fenno, L. E. *et al.* Targeting cells with single vectors using multiple-feature Boolean logic. *Nat Methods* **11**, 763-772 (2014). https://doi.org:10.1038/nmeth.2996

70 Kramer, D. J. *et al.* Generation of a DAT-P2A-Flpo mouse line for intersectional genetic targeting of dopamine neuron subpopulations. *Cell Rep* **35**, 109123 (2021). https://doi.org:10.1016/j.celrep.2021.109123

71 Poulin, J. F. *et al.* Mapping projections of molecularly defined dopamine neuron subtypes using intersectional genetic approaches. *Nat Neurosci* **21**, 1260-1271 (2018). https://doi.org:10.1038/s41593-018-0203-4

72 Ren, J. *et al.* Single-cell transcriptomes and whole-brain projections of serotonin neurons in the mouse dorsal and median raphe nuclei. *Elife* **8** (2019). https://doi.org:10.7554/eLife.49424

73 Okaty, B. W. *et al.* A single-cell transcriptomic and anatomic atlas of mouse dorsal raphe Pet1 neurons. *Elife* **9** (2020). https://doi.org:10.7554/eLife.55523

74 Cadwell, C. R. *et al.* Electrophysiological, transcriptomic and morphologic profiling of single neurons using Patch-seq. *Nat Biotechnol* **34**, 199-203 (2016). https://doi.org:10.1038/nbt.3445

75 Fuzik, J. *et al.* Integration of electrophysiological recordings with single-cell RNA-seq data identifies neuronal subtypes. *Nat Biotechnol* **34**, 175-183 (2016). https://doi.org:10.1038/nbt.3443

76 Sloan, S. A. *et al.* Human Astrocyte Maturation Captured in 3D Cerebral Cortical Spheroids Derived from Pluripotent Stem Cells. *Neuron* **95**, 779-790 e776 (2017). https://doi.org:10.1016/j.neuron.2017.07.035

77 Guttenplan, K. A. *et al.* Knockout of reactive astrocyte activating factors slows disease progression in an ALS mouse model. *Nat Commun* **11**, 3753 (2020). https://doi.org:10.1038/s41467-020-17514-9

78 Guttenplan, K. A. *et al.* Neurotoxic reactive astrocytes induce cell death via saturated lipids. *Nature* (2021). https://doi.org:10.1038/s41586-021-03960-y

79 Hasel, P., Rose, I. V. L., Sadick, J. S., Kim, R. D. & Liddelow, S. A. Neuroinflammatory astrocyte subtypes in the mouse brain. *Nat Neurosci* **24**, 1475-1487 (2021). https://doi.org:10.1038/s41593-021-00905-6

80 Banker, G. A. Trophic interactions between astroglial cells and hippocampal neurons in culture. *Science* **209**, 809-810 (1980). https://doi.org:10.1126/science.7403847

81 Meyer-Franke, A., Kaplan, M. R., Pfrieger, F. W. & Barres, B. A. Characterization of the signaling interactions that promote the survival and growth of developing retinal ganglion cells in culture. *Neuron* **15**, 805-819 (1995). https://doi.org:10.1016/0896-6273(95)90172-8

82 Allen, N. J. *et al.* Astrocyte glypicans 4 and 6 promote formation of excitatory synapses via GluA1 AMPA receptors. *Nature* **486**, 410-414 (2012). https://doi.org:10.1038/nature11059

83 Kucukdereli, H. *et al.* Control of excitatory CNS synaptogenesis by astrocyte-secreted proteins Hevin and SPARC. *Proc Natl Acad Sci U S A* **108**, E440-449 (2011). https://doi.org:10.1073/pnas.1104977108

84 Fuentes-Medel, Y. *et al.* Glia and muscle sculpt neuromuscular arbors by engulfing destabilized synaptic boutons and shed presynaptic debris. *PLoS Biol* **7**, e1000184 (2009). https://doi.org:10.1371/journal.pbio.1000184

85 Chung, W. S. *et al.* Astrocytes mediate synapse elimination through MEGF10 and MERTK pathways. *Nature* **504**, 394-400 (2013). https://doi.org:10.1038/nature12776

86 La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494-498 (2018). https://doi.org:10.1038/s41586-018-0414-6

87 Bergen, V., Soldatov, R. A., Kharchenko, P. V. & Theis, F. J. RNA velocity-current challenges and future perspectives. *Mol Syst Biol* **17**, e10282 (2021). https://doi.org:10.15252/msb.202110282

