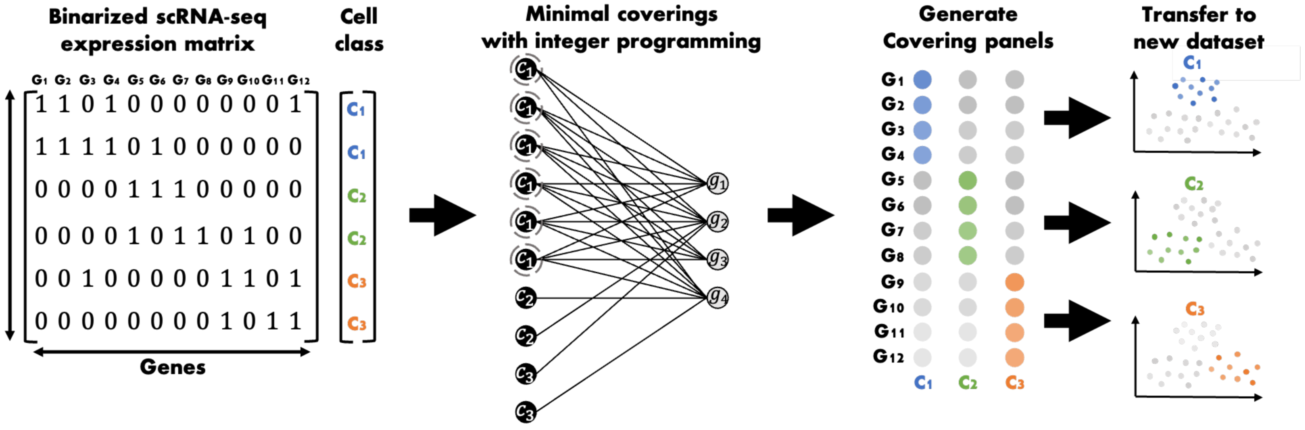
CellCover Defines Conserved Cell Types and Temporal Progression in scRNA-seq Data across Mammalian Neocortical Development

Lanlan Ji[1], An Wang[1], Shreyash Sonthalia[2], Daniel Q. Naiman[1], Laurent Younes[1,3], Carlo Colantuoni[2,4*], and Donald Geman[1,3*]

[1]Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, MD, USA
[2]Departments of Neurology and Neuroscience, Johns Hopkins University, Baltimore, MD, USA
[3]Center for Imaging Science, Johns Hopkins University, Baltimore, MD, USA
[4]Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD, USA

*To whom correspondence should be adressed:
Donald Geman, PhD; geman@jhu.edu Carlo Colantuoni, PhD; ccolantu@jhmi.edu

Graphical Abstract



G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12	Cell class
1	1	0	1	0	0	0	0	0	0	0	1	C1
1	1	1	1	0	1	0	0	0	0	0	0	C1
0	0	0	0	1	1	1	0	0	0	0	0	C2
0	0	0	0	1	0	1	1	0	1	0	0	C2
0	0	1	0	0	0	0	0	1	1	0	1	C3
0	0	0	0	0	0	0	0	1	0	1	1	C3

Contents

1 Abstract

Accurate identification of cell classes across the tissues of living organisms is central in the analysis of growing atlases of single-cell RNA sequencing (scRNA-seq) data across biomedicine. Such analyses are often based on the existence of highly discriminating "marker genes" for specific cell classes which enables a deeper functional understanding of these classes as well as their identification in new, related datasets. Currently, marker genes are defined by methods that serially assess the level of differential expression (DE) of individual genes across landscapes of diverse cells. This serial approach has been extremely useful, but is limited because it ignores possible redundancy or complementarity across genes, that can only be captured by analyzing several genes at the same time. We wish to identify discriminating *panels* of genes. To efficiently explore the vast space of possible marker panels, leverage the large number of cells often sequenced, and overcome zero-inflation in scRNA-seq data, we propose viewing panel selection as a variation of the "minimal set-covering problem" in combinatorial optimization which can be solved with integer programming. In this formulation, the covering elements are genes, and the objects to be covered are cells of a particular class, where a cell is covered by a gene if that gene is expressed in that cell. Our method, CellCover, identifies a panel of marker genes in scRNA-seq data that covers one class of cells within a population. We apply this method to generate covering marker gene panels which characterize cells of the developing mouse neocortex as postmitotic neurons are generated from neural progenitor cells (NPCs). We show that CellCover captures cell class-specific signals distinct from those defined by DE methods and that CellCover's compact gene panels can be expanded to explore cell type specific function.Transfer learning experiments exploring these covering panels across *in vivo* mouse, primate, and human scRNA-seq datasets demonstrate that CellCover identifies markers of conserved cell classes in neurogenesis, as well as markers of temporal progression in the molecular identity of these cell types across development of the mammalian neocortex. The gene covering panels we identify across cell types and developmental time can be freely explored in visualizations across all the public data we use in this report at [1] with NeMo Analytics [2]. The code for CellCover is written in R and the Gurobi R interface and is available at [3].

2 Introduction

scRNA-seq technologies provide measurements of RNA molecules in many thousands of individual cells, a rich source of information for determining attributes of cell populations, such as cell types and the variation in gene expression from cell to cell, which are not available from bulk RNA sequencing data [4, 5, 6, 7, 8]. After standard preprocessing and normalization, a core challenge in the analysis of scRNA-seq data is to annotate the cells in a given dataset with a label indicating cell class, type or state. This usually involves multiple phases of processing, notably dimension reduction, clustering, and finding "marker genes" for the labels of interest. Markers are usually selected by first ranking the genes based on a statistical test for differential expression (DE) across labels and then selecting a panel of top-ranking genes based on manual curation, estimates of predictive capacity, a priori information (e.g., cell-surface proteins), or panel size [9, 10, 11, 12, 13, 14, 8, 15]. Importantly, the construction of such marker panels incorporates data from only one gene at a time, i.e. the panel is assembled "gene by gene," and does not take into account possible gene combinations that could better distinguish among cell classes.

Complications arise because scRNA-seq data are high-dimensional [16], highly heterogeneous, generally noisier than bulk RNA-seq, and stochastically zero-inflated. Complex pipelines with many choices require manual intervention [9, 17], and raise issues of bias and reproducibility [18, 19]. In order for downstream analyses to be computationally feasible, restricting dimension, such as using small panels of marker genes, is necessary. Analysis of scRNA-seq data is generally done at the univariate (single gene) level, even though the relevant biology is often decidedly multivariate. In particular, probability distributions, when proposed, are for the expression of individual genes (i.e., marginal distributions) not for gene panels (i.e., higher-order marginals). These analyses do not take into account how the individual marker genes may be cooperating to determine cell labels, i.e., what are the statistical interactions among the marker genes which characterize the label.

To avoid these limitations, we formulate marker gene selection as a variation of the well-known "minimal set-covering problem" in combinatorial optimization. In our case, the "covering" elements are genes, and the objects to be covered are the cells in some class or population. A cell is covered at depth $\boldsymbol{d}$ by a gene panel if at least $\boldsymbol{d}$ genes in this panel are expressed in the cell. A set of genes is then a depth $\boldsymbol{d}$ marker panel for a population if nearly all cells are covered at depth $\boldsymbol{d}$; the fraction of uncovered cells can be interpreted as a false negative rate; see Methods (section 5) The optimal marker panel then minimizes a performance measure, described in Methods, based on the discriminative power of each gene. Moreover, this can be found by integer linear programming. Consequently, our method only uses binarized raw mRNA counts, separating genes into those which are "expressed" (positive count) or "not expressed" (zero count). We are then covering (nearly) all cells of a class with a panel of genes such that for every cell in the class, at least $\boldsymbol{d}$ genes in the panel are expressed. In addition to enabling the link to set covering, binarization facilitates the biological interpretation of marker genes and the manner in which they characterize and discriminate among types of cells. Using simple and transparent cell functions, such as the number of covering panel genes expressed in a cell, the covering paradigm enables the transfer of covering marker panels to related expression datasets for the identification of conserved cell classes and shared cellular processes across diverse datasets. Additionally, binarization obviates the need for complex data normalization that is often necessary in exploring continuous scRNA-seq data (especially across differing technologies), and we show evidence that binarization may also mitigate noise and batch/artefactual effects.

The minimal covering approach in CellCover is designed specifically to leverage the high number of cells routinely observed in scRNA-seq data in order to be robust to the zero-inflation known to be rampant in this data modality. Because the covering algorithm optimizes a set of genes (rather than individual markers) and different true marker genes will have drop-outs in different cells, this stochastic zero-inflation can be overcome if many cells of a class are sequenced. In contrast, standard methods, based on selecting the markers to be those genes for which the p-value associated with a Wilcoxon Rank Sum (or some other test statistic) falls below a threshold, cannot borrow strength across genes in a marker panel when selected gene-by-gene. CellCover searches for small panels of covering marker genes that precisely define cell classes together as a set, and are also expandable, i.e., can be readily linked to additional related genes for the deeper exploration of cell-type specific function. In addition, CellCover includes user-adjustable constraints to promote specificity at high sensitivity, as well as tools to facilitate gene set over-representation analysis and the exploration of heterogeneity within individual cell classes. Here we use CellCover to explore cell classes across neurogenesis in the excitatory neuronal lineage of the neocortex in mammals.