88 Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat Biotechnol* **38**, 1408-1414 (2020). https://doi.org:10.1038/s41587-020-0591-3

89 Frazel, P. W. *et al.* Longitudinal scRNA-seq analysis in mouse and human informs optimization of rapid mouse astrocyte differentiation protocols. *Nat Neurosci* **26**, 1726-1738 (2023). https://doi.org:10.1038/s41593-023-01424-2

90 Lange, M. *et al.* CellRank for directed single-cell fate mapping. *Nat Methods* **19**, 159-170 (2022). https://doi.org:10.1038/s41592-021-01346-6

91 Kamimoto, K. *et al.* Dissecting cell identity via network inference and in silico gene perturbation. *Nature* **614**, 742-751 (2023). https://doi.org:10.1038/s41586-022-05688-9

92 Patani, R., Hardingham, G. E. & Liddelow, S. A. Functional roles of reactive astrocytes in neuroinflammation and neurodegeneration. *Nat Rev Neurol* **19**, 395-409 (2023). https://doi.org:10.1038/s41582-023-00822-1

93 Jeon, H. *et al.* Statistical Power Analysis for Designing Bulk, Single-Cell, and Spatial Transcriptomics Experiments: Review, Tutorial, and Perspectives. *Biomolecules* **13** (2023). https://doi.org:10.3390/biom13020221

94 Svensson, V. *et al.* Power analysis of single-cell RNA-sequencing experiments. *Nat Methods* **14**, 381-387 (2017). https://doi.org:10.1038/nmeth.4220

95 Su, K., Wu, Z. & Wu, H. Simulation, power evaluation and sample size recommendation for single-cell RNA-seq. *Bioinformatics* **36**, 4860-4868 (2020). https://doi.org:10.1093/bioinformatics/btaa607

96 Zeng, H. What is a cell type and how to define it? *Cell* **185**, 2739-2755 (2022). https://doi.org:10.1016/j.cell.2022.06.031

97 Yuste, R. *et al.* A community-based transcriptomics classification and nomenclature of neocortical cell types. *Nat Neurosci* **23**, 1456-1468 (2020). https://doi.org:10.1038/s41593-020-0685-8

98 Hodge, R. D. *et al.* Conserved cell types with divergent features in human versus mouse cortex. *Nature* **573**, 61-68 (2019). https://doi.org:10.1038/s41586-019-1506-7
99 Arendt, D., Bertucci, P. Y., Achim, K. & Musser, J. M. Evolution of neuronal types and families. *Curr Opin Neurobiol* **56**, 144-152 (2019). https://doi.org:10.1016/j.conb.2019.01.022
100 Wagner, D. E. & Klein, A. M. Lineage tracing meets single-cell omics: opportunities and challenges. *Nat Rev Genet* **21**, 410-427 (2020). https://doi.org:10.1038/s41576-020-0223-2
101 Bandler, R. C. *et al.* Single-cell delineation of lineage and genetic identity in the mouse brain. *Nature* **601**, 404-409 (2022).
102 Cadwell, C. R. *et al.* Cell type composition and circuit organization of clonally related excitatory neurons in the juvenile mouse neocortex. *Elife* **9** (2020). https://doi.org:10.7554/eLife.52951
103 Mayer, C. *et al.* Developmental diversification of cortical inhibitory interneurons. *Nature* **555**, 457-462 (2018). https://doi.org:10.1038/nature25999
104 Nowakowski, T. J. *et al.* Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. *Science* **358**, 1318-1323 (2017). https://doi.org:10.1126/science.aap8809
105 Schmitz, M. T. *et al.* The development and evolution of inhibitory neurons in primate cerebrum. *Nature* **603**, 871-877 (2022). https://doi.org:10.1038/s41586-022-04510-w
106 Woych, J. *et al.* Cell-type profiling in salamanders identifies innovations in vertebrate forebrain evolution. *Science* **377**, eabp9186 (2022). https://doi.org:10.1126/science.abp9186
107 Hahn, J. *et al.* Evolution of neuronal cell classes and types in the vertebrate retina. *bioRxiv* (2023). https://doi.org:10.1101/2023.04.07.536039
108 Krienen, F. M. *et al.* Innovations present in the primate interneuron repertoire. *Nature* **586**, 262-269 (2020). https://doi.org:10.1038/s41586-020-2781-z
109 Jorstad, N. L. *et al.* Comparative transcriptomics reveals human-specific cortical features. *bioRxiv*, 2022.2009.2019.508480 (2022). https://doi.org:10.1101/2022.09.19.508480
110 Tosches, M. A. *et al.* Evolution of pallium, hippocampus, and cortical cell types revealed by single-cell transcriptomics in reptiles. *Science* **360**, 881-888 (2018). https://doi.org:10.1126/science.aar4237
111 Peng, Y. R. *et al.* Molecular Classification and Comparative Taxonomics of Foveal and Peripheral Cells in Primate Retina. *Cell* **176**, 1222-1237 e1222 (2019). https://doi.org:10.1016/j.cell.2019.01.004
112 Hain, D. *et al.* Molecular diversity and evolution of neuron types in the amniote brain. *Science* **377**, eabp8202 (2022). https://doi.org:10.1126/science.abp8202
113 Lust, K. *et al.* Single-cell analyses of axolotl telencephalon organization, neurogenesis, and regeneration. *Science* **377**, eabp9262 (2022). https://doi.org:10.1126/science.abp9262
114 Tarashansky, A. J. *et al.* Mapping single-cell atlases throughout Metazoa unravels cell type evolution. *Elife* **10** (2021). https://doi.org:10.7554/eLife.66747
115 Chen, J. *et al.* A quantitative framework for characterizing the evolutionary history of mammalian gene expression. *Genome Res* **29**, 53-63 (2019). https://doi.org:10.1101/gr.237636.118
116 Bakken, T. E. *et al.* Comparative cellular analysis of motor cortex in human, marmoset and mouse. *Nature* **598**, 111-119 (2021). https://doi.org:10.1038/s41586-021-03465-8
117 Caglayan, E. *et al.* Molecular features driving cellular complexity of human brain evolution. *Nature* **620**, 145-153 (2023). https://doi.org:10.1038/s41586-023-06338-4