3 Results

3.1 scRNA-seq datasets in mammalian neocortical neurogenesis

We illustrate the use of CellCover in the context of dorsal pallial neurogenesis during mid-gestation, when the majority of excitatory neurons in the neocortex are produced. The mature neocortex in mammals is made up of 6 layers of post-mitotic neurons. Neurons of each new layer are produced in succession, with neurons of the new layer migrating past previously created layers, leading to an "inside-out" arrangement of neurons by birthdate [20], with deep layers (VI-V) being born first and superficial layers (II-IV) born last. Neurons are generated sequentially from radial glia (RG), the neural stem cells of the telencephalon, and intermediate progenitor cells (IPCs) through both symmetric and asymmetric cell division [21]. The number and diversity of neural stem and progenitor cells in the cortex have been greatly expanded in the evolution of the primate lineage. In particular, the appearance of the outer subventricular zone (OSVZ) in primates appears to underly the development of the supragranular layers of the cortex (layers II and III) which contribute to higher cognition through extensive cortico-cortico connections [22]. Notably, this includes greater proliferative capacity of neural stem cells and the expansion of outer radial glia (oRG, or basal radial glia, bRG), which are rare in rodents but numerous in primates and especially in humans [23]. As each cortical layer is generated, newborn neurons arise contemporaneously from progenitor cells in the germinal zones (GZ) and migrate radially along processes of the RG, outward toward the surface of the cortex to the cortical plate (CP), coming to rest together at their characteristic position in the developing cortex [24]. Neurons arising at the same time and place, and therefore from precursors with common molecular features, share unique morphological and functional characteristics. We are interested both in the general process by which neural precursor cells yield post-mitotic neurons, and in the more nuanced dynamics by which specific properties are imparted to neurons which are born together and which come to rest in the same cellular microenvironment, i.e., cortical region and layer.

How the molecular dynamics of neural progenitor identity across this progression impact the fate and function of the neurons being produced is not fully understood. To explore both the principal cell identities and the developmental progression within cell types, we first focus on scRNA-seq data (Telley et al. [25]) from cells extracted during neocortical neurogenesis in the mouse, and for which highly precise temporal labels are available. This dataset, referred to as the "Telley dataset" from now on, contains expression data from cells produced across four days of mid-gestational embryonic development during the peak of excitatory neocortical neurogenesis spanning embryonic days E12, E13, E14, E15 and sequenced using Fluidigm C1- based methods. Using a technique to specifically label ventricular neural progenitor cells during their final cell division as post-mitotic neurons are produced, cells born on each of the four embryonic days were sampled at one hour, one day, and four days after this terminal division, yielding a two-dimensional temporal indexing of the data based on age and embryonic date. We define covering marker gene panels in this dataset and validate their ability to capture conserved cell-type specific signals in additional scRNA-seq data from developing mouse [26], primate [27], and human [28, 29, 30] cortical tissue.

3.2 Binarizing scRNA-seq count data

Binarization of gene expression data is the first step in exploring minimal coverings of cell types with CellCover, and has been proposed in several contexts, including the exploration of cell-type specific transcriptional dynamics [31]. Before moving on to the definition of marker genes with CellCover, we use the Telley dataset [25] to show that binarization (where zero counts remain zero and all positive counts map to 1) of scRNA-seq data preserves key cell type specific structure in the expression data. The Telley data contain ground truth cell class labels based on the precise time since a progenitor cell's terminal division (1, 24, or 96 hours prior to sequencing) and the age of the pup at the time of that terminal division (Embryonic days 12, 13, 14, or 15). Hence, the scRNA-seq data contain progenitors (1H) and neurons (96H), as well as cells transitioning between these states (24H) across a range of developmental ages (E12-E15). We evaluated the capacity of binarized data to distinguish among the 12 classes, namely 3 (times since final division) × 4 (pup ages) using two methods:

(1) prediction/recovery of ground truth cell labels using supervised machine learning (Table 1 and Supplementary Table 1) and

(2) unsupervised dimension reduction and cell clustering (Supplementary Figure 1 and Supplementary Table 2).

XGB	$\mathbf{XGB_b}$	$\mathbf{XGB^{PCA}}$	$\mathbf{XGB_b^{PCA}}$	**NN1**	$\mathbf{NN1_b}$	**NN2**	$\mathbf{NN2_b}$	**NN3**	$\mathbf{NN3_b}$
95.0	93.6	86.5	91.2	93.8	97.1	93.3	96.4	87.9	94.9

Table 1: **Average accuracies of different data types and supervised classifiers in recovering the 12 time-based cell labels in the Telley data.** Subscript "b" indicates binary data; all other trials used normalized log2 count level data. XGB = XGBOOST [32]. "PCA" indicates using only the 1st 100 principal components. NNx refers to a neural network with x hidden layers and 64 total nodes in the network. The size of the training set is 2342, and the size of the test set is 414.

Among the combinations of data type and prediction model, the PCA-processed binarized data, trained with a one-layer neural network, achieves the highest average accuracy in predicting the 12 developmental stages in the Telley data [25] (Table 1). Similar results are observed in the dataset from Polioudakis et al. [28], where binarized data consistently outperform the $\log_2$ count level data in predicting the cell location (germinal zone vs. cortical plate). We also assessed the impact of binarization as input to non-linear embedding algorithms such as t-SNE [33] or UMAP [34], followed by cell type clustering [35]. Using the popular Seurat package in R [36], we calculated the PCs, t-SNE embeddings, and Louvain cell clusters in the Telley dataset with 12 ground truth developmental cell labels. Supplementary Figure 1 shows the results of this analysis using count level data versus binarized expression data in a standard Seurat [36] pipeline analysis. Strikingly, the binary data achieve clearer separation of cells based on both the time since terminal division and the pup age at the time of that division. This is shown quantitatively in Supplementary Table 2 using adjusted Rand indices (ARIs [37]) as a measure of concordance between the ground truth labels and the unsupervised clusters obtained with continuous or binary data.

3.3 Marker gene panels for stages in mammalian neocortical development

We apply the concept of minimal covering to binarized scRNA-seq data, and show that cell developmental stages can be inferred by simply counting how many marker genes are "on" (expressed) in one stage and "off" (zero counts) at other stages. In contrast to standard gene marker identification in scRNA-seq [17, 36], which assesses the utility of each gene individually, our approach identifies an optimal panel of genes which functions together to most precisely discriminate one stage from the others. We formulate this search as a constrained optimization problem amenable to integer programming (Ke et al. [38]; see Methods). In this example, the objects to be covered are the cells of some stage of neural development and a cell is covered by a marker gene if that gene is expressed, i.e., has a non-zero raw count, in that cell. Two basic parameters define admissible coverings: the *depth*, denoted $\boldsymbol{d}$, which is the minimum number of genes in the marker panel for a stage that we expect to be expressed; and the *false negative covering rate*, denoted α, which is the proportion of cells in the class of interest (here stages) which are not covered at depth $\boldsymbol{d}$.

Hence, the algorithm aims to optimize a performance measure over gene sets for which the fraction of cells expressing at least $\boldsymbol{d}$ genes in the panel is at least $1 - \alpha$.

Using this approach, we define marker gene panels which cover the primary cell stages across neurogenesis in the excitatory neuronal lineage of the mouse neocortex using data visualized in Figure 1A [25]. Only cells undergoing their terminal division at the ventricular surface were sequenced in this study, and the exact timing of this division is annotated for each cell, enabling the precise staging of cells through the progression of NPCs to postmitotic neurons. Therefore, rather than using scRNA-seq data-driven clustering, or biologist-annotated cell labels, we employ experimental timing variables to define cell classes. Accordingly, all cells have one of the three labels “1H”, “24H”, “96H” representing the number of hours since terminal division, which we use as ground truth cell class labels for defining covering marker gene panels. Figure 1A depicts the proportion of cells in each class which express the derived marker genes at non-zero levels. The original authors noted heterogeneity in the 24H cell class, which can be seen as more frequent expression of the 24H cell markers in the other time classes (Figure 1A, 24H marker panel in center).

These covering marker gene panels contain canonical markers of neural progenitors, intermediate progenitors and neurons, including Sox3, Neurod1, and Mef2c in 1H, 24H, and 96H cells, respectively. The panels also contain many lesser known genes (Gm . . .), pseudogenes (. . . -ps), and mitochondrialy encoded genes (mt-. . .) which are expressed at much lower levels, but together precisely distinguish these neurogenic cell classes. Mitochondrial transcripts, and tRNAs in particular, are often disregarded as uninformative “housekeeping” genes or even indications of low-quality RNA samples. These are dangerous assumptions, especially when exploring neurogenesis where the mitochondrial content of cells changes dramatically as cells transition from stem and progenitor states to post-mitotic neurons. The expression of mt-Tv, for example, is highly specific to 96H cells (Figure 1A), and while often disregarded as a housekeeping tRNA, it also forms a structural component of the mitochondrial ribosome [39] and is transcribed in higher abundance as energy demands of nascent neurons increase. This suggests that many genes not often selected by conventional marker gene finding methods (and often intentionally filtered out by biologists) hold significant cell-type specific information. To assess this possibility and the utility of covering marker gene panels in general, we examine the ability of these panels to define cell types in additional datasets in mammalian neocortical neurogenesis.

3.4 Transfer of covering marker gene panels #1: Conserved cell type markers in mammalian neocortical neurogenesis

To assess the ability of covering marker gene panels to capture consistent cell-type specific signals in murine neurogenesis we explored additional *in vivo* data from the dorsal mouse telencephalon. We examined the expression of covering panels derived from the Telley mouse dataset in a recent atlas of the developing mouse brain (La Manno et al. [26]; Figure 1B and 1C). For both these mouse scRNA-seq datasets the precise gestational age of the mouse pup of origin is known for all cells collected. However, while there are ground truth labels of the time from ventricular progenitor terminal cell division in the Telley data, in the atlas we must rely on cell type annotation provided by the researchers, which include radial glia, neuroblast, and neuron classes sequentially along the neurogenic trajectory. As expected, marker genes derived from progenitors labeled one hour after their terminal cell division in the Telley data, i.e., while still in the progenitor state, were most frequently expressed in the progenitors (i.e., radial glia cells) of the mouse dorsal forebrain atlas. Correspondingly, markers of 24H cells were most frequently expressed in neuroblasts (and some neurons, as noted by original authors) and markers of 96H cells in neurons of the LaManno dataset. These observations across datasets indicate that the covering marker gene panels in the Telley data accurately define the primary cells states traversed during neocortical neurogenesis in the mouse.

To explore if these gene panels define cortical cell types conserved across species in mammalian neurogenesis, we also examined expression of the panels derived from the mouse in scRNA-seq data from human mid-gestational cortical tissue (Polioudakis et al. [28]; Figure 1D and 1E). Markers of 1H mouse cells are most frequently and specifically expressed in the progenitor cells of the human neocortex. Markers of 24H cells are most frequently expressed in intermediate progenitor cells (IPCs) and earliest neurons that have not migrated out of the germinal zone, while 96H markers show greatest frequency in later neurons resident in the cortical plate. This clear mapping of the murine covering markers onto human cortical data demonstrates that these gene panels capture neurogenic cell states conserved across mammalian neocortical development.

These transfers of the Telley 1H, 24H, and 96H covering marker gene panels into single-cell data of the mouse and human neocortex provide a cell-type level perspective on the marker panels. To further understand the signals captured in these panels, we transferred them into additional data with much less cell resolution but greater temporal and anatomical resolution. In bulk RNA-seq of human postmortem prefrontal cortical tissue [30], we observed a

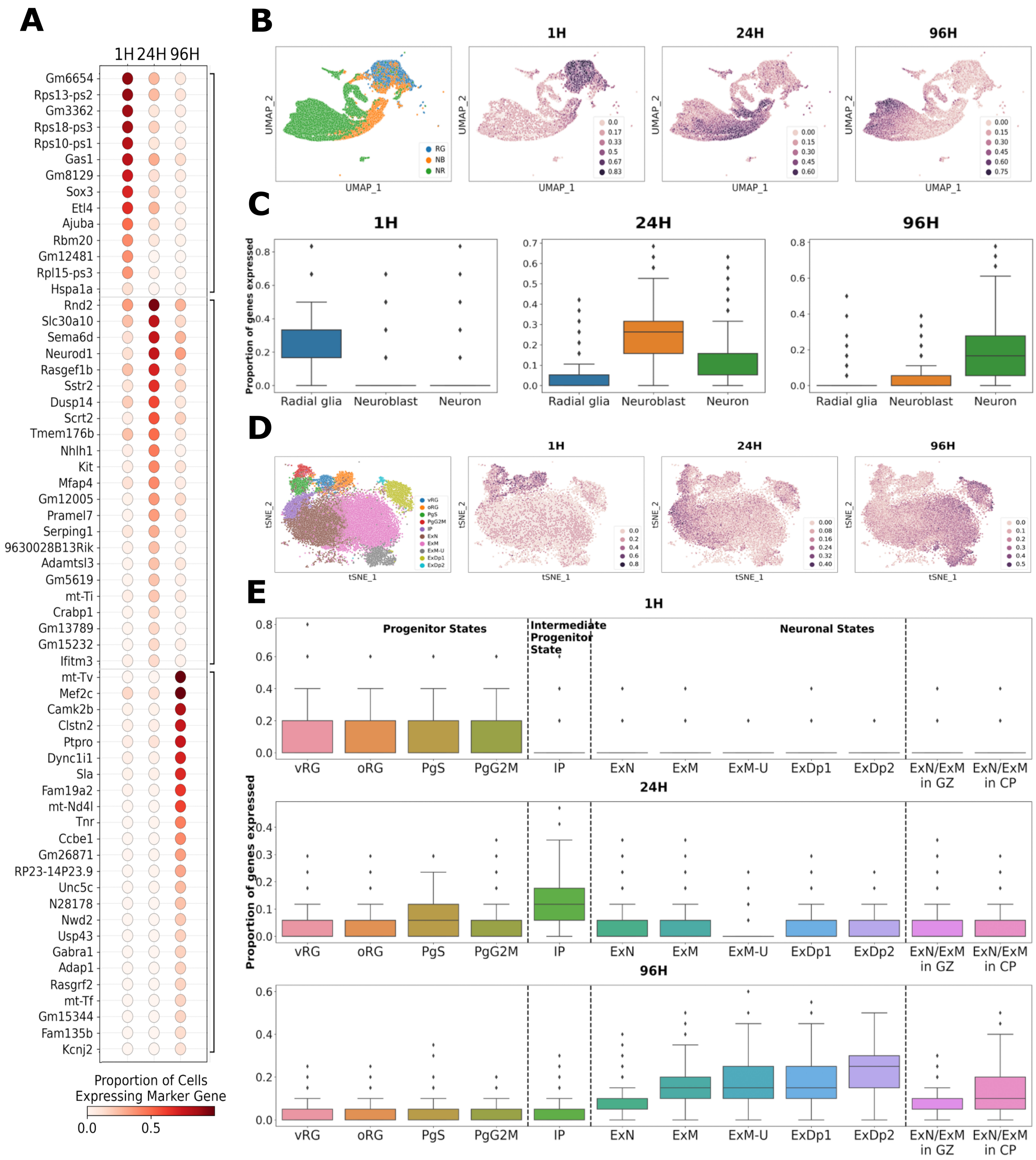

A
1H
24H
96H
Gm6654
Rps13-ps2
Gm3362
Rps18-ps3
Rps10-ps1
Gas1
Gm8129
Sox3
Etl4
Ajuba
Rbm20
Gm12481
Rpl15-ps3
Hspa1a
Rnd2
Slc30a10
Sema6d
Neurod1
Rasgef1b
Sstr2
Dusp14
Scrt2
Tmem176b
Nhlh1
Kit
Mfap4
Gm12005
Pramel7
Serping1
9630028B13Rik
Adamtsl3
Gm5619
mt-Ti
Crabp1
Gm13789
Gm15232
Ifitm3
mt-Tv
Mef2c
Camk2b
Clstn2
Ptpro
Dync1i1
Sla
Fam19a2
mt-Nd4l
Tnr
Ccbe1
Gm26871
RP23-14P23.9
Unc5c
N28178
Nwd2
Usp43
Gabra1
Adap1
Rasgrf2
mt-Tf
Gm15344
Fam135b
Kcnj2
Proportion of Cells
Expressing Marker Gene
0.0
0.5
B
1H
24H
96H
UMAP_1
UMAP_2
RG
NB
NR
0.0
0.17
0.33
0.5
0.67
0.83
0.00
0.15
0.30
0.45
0.60
0.00
0.15
0.30
0.45
0.60
0.75
C
1H
24H
96H
Proportion of genes expressed
Radial glia
Neuroblast
Neuron
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
D
1H
24H
96H
tSNE_1
tSNE_2
vRG
oRG
PgS
PgG2M
IP
ExN
ExM
ExM-U
ExDp1
ExDp2
0.0
0.2
0.4
0.6
0.8
0.00
0.08
0.16
0.24
0.32
0.40
0.0
0.1
0.2
0.3
0.4
0.5
E
1H
24H
96H
Progenitor States
Intermediate
Progenitor
State
Neuronal States
Proportion of genes expressed
vRG
oRG
PgS
PgG2M
IP
ExN
ExM
ExM-U
ExDp1
ExDp2
ExN/ExM
in GZ
ExN/ExM
in CP
0.0
0.1
0.2
0.3
0.4
0.6
0.8