118 Arendt, D. *et al.* The origin and evolution of cell types. *Nat Rev Genet* **17**, 744-757 (2016). https://doi.org:10.1038/nrg.2016.127

119 Colquitt, B. M., Merullo, D. P., Konopka, G., Roberts, T. F. & Brainard, M. S. Cellular transcriptomics reveals evolutionary identities of songbird vocal circuits. *Science* **371** (2021). https://doi.org:10.1126/science.abd9704

120 Kebschull, J. M. *et al.* Cerebellar nuclei evolved by repeatedly duplicating a conserved cell-type set. *Science* **370** (2020). https://doi.org:10.1126/science.abd5059

121 Lamanna, F. *et al.* Reconstructing the ancestral vertebrate brain using a lamprey neural cell type atlas. *bioRxiv*, 2022.2002.2028.482278 (2022). https://doi.org:10.1101/2022.02.28.482278

122 Shafer, M. E. R., Sawh, A. N. & Schier, A. F. Gene family evolution underlies cell-type diversification in the hypothalamus of teleosts. *Nat Ecol Evol* **6**, 63-76 (2022). https://doi.org:10.1038/s41559-021-01580-3

123 Tarashansky, A. J., Xue, Y., Li, P., Quake, S. R. & Wang, B. Self-assembling manifolds in single-cell RNA sequencing data. *Elife* **8** (2019). https://doi.org:10.7554/eLife.48994

124 Agboola, O. S., Hu, X., Shan, Z., Wu, Y. & Lei, L. Brain organoid: a 3D technology for investigating cellular composition and interactions in human neurological development and disease models in vitro. *Stem Cell Res Ther* **12**, 430 (2021). https://doi.org:10.1186/s13287-021-02369-8

125 Lancaster, M. A. *et al.* Cerebral organoids model human brain development and microcephaly. *Nature* **501**, 373-379 (2013). https://doi.org:10.1038/nature12517

126 Amiri, A. *et al.* Transcriptome and epigenome landscape of human cortical development modeled in organoids. *Science* **362** (2018). https://doi.org:10.1126/science.aat6720

127 Bhaduri, A. *et al.* Cell stress in cortical organoids impairs molecular subtype specification. *Nature* **578**, 142-148 (2020). https://doi.org:10.1038/s41586-020-1962-0