Figure 1: **Transfer of CellCover marker gene panels to additional data in the developing mammalian neocortex. A.** Dot plot of empirical conditional expression probabilities of covering marker panels of each cell age. The marker genes of each cell age are grouped along the horizontal axis and sorted by expression frequency in the cell class of interest. The color of the dots represents the expression probability of covering markers conditioned on time since a cell's terminal division, i.e., the proportion of cells within a class expressing the marker gene. For this analysis, cells of each time point were pooled across embryonic ages (E12-15). The panel is obtained using the CellCover with $\alpha = 0.02$ and $\boldsymbol{d} = 5$. **B.** Transfer of covering marker gene panels from the Telley data [25] to a second mouse neocortex dataset shows consistent identification of the primary cell types in mouse neurogenesis. Left: UMAP representations of cells from the dorsal forebrain excitatory lineage in the LaManno atlas of mouse brain development [26], colored by cell labels assigned by the original authors. This is followed by the same UMAPs, now showing the proportion of gene panels derived from the Telley dataset, labeled 1H, 24H and 96H, expressed at non-zero levels in each individual cell. These last three plots illustrate the transfer of covering marker gene panels derived from the Telley data to the LaManno data. **C.** Box plots of these same proportions broken down by cell-type labels provided by the original authors. **D.** Transfer of covering marker gene panels from the Telley data in mouse to data in the developing human neocortex from Polioudakis et al. [28], shows the identification of conserved cell types in neurogenesis across mammalian species. Left: tSNE map of cells from the dorsal forebrain excitatory lineage in the Polioudakis data. This is followed by the same tSNE maps, now showing the transfer of covering marker gene panels derived from progenitors labeled 1, 24, and 96 hours after terminal cell division in the Telley data. The map is colored by the proportion of each gene panel that is expressed at non-zero levels in each individual cell of the human Polioudakis dataset. **E.** Box plots of these same proportions broken down by cell type labels and microdissection information provided by the original authors. The final two boxes indicate expression in neuronal subtypes segregated by physical location: germinal zone (GZ) or cortical plate (CP) microdissection.

decrease in the expression of 1H covering markers across human neocortical development *in utero* (Figure 2A). This is consistent with the progressive disappearance of neural progenitors as post-mitotic neurons are generated during peak neurogenesis in mid-gestation. The expression of 24H markers appears to peak around GW12. Covering markers of 96H cells increase in expression during fetal development, plateauing around GW16, consistent with the progressive accumulation of newly born neurons in the developing cortex during this period.

Cell states across mammalian neurogenesis are tightly linked to the position in the developing cortex. To explore the precise positional association of signals captured by covering panels from the developing mouse cortex, we transferred the markers derived from the Telley data into expression data from hundreds of laser microdissected regions of the developing primate cortex (Bakken et al. [27]; Figure 2B). Markers of the 1H mouse neural progenitor cells show the highest expression in the ventricular zone (VZ) of the macaque and decrease over fetal development (Figure 2B left panel), consistent with the widely held notion that ventricular radial glia are the earliest and most conserved neural progenitor state in the developing mammalian cortex. In contrast, markers of 24H cells in the mouse are most highly expressed in the cells of the inner subventricular zone (iSVZ) of the primate (Figure 2B center panel), indicating a progression from the ventricular zone state to delaminated intermediate progenitor (IPCs) / basal progenitor (BPs) states, but not reaching the primate-specific outer subventriclar zone (oSVZ) [22] / outer radial glial (oRG) states [23]. Genes marking 96H cells are most highly expressed in the post-mitotic neurons of the fetal cortex and decrease following birth, indicating that some of the genes in the 96H panel are specific to newborn neurons. It is of interest that in the VZ, the 1H signal is high and falling while the 96H signal is low but rising. These opposing dynamic signatures within mammalian neural progenitors appear to indicate that progenitor and neuron states begin to converge over development. Elements of this have been noted elsewhere [25, 40] and become clearer later in this report as we perform additional transfers of covering marker gene panels at higher resolution (Figures 5 and 6).

3.5 Using deeper covering analysis to more broadly explore cell function

We propose an additional method in section 5.2 to leverage the cell type specific signals captured by CellCover in order to extend the exploration of what cells are (classes) to include what cells are doing (cell type specific functions). From a small covering panel, this will create a broader list of genes that can be used to explore cell-type specific function via gene set over-representation analysis and more complex transfer learning experiments. To expand a covering gene marker panel for a particular cell class, the CellCover algorithm is simply run again, taking the initial covering set as a starting point along with new, broader covering constraints (depth, $\boldsymbol{d}$ and proportion of uncovered

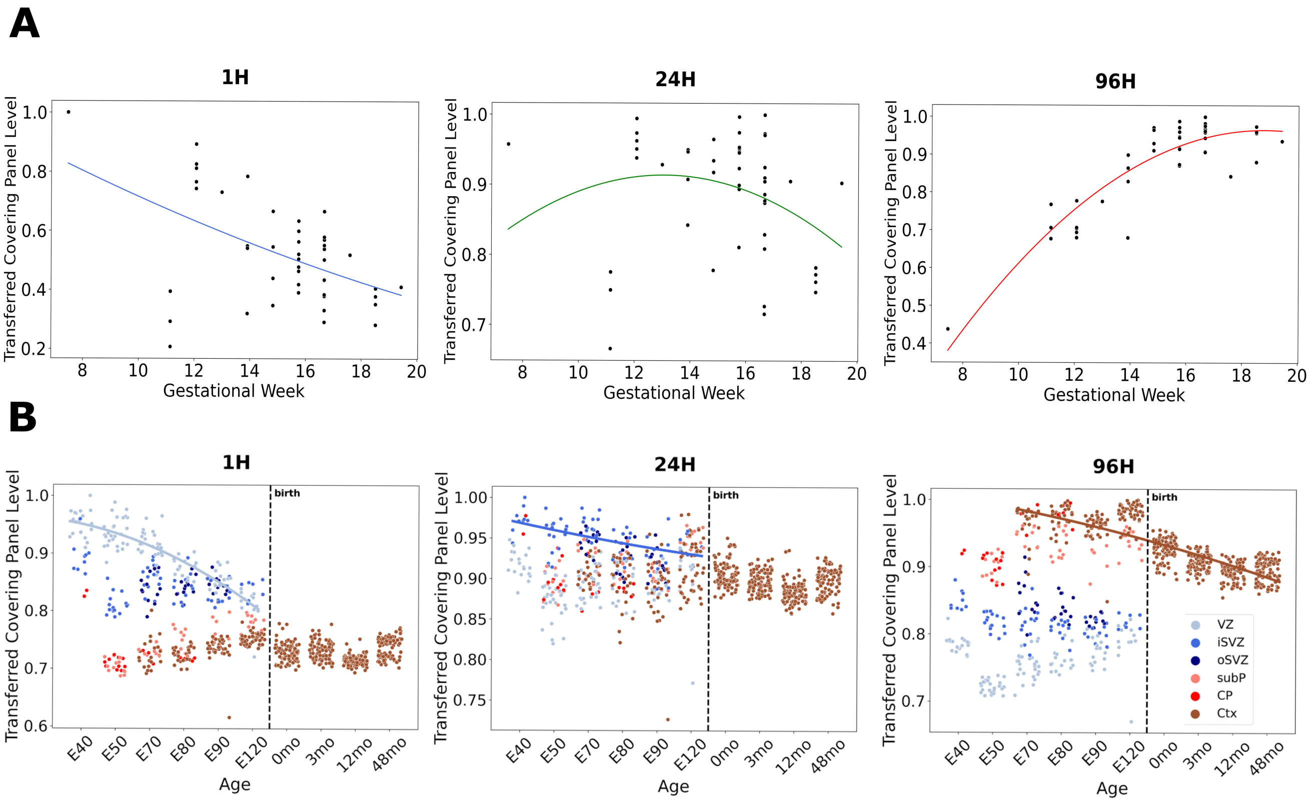


Figure 2: **A.** Transfer of covering marker gene panels from the Telley data in mouse [25] into bulk RNAseq from human fetal cortical tissue [30]. Because we cannot calculate covering rates as we did in single-cell data, here, the transferred values of covering panels were assessed as the sum of covering panel gene expression levels in each individual sample divided by the maximum sum observed in the samples. Hence, there is always a single maximum transfer value of 1 unit in each set of Y-axis values. **B.** Transfer of covering marker gene panels from the Telley data in mouse into microarray data from laser microdissected regions of the developing macaque neocortex [27]. Nonlinear fits in 1H, 24H, and 96H panels are to VZ, iSVZ, and Ctx data, respectively. X-axis ages are expressed as embryonic (E) days after conception and months (mo) after birth. Transferred covering panel levels were calculated as in panel A. VZ = ventricular zone, iSVZ= inner subventricular zone, oSVZ = outer subventricular zone, subP = subplate, CP = cortical plate, Ctx = cortex.

cells, α) as input. Enforcing the inclusion of the previously defined covering panel in this nested run ensures the original compact list of genes for cell identity will be a subset of the expanded panel to be used in the exploration of cell function and also reduces the CellCover run time (compared to a de novo run with the same expanded constraints).

Supplemental Table 4 lists expanded covering marker gene panels for each of the three cell ages in the Telley data that we have been exploring. The original covering gene marker panels of depth 3 were expanded to depths of 7, 15, 20, 25, 30, 35, and 40. To assess the capacity of the covering marker gene panels of differing size to characterize cell type specific function, we performed over-representation analysis using Gene Ontology (GO) on all these covering panels in addition to marker panels of equal size derived from conventional DE marker selection (Figure 3). In all cases, increasing panel sizes resulted in the detection of increasing numbers of significantly enriched gene sets. The one exception to this is in the 24H panels where, after an initial increase, a reduction in the number of hits is observed in both methods, before hits again begin to rise with increasing panel size. Strikingly, as the number of significant hits for both methods grows with increasing panel sizes, the specific gene sets found to be enriched by the two methods are quite distinct: 40-50% of the gene sets detected by one method are not found to be significant by the other. This is consistent with our other observations indicating that the two methods of gene marker definition extract fundamentally different signals from single-cell data.

3.6 Covering marker gene panels in mammalian neocortical development #2: Temporal Progression

Aiming to define cell type specific dynamics at higher resolution, we extended our previous covering analysis of the Telley data from the level of three cell ages (1, 24, and 96 hours) aggregated across embryonic days (E12-15), to 12 cell class labels, each representing a single cell age from one embryonic day (3 cell ages × 4 embryonic days = 12 cell classes). Supplementary Figure 4 depicts the individual genes in each of these 12 covering marker gene panels and the proportion of cells of each type which express them. Within each cell age (1H, 24H, or 96H), there is appreciable expression of marker genes across embryonic days. That is, markers for cells of one age at one embryonic day are often expressed in cells of the same age at other embryonic days. This indicates common molecular dynamics in progenitors following their terminal division at the ventricle, regardless of the embryonic day on which that division took place. These shared elements are most clear in E13 cells of all ages. At E14, 24H cells begin to become distinct from 24H cells of other embryonic ages, while 1H and 96H cells at E14 share dynamics across embryonic days. In contrast, at E12, 1H cells are most distinct from other 1H cells. This E12 1H panel includes Crabp2, which also marks a specific early NPC state in the primate: [41]). At E15, 24H and 96H cells are most distinct, paralleling the shift from deep to upper layer neuron generation (in section 3.8 we gain more granular insight into these observations as we use the transfer of the marker gene panels into additional primate and human data).

3.7 Genes in covering panels and marker genes defined by differential expression (DE) capture distinct elements of cell-type specific gene expression

Currently, the vast majority of cell type markers derived from scRNA-seq data are selected from the top of a ranked list of genes differentially expressed between a cell class of interest versus other cell types [17, 36]. This approach has yielded many insightful results across tissue systems and species. While this method optimizes individual genes as cell type markers (by their DE rank) and creates a set of markers by selecting top individual genes that distinguish a cell type, CellCover optimizes a panel that functions together to define a cell class. This fundamental difference in the search for marker genes results in the capture of distinct cell-type specific signals. Therefore we put forth CellCover as a complementary approach to conventional DE marker gene finding, rather than an optimized method for cell-type marker gene discovery.

Notable among the additional differences in these marker-finding methods is that CellCover (1) obviates the need for normalization, scaling or feature selection (as a function of utilizing binarized expression data), and (2) can directly capture heterogeneity within a cell class of interest. We show the latter empirically by comparing markers derived from CellCover and DE methods in the Telley data. Figure 4 summarizes the overlap of covering marker gene panels and the Seurat DE ranked genes for all 12 cell classes. In the majority of cell classes, there is appreciable divergence in the marker genes that the two methods identify as cell-type specific (represented by vertical gaps between the red and blue lines in each plot). Concordance between 96H cell markers is greatest (perhaps indicating lower within-class heterogeneity in neurons), with the shortest list of top DE selected markers that include an entire

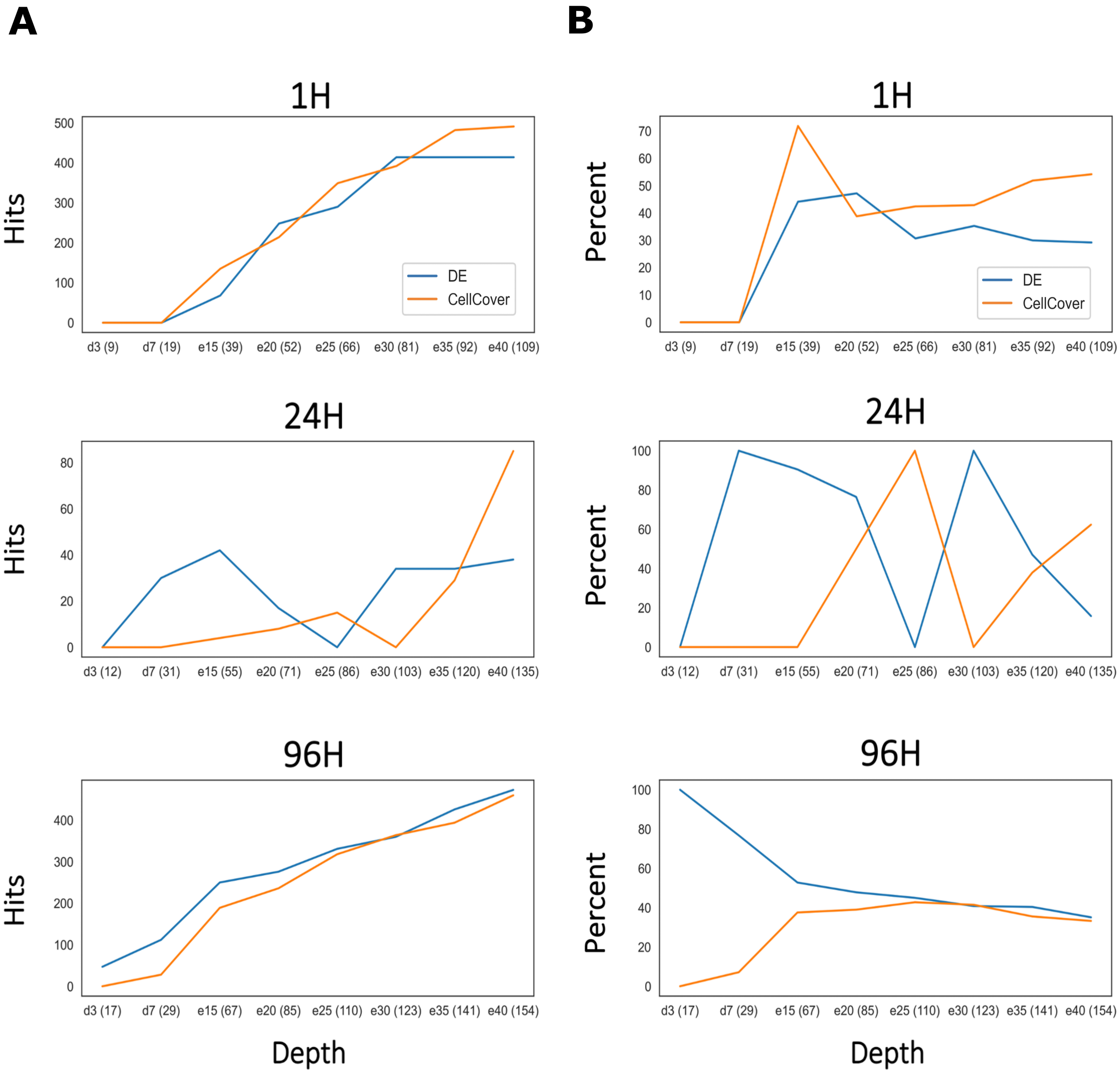


Figure 3: Over-representation analysis in gene marker panels of differing size and methodology, using the GO database. **A.** The number of GO gene sets found to be significantly enriched in the gene marker panels derived from CellCover and DE methods across different matched panel sizes. The X-axis indicates the depth parameter d that was set for the CellCover run (preceding the numbers on the X-axis, “d” indicates the depth in de novo CellCover runs, “e” indicates expanded runs that use marker panels from the preceding CellCover run as a starting point). Corresponding numbers of DE markers were selected for each depth so that marker panels of equal size were compared across CellCover and DE methods. Numbers in parentheses indicate the number of genes in each marker panel. **B.** The proportion of gene sets that were found to be significant by one method and not the other.