128 Camp, J. G. *et al.* Human cerebral organoids recapitulate gene expression programs of fetal neocortex development. *Proc Natl Acad Sci U S A* **112**, 15672-15677 (2015). https://doi.org:10.1073/pnas.1520760112

129 Gordon, A. *et al.* Long-term maturation of human cortical organoids matches key early postnatal transitions. *Nat Neurosci* **24**, 331-342 (2021). https://doi.org:10.1038/s41593-021-00802-y

130 Pollen, A. A. *et al.* Establishing Cerebral Organoids as Models of Human-Specific Brain Evolution. *Cell* **176**, 743-756 e717 (2019). https://doi.org:10.1016/j.cell.2019.01.017

131 Velasco, S. *et al.* Individual brain organoids reproducibly form cell diversity of the human cerebral cortex. *Nature* **570**, 523-527 (2019). https://doi.org:10.1038/s41586-019-1289-x

132 Qian, X. *et al.* Sliced Human Cortical Organoids for Modeling Distinct Cortical Layer Formation. *Cell Stem Cell* **26**, 766-781 e769 (2020). https://doi.org:10.1016/j.stem.2020.02.002

133 Xiang, Y. *et al.* Fusion of Regionally Specified hPSC-Derived Organoids Models Human Brain Development and Interneuron Migration. *Cell Stem Cell* **21**, 383-398 e387 (2017). https://doi.org:10.1016/j.stem.2017.07.007

134 Pinson, A. *et al.* Human TKTL1 implies greater neurogenesis in frontal neocortex of modern humans than Neanderthals. *Science* **377**, eabl6422 (2022). https://doi.org:10.1126/science.abl6422

135 Trujillo, C. A. *et al.* Reintroduction of the archaic variant of NOVA1 in cortical organoids alters neurodevelopment. *Science* **371** (2021). https://doi.org:10.1126/science.aax2537

136 Kanton, S. *et al.* Organoid single-cell genomic atlas uncovers human-specific features of brain development. *Nature* **574**, 418-422 (2019). https://doi.org:10.1038/s41586-019-1654-9

137 Uzquiano, A. *et al.* Proper acquisition of cell class identity in organoids allows definition of fate specification programs of the human cerebral cortex. *Cell* **185**, 3770-3788 e3727 (2022). https://doi.org:10.1016/j.cell.2022.09.010
138 Fleck, J. S. *et al.* Inferring and perturbing cell fate regulomes in human brain organoids. *Nature* (2022). https://doi.org:10.1038/s41586-022-05279-8
139 Li, C. *et al.* Single-cell brain organoid screening identifies developmental defects in autism. *bioRxiv*, 2022.2009.2015.508118 (2022). https://doi.org:10.1101/2022.09.15.508118
140 Bershteyn, M. *et al.* Human iPSC-Derived Cerebral Organoids Model Cellular Features of Lissencephaly and Reveal Prolonged Mitosis of Outer Radial Glia. *Cell Stem Cell* **20**, 435-449 e434 (2017). https://doi.org:10.1016/j.stem.2016.12.007
141 Trevino, A. E. *et al.* Chromatin accessibility dynamics in a model of human forebrain development. *Science* **367** (2020). https://doi.org:10.1126/science.aay1645
142 Mora-Bermudez, F. *et al.* Differences and similarities between human and chimpanzee neural progenitors during cerebral cortex development. *Elife* **5** (2016). https://doi.org:10.7554/eLife.18683
143 Tanaka, Y., Cakir, B., Xiang, Y., Sullivan, G. J. & Park, I. H. Synthetic Analyses of Single-Cell Transcriptomes from Multiple Brain Organoids and Fetal Brain. *Cell Rep* **30**, 1682-1689 e1683 (2020). https://doi.org:10.1016/j.celrep.2020.01.038
144 Werner, J. M. & Gillis, J. Preservation of co-expression defines the primary tissue fidelity of human neural organoids. *bioRxiv* (2023). https://doi.org:10.1101/2023.03.31.535112
145 Sarieva, K. & Mayer, S. The Effects of Environmental Adversities on Human Neocortical Neurogenesis Modeled in Brain Organoids. *Front Mol Biosci* **8**, 686410 (2021). https://doi.org:10.3389/fmolb.2021.686410
146 Czerminski, J. T., King, O. D. & Lawrence, J. B. Large-scale organoid study suggests effects of trisomy 21 on early fetal neurodevelopment are more subtle than variability between isogenic lines and experiments. *Front Neurosci* **16**, 972201 (2022). https://doi.org:10.3389/fnins.2022.972201
147 Lee, J. H. *et al.* Cell-line dependency in cerebral organoid induction: cautionary observations in Alzheimer's disease patient-derived induced pluripotent stem cells. *Mol Brain* **15**, 46 (2022). https://doi.org:10.1186/s13041-022-00928-5
148 Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* **36**, 411-420 (2018). https://doi.org:10.1038/nbt.4096
149 Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat Biotechnol* **36**, 421-427 (2018). https://doi.org:10.1038/nbt.4091
150 Johansen, N. & Quon, G. scAlign: a tool for alignment, integration, and rare cell identification from scRNA-seq data. *Genome Biol* **20**, 166 (2019). https://doi.org:10.1186/s13059-019-1766-4
151 Polanski, K. *et al.* BBKNN: fast batch alignment of single cell transcriptomes. *Bioinformatics* **36**, 964-965 (2020). https://doi.org:10.1093/bioinformatics/btz625
152 Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902 e1821 (2019). https://doi.org:10.1016/j.cell.2019.05.031
153 Welch, J. D. *et al.* Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell* **177**, 1873-1887 e1817 (2019). https://doi.org:10.1016/j.cell.2019.05.006
154 Mostajo-Radji, M. A., Schmitz, M. T., Montoya, S. T. & Pollen, A. A. Reverse engineering human brain evolution using organoid models. *Brain Res* **1729**, 146582 (2020). https://doi.org:10.1016/j.brainres.2019.146582