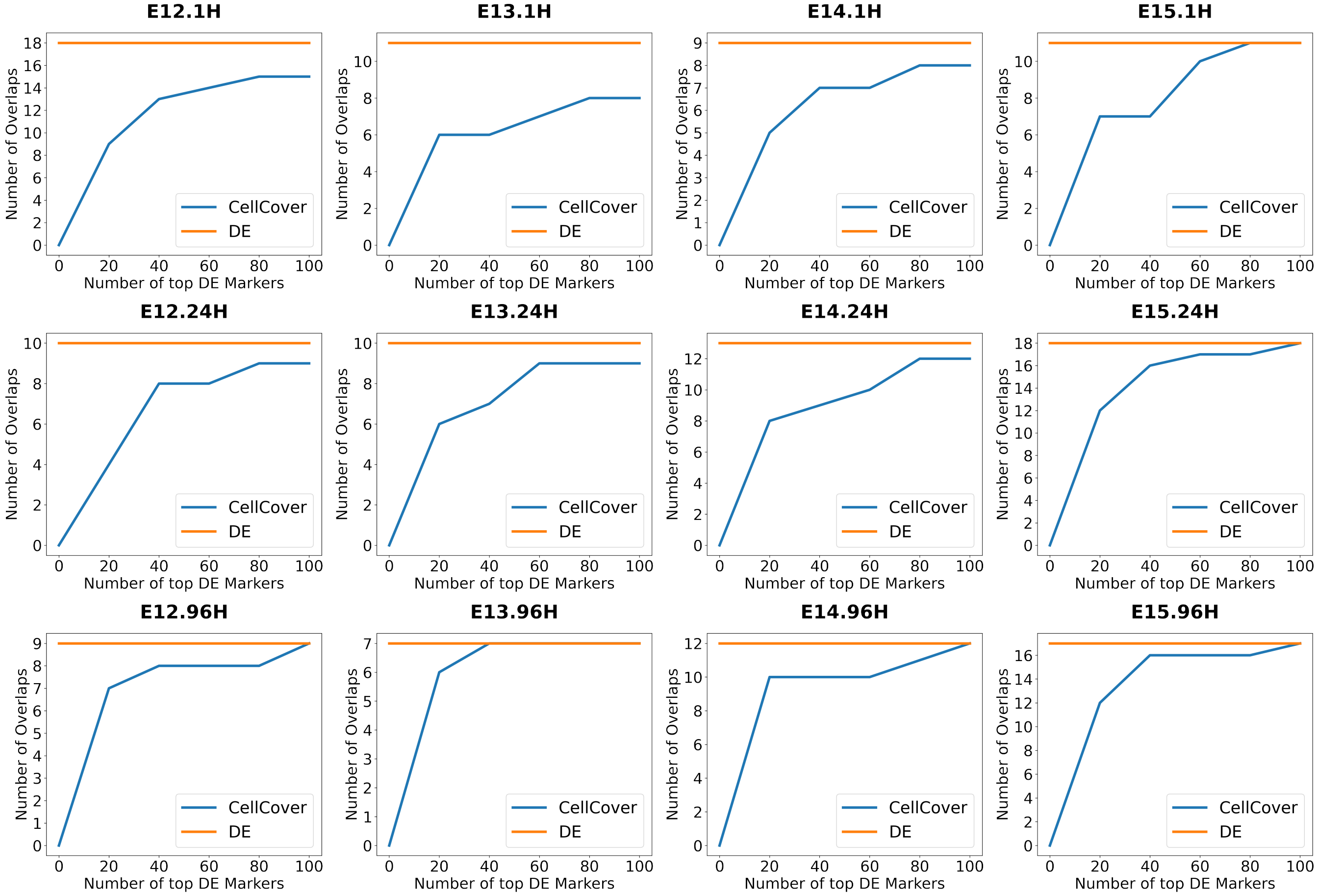


Figure 4: **Overlap and discordance between marker genes selected by CellCover and DE methods.** For each of the 12 CellCover gene panels defined in the Telley mouse data (depicted in Figure 4), the number of overlapping genes found by DE methods was quantified. In each panel, the red line indicates the number of genes in the CellCover gene panel for that cell class. The blue line indicates the cumulative number of genes also found in the ranked DE gene marker list as increasingly large DE gene lists are considered (X-axis; using the Seurat marker pipeline). The X-axis value where the red and blue lines meet indicates the number of Seurat markers that would need to be selected to include all the CellCover markers.

covering marker gene panel at 40 genes in length (E13.96H). These observations indicate that in practice, the two methods of marker gene discovery produce quite divergent results.

To further explore the discordance in marker genes selected by the two methods, we visualized expression levels of E12.1H covering markers with high p-value rank in the DE (Supplementary Figure 5 top panel) and those that are in the CellCover panel, but rank lowly in the DE ranked list (Supplementary Figure 5 bottom panel). Markers appearing in both lists are frequently expressed in E12.1H cells and are clearly expressed in other cell types, though less frequently. In contrast, some genes appearing in the CellCover solution but ranking lowly in the DE are less frequently expressed but are more specific to the E12.1H class. For example, genes Nptx2, Pctp2, and Clnd9 are not as frequently expressed in E12.1H cells as other markers. However, these genes are rarely expressed in other cell classes. These marker genes with high specificity are very discriminative for the purpose of defining cell classes when used together but may be omitted by DE marker selection methods in the feature selection or DE calculation steps due to low and/or less frequent expression, i.e., low sensitivity as an individual marker. By optimizing the panel of marker genes rather than individual markers, CellCover can borrow power across multiple genes that show greater specificity but which individually do not identify all cells in a class.

To show this more systematically, we ran differential expression analysis on the 1H, 24H, and 96H cells in the Telley data [25] and showed the rank of the CellCover markers (Figure 1A) based on the p-values returned from the DE analysis in Supplementary Figure 6. We found that across all the cell ages, the CellCover markers that have low sensitivity also have a lower rank in the DE. For example, in the 1H, the CellCover marker Hspa1a has the lowest sensitivity among all markers, expressed only near 13% percent of the time in this cell age, but its rank from DE is more than 500 (Supplementary Table 3). Although this gene has low sensitivity, it has high specificity, rendering it a good candidate for a marker gene when used with other complementary genes. Genes with these properties appear at much lower ranks in DE and are hence not selected when using that method.

3.8 Transfer of covering marker gene panels #2: Conserved cell type specific temporal progression in mammalian neocortical development

Using covering marker gene panels for the three principal cell ages in the Telley data, we have shown that CellCover can precisely define the major cell classes in mammalian neurogenesis across multiple scRNA-seq data sets (Figures 1 and 2). To evaluate the ability of covering marker gene panels to delineate temporal change within these cell types across cortical development, we have conducted similar transfer experiments using CellCover marker panels for all 12 cell classes in the Telley data as described above (Figures 1-2). Figure 5A shows the transfer of these 12 covering panels into the LaManno mouse brain atlas data [26]. As expected for each of the three main cell types across neurogenesis: (1) radial glia in the LaManno data show the highest covering rates of gene panels derived from 1H cells in the Telley data (Figure 5A, left panel, in blue), (2) neuroblasts in LaManno have highest covering rates of 24H markers (Figure 5A, center panel, in green), (3) neurons have the highest rates of 96H markers (Figure 5A, right panel, in red).

Importantly, in the radial glia of the LaManno data, the four 1H cell class panels spanning E12-E15 display sequentially ordered temporal patterns across developmental ages (sequential peaks of blue lines in Figure 5A, left panel). Notably, radial glia beyond E15 in the laManno data begin to express neuronal (96H) markers. This is consistent with conclusions of Telley et al. [25] and and a previous report of ours [40], which note that late NPCs begin to express markers of postmitotic neurons. Here, we extend this observation by showing that this is not the emergence of a non-specific neuronal identity but rather that late NPCs specifically begin to express markers of the later-born neurons they are about to produce (Figure 5A, left panel, E15.96H markers in the red dotted line).

While 24H markers show highest covering rates in neuroblasts of the laManno data (Figure 5A, center panel, in green), neurons also show notable levels of these markers (Figure 5A, right panel, in green). This is consistent with the annotation by Telley et al. [25] of both progenitor and neuronal sub-populations within their 24H cells. That is, 24 hours following terminal division, some NPCs have progressed to an early neuronal state, while others have not fully exited the progenitor state (denoted as basal progenitors by Telley et al. [25]). It is also important to note in this analysis that Eday labels from the Telley data indicate the age of animals when cells were labeled. Hence E12.96H cells were labeled on E12 but harvested and sequenced on E16. This explains the later, and still clear, temporal progression of the Telley 96H covering marker gene panels in the neurons of the laManno data (Figure 5A, right panel, in red).

We also transferred the CellCover marker panels from the twelve Telley cell classes into bulk RNA-seq data from the developing human neocortex [30] (Figure 5B). Levels of the 1H covering panels from E12-14 are highest at the earliest time point (at GW8) and descend through the second trimester, indicating peak expression of these panels lies prior to the window of human cortical development studied here (Figure 5B, top row of panels). The gene panel from 1H cells derived from E15, in contrast, peaks at GW14, indicating that the progenitor state captured in the mouse E15.1H cells is conserved as a more advanced state in the human, potentially indicating a shift to gliogenesis. Transferred expression levels of the 24H covering panels all peak in the window of development captured in the human data, with panels from E12 and E13 peaking (at GW12) before the E14 and E15 panels (at GW16; Figure 5B middle row of panels). These sequential dynamics demonstrate a parallel between the temporal progression of delaminating ventricular progenitors or basal progenitors (BPs) in the mouse and human which map to specific ages in both species: the shift from E12/13 to E14/E15 BPs in the mouse corresponds to a change in human BPs that occurs between GW12 and GW16. In a trend inverse of that observed in the transfer of the 1H covering panels, the 96H panels show coordinated increases from GW12-16 and appear to begin to plateau after GW16 (Figure 5B, bottom row of panels). Consistent with a later peaking state, there is less plateauing of the E15.96H covering panel signal in the human data.

Transfer of the twelve mouse covering panels into laser microdissected tissue from the developing macaque [27] also revealed conserved temporal elements of neocortical development (Figure 6A). As we observed in the transfer of the aggregated covering panels (Figure 2B), here we see the 1H panels with high expression in the cells of the VZ, 24H panels highest in the iSVZ, and 96H panels in the cortex (Ctx). Examination of the temporal patterns of these covering panels in their corresponding cortical laminae shows that the temporal progression of these cell types that were captured in the mouse is conserved in the macaque (Figure 6A, note the progressively later shifts in nonlinear fits within each lamina). This definition of conserved molecular progression in shared mammalian neural progenitors populations may help elucidate how neurons sharing defining molecular, morphological, and physiological features are produced contemporaneously in the mammalian cortex.

The E15.96H panel has a particularly distinct pattern, not neuron-specific like the earlier 96H panels, but rather a pan-cellular signature of the maturing cortex (Figure 6A, far right in bottom row of panels). Transfer of this E15.96H panel into additional late fetal and postnatal human cortical data shows that, unlike the earlier E12-E14.96H panels, expression of this panel continues to increase dramatically in the third trimester and after birth (Figure 6B). Together, these observations may reflect the shift from neurogenic-only states in earlier 96H cells to E15 when gliogenesis has begun in addition to neurogenesis. Further supporting this conclusion, levels of E12, E13 and E14 96H covering panels are lowest in white matter, while the E15.96H panel shows quite high levels in white matter, indicating that this latest covering panel has captured the beginning of gliogenesis (Figure 6C, gray points). Also striking are the elevated levels of this E15.96H panel specifically in layer I (containing the most superficial and latest-born neurons of the cortex), which contrast with the deep layer enriched signal of the E12.96H panel (Figure 6C, nonlinear fits). These transfers of the covering gene marker panels map the timing of cell identities we have learned from the mouse data onto the temporal progression of cortical development in the primate and human. We propose CellCover coupled with transfers of this kind as a general tool in mapping cell states across experimental systems and species.

4 Discussion

CellCover represents a novel algorithm for the identification of cell type marker gene panels in scRNA-seq data. Here, using data from the developing mammalian neocortex, we have shown that CellCover captures cell type specific signals in mouse that are conserved across primate and human cortical development. Additionally, using ground truth temporal labeling of cells in the mouse cortex, we have mapped temporal progressions of murine neocortical development onto the non-human primate and human developmental time courses.

We invite researchers to interrogate all the public data resources we examine here in the NeMO Analytics multi-omics data exploration environment that is designed to provide biologists with no programming expertise the ability to perform powerful analyses across collections of data in brain development: [1]. Users can visualize individual genes or groups of genes simultaneously across multiple datasets, including automated loading of the CellCover gene marker panels of murine cell types we have defined here. The code for CellCover is written in R and the Gurobi R interface and is available at [3].

5 Methods

5.1 Computing marker sets

5.1.1 Goals

The goal of our method is to find a relatively small set of genes for each cell type that is sufficient to label each individual cell. We associate to the observation of a (random) cell of a certain tissue a random variable $X = (X_g, g \in G)$ where X_g denotes the expression level of gene g. We denote the cell type labels of cells by a categorical random variable $Y \in \{1, 2, \dots K\}$.

Fix a target cell type, say, k. We will declare that a set of genes, say, M, provides a good marker set for this cell type if (1) with high probability, a sufficient number of genes in M are expressed if the cell type is k, and (2) genes in M are expressed with low probability if the cell type is not k. We note that specifying M is equivalent to specifying a family of binary variables $z = (z_g, g \in G)$, taking $M = \{g : z_g = 1\}$.

We will dedicate most of this discussion to the formulation of optimization problems that quantify conditions (1) and (2). But we can already note that these conditions differ from what is most often accepted as a definition of a marker gene, as a gene which is expressed with a significantly larger probability in type k than in any other type, or, alternatively, as a gene whose expression is significantly larger in type k than in any other type. Such definitions, leading, e.g., to marker genes having large differential expression, define an individual gene as marker or not. Our definition, however, defines *marker sets*, i.e., relies on the collective behavior of the genes selected in M.

We now make more precise our definition of marker sets, which will take us to an integer linear problem (ILP).

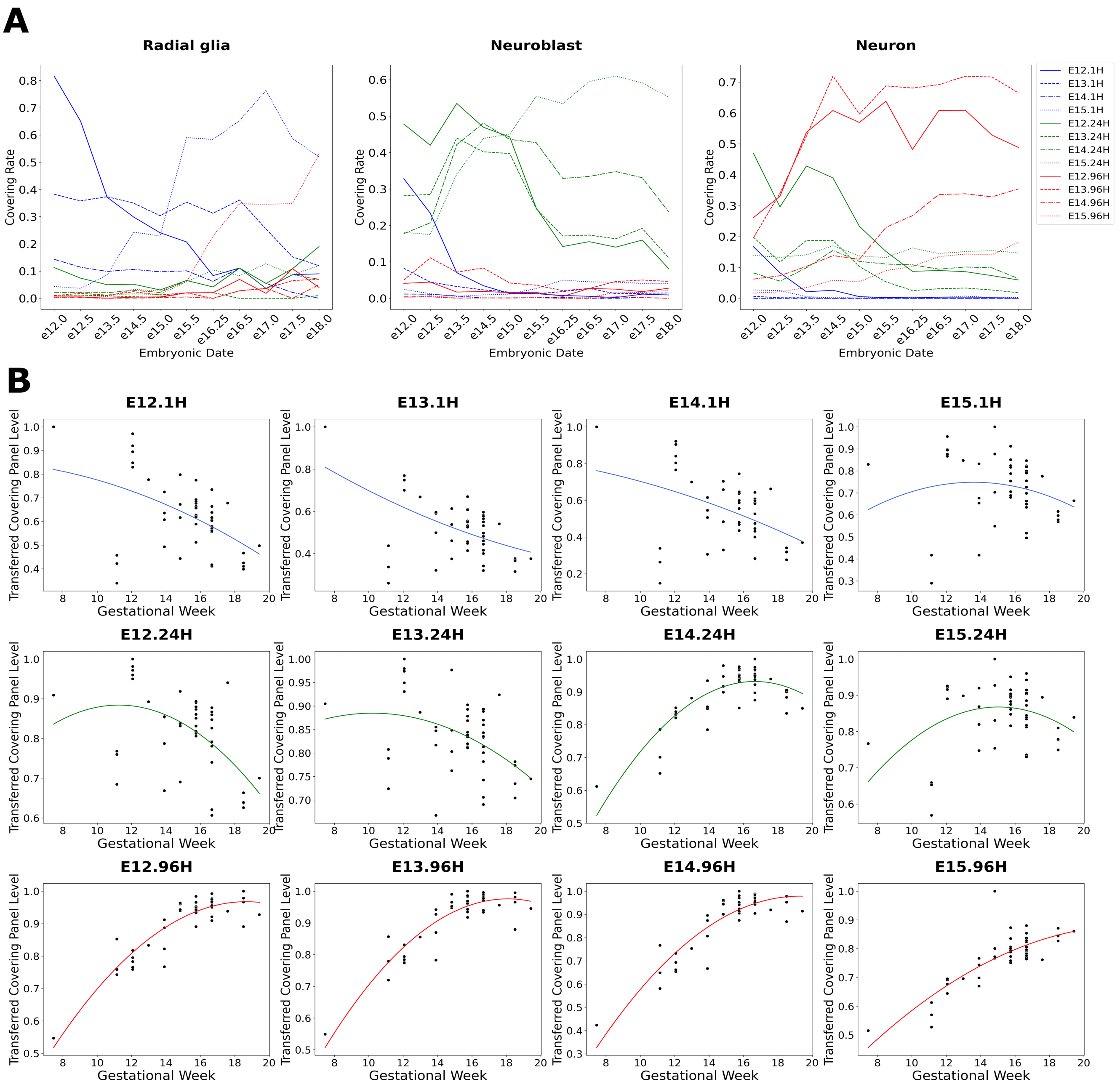


Figure 5: **Transfer of 12 cell class covering panels from the Telley data into additional mouse, human and macaque developing cortex data. A.** Transfer of Telley covering gene panels into the radial glia (left panel), neuroblasts (center panel) and neurons (right panel) from the developing mouse brain atlas (La Manno et al. [26]). In all cases, covering rates of transferred panels from 1H cells are depicted in blue, 24H in green, and 96H in red. Transferred levels were calculated as the proportion of cells of each type expressing more than 3 marker genes in the covering panel. **B.** Transfer of the 12 covering gene panels into bulk RNA-seq data from the human fetal cortex [30]. Transferred values of covering panels were assessed as the sum of covering panel gene expression levels in each individual sample divided by the maximum sum observed in the samples (as in Figure 2).