155 Krienen, F. M. *et al.* A marmoset brain cell census reveals persistent influence of developmental origin on neurons. *bioRxiv*, 2022.2010.2018.512442 (2022). https://doi.org:10.1101/2022.10.18.512442
156 Caglayan, E., Liu, Y. & Konopka, G. Neuronal ambient RNA contamination causes misinterpreted and masked cell types in brain single-nuclei datasets. *Neuron* **110**, 4043-4056 e4045 (2022). https://doi.org:10.1016/j.neuron.2022.09.010
157 Soreq, L. *et al.* Major Shifts in Glial Regional Identity Are a Transcriptional Hallmark of Human Brain Aging. *Cell Rep* **18**, 557-570 (2017). https://doi.org:10.1016/j.celrep.2016.12.011
158 Gorin, G., Fang, M., Chari, T. & Pachter, L. RNA velocity unraveled. *PLoS Comput Biol* **18**, e1010492 (2022). https://doi.org:10.1371/journal.pcbi.1010492
159 Norimoto, H. *et al.* A claustrum in reptiles and its role in slow-wave sleep. *Nature* **578**, 413-418 (2020). https://doi.org:10.1038/s41586-020-1993-6

ACKNOWLEDGEMENTS:

The authors thank [list of co-authors of the other two pieces not directly involved in this one] for the insightful feedback on this work. M.C. would like to thank Alina Ulezko Antonova for helpful additional discussions. Additional funding acknowledgements: G.K. is a Jon Heighten Scholar in Autism Research and Townsend Distinguished Chair in Research on Autism Spectrum Disorders at UT Southwestern, and was partially supported by the James S. McDonnell Foundation 21st Century Science Initiative in Understanding Human Cognition Scholar Award (220020467), Simons Foundation for Autism Research Award (947591), and NIH grants (NS126143, HG011641, MH126481, NS115821). S.A.L. is supported by the Carol and Gene Ludwig Family Foundation, The Cure Alzheimer's Fund, the National Multiple Sclerosis Society, and NIH grants (R01EY033353, R03NS127079). C.R.C. is supported by the Weill Neurohub, the Shurl and Kay Curci Foundation, the CURE Epilepsy Taking Flight Award, and grants from the NIH (K08NS126573 and U01NS132353). E.C. is a Neural Scientist Training Program Fellow in the Peter O'Donnell Brain Institute at UT Southwestern. J.G. is supported by NIH grant R01MH113005. M.S. is supported by a HHMI Hanna H. Gray Fellowship, a SFARI Pilot Award, and a NYSCF Druckenmiller Fellowship. J.W. is supported by NIH grant R01MH113005. M.E.U. is supported by the National Science Foundation (NSF) Graduate Research Fellowship Program (GRFP) and a Hertz Foundation Fellowship Program.