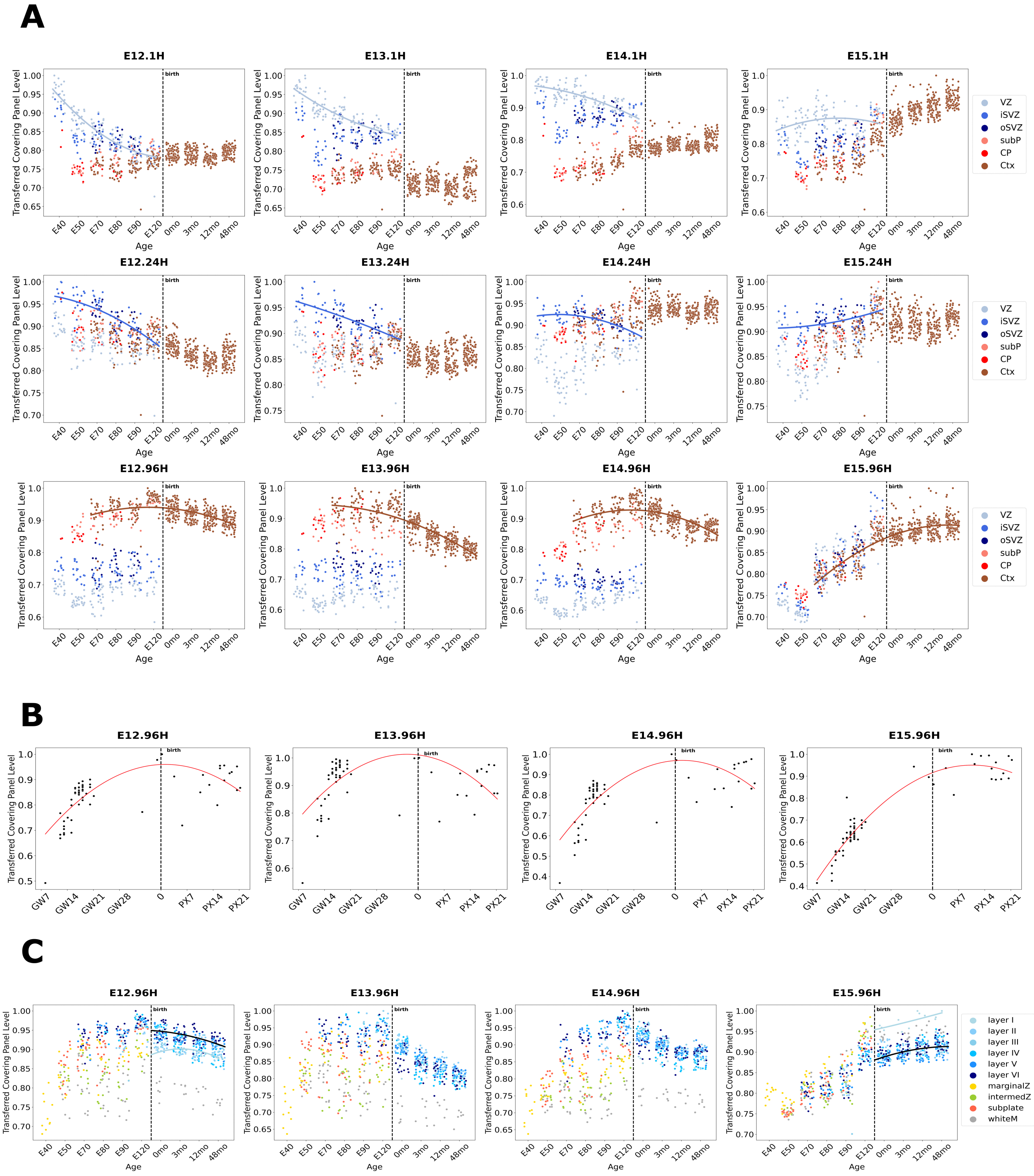


Figure 6: **A.** Transfer of the 12 covering panels into microarray data from microdissected regions of the developing macaque neocortex [27]. X-axis ages are expressed as embryonic (E) days after conception and months (mo) after birth. Transferred covering panel levels were calculated as in Figure 2. VZ = ventricular zone, iSVZ= inner subventricular zone, oSVZ = outer subventricular zone, subP = subplate, CP = cortical plate, Ctx = cortex. **B.** Repeated transfer of the 96H covering panels into the human cortex data [30], showing additional late fetal and early postnatal samples in postnatal development. **C.** Repeated transfer of the 96H covering panels into the microdissected macaque data, using additional labeling of dissections by cortical layer.

5.1.2 Binarization

We use a very simple definition of a gene being expressed, namely g is expressed if and only if $X_g > 0$. This operation can be refined by replacing the zero threshold by some low activity number θ (or θ_g), but our experiments with single-cell data that typically contains a high proportion of zeros show that this simple binarization rule already retains a large quantity of relevant information (see section 3.2). For convenience, we introduce the random variable U_g equal to 1 if $X_g > 0$ and to 0 otherwise.

5.1.3 First condition

We now formulate condition (i) as a constraint on the set M and introduce two parameters of our method, namely an integer $\boldsymbol{d}$ (or "depth") and a scalar $0 \leq \alpha \ll 1$ to quantify this condition as "the probability that at least $\boldsymbol{d}$ genes in M are expressed given that the cell type is k is larger than $1 - \alpha$," or

$$P\Big(\sum_{g\in M} U_g \geq \boldsymbol{d} \Big| Y = k\Big) \geq 1 - \alpha. \tag{1}$$

The left-hand side of the inequality (1) is the rate at which the set M covers the class k, or covering rate. Obviously, this constraint can be satisfied for some $M \subset G$ if and only if it is satisfied for $M = G$, so that α needs to be chosen such that

$$P\Big(\sum_{g\in G} U_g \geq \boldsymbol{d} \Big| Y = k\Big) \geq 1 - \alpha.$$

Note also that, using the notation introduced in the previous section,

$$\sum_{g\in M} U_g = \sum_{g\in G} U_g z_g.$$

This condition can also be interpreted in terms of a classification error. Indeed, define $\Phi_{M,\boldsymbol{d}}$ as the binary classifier equal to one if $\sum_{g\in M} U_g \geq \boldsymbol{d}$ and to 0 otherwise. Then (1) expresses the fact that this classifier's sensitivity is at least $1 - \alpha$.

5.1.4 Second condition

Having expressed the first condition as a constraint, we now express the second one as a quantity $M \mapsto F(M)$ to optimize. Ideally, one would like to minimize the probability that, conditionally to $Y_k = 0$, there exists at least one active gene in M, namely

$$P\Big(\sum_{g\in M} U_g \geq 1 \Big| Y \neq k\Big).$$

This quantity, however, is not straightforward to minimize, but its Bonferroni upper bound

$$F_0(M) = \sum_{g\in M} P(U_g = 1|Y \neq k) = \sum_{g\in G} P(U_g = 1|Y \neq k) z_g$$

is linear in the binary vector z and will be amenable to an ILP formulation. ($F_0(M)$ is also equal to the expectation of $\sum_{g\in M} U_g$ given $Y \neq k$.)

Note that, obviously

$$P\Big(\sum_{g\in M} U_g \geq r \Big| Y \neq k\Big) \leq P\Big(\sum_{g\in M} U_g \geq 1 \Big| Y \neq k\Big) \leq \sum_{g\in G} P(U_g = 1|Y \neq k) z_g$$

so that $1 - F_0(M)$ controls the sensibility of the previous classifier $\Phi_{M,r}$.

If the variables U_g are independent (conditionally to Y), then

$$P\Big(\sum_{g\in M} U_g \geq 1 \Big| Y \neq k\Big) = 1 - \prod_{g\in M} (1 - P(U_g = 1|Y \neq k)),$$

which suggests using

$$F_1(M) = -\sum_{g\in M} \log(1 - P(U_g = 1|Y \neq k))$$

as an alternative objective.

5.1.5 Margin weights

More generally, we will consider objective functions F taking the form

$$F(M) = \sum_{g\in G} w_g z_g$$

for some choice of weights $(w_g, g \in G)$. The function F_0 above uses $w_g = P(U_g = 1|Y \neq k)$, and F_1 uses $w_g = -\log(1 - P(U_g = 1|Y \neq k))$, but other choices are possible. We will in particular work with weights reflecting the ability of the variable U_g to identify class k, such as

$$w_g = \frac{\sum_{l\neq k} P(U_g = 1|Y = l)/(K-1)}{P(U_g = 1|Y = k)}. \tag{2}$$

or

$$w_g = \frac{\max_{l\neq k} P(U_g = 1|Y = l)}{P(U_g = 1|Y = k)}. \tag{3}$$

where K is the total number of classes. Small values of w_g reflect a large sensibility of the variable U_g, when differentiating type k from any other type. In our algorithm, we implement Equation (2) as the default while offering users the flexibility to choose their desired weight schemes. Going further, it is natural to eliminate genes for which one of the ratios above is larger than 1, which is achieved by enforcing $z_g = 0$ if $w_g > 1$ (or, equivalently, replacing w_g by $+\infty$ when $w_g > 1$). This choice has, in addition, the merit of reducing the number of free variables z_g and accelerating the optimization process. Note that these margin weights ensure that $F(M) \geq F_0(M)$ and therefore also control sensibility.

5.1.6 Marker set estimation algorithm

We can now conclude our discussion with the description of our algorithm, based on the observed training data, provided by a collection of N cells, with expressions $(x_g^{(c)}, g \in G)$ for $c = 1, \ldots, N$, yielding binary variables $(u_g^{(c)}, g \in G)$, and types $y^{(c)}$ for $c = 1, \ldots, N$. The following returns a marker set for type k, with parameters $\boldsymbol{d}$ and α. Let S_k be the set of cells with type k.

(1) Estimate empirical probabilities $\hat{P}(U_g = 1|Y = l)$ for $g \in G$ and $l = 1, \ldots, K$ based on training data.

(2) Compute weights $w_g, g \in G$, using equation (2) (or use user-supplied weights).

(3) Let $G_k = \{g : w_g \leq 1\} \cap \{g : P(U_g = 1|Y = k) \geq \rho\}$.

(4) Use an ILP software to minimize, with respect to binary variables $z_g, g \in G_0$ and $s^{(c)}, c \in S_k$, the function $\sum_{g\in G_0} w_g z_g$ subject to the constraints

$$\sum_{g\in G_k} z_g u_g^{(c)} \geq \boldsymbol{d}(1 - s^{(c)}), \quad c \in S_k$$

and

$$\sum_{c\in S_k} s^{(c)} \leq \alpha|S_k|.$$

(5) Return the set $M_k = \{g \in G_k : z_g = 1\}$

5.2 Nested marker set expansion

Step (4) can be modified to estimate a covering as a superset of a previous one (obtained e.g., with a smaller depth): Let $(z_g^{(0)})$ be a vector of binary variables (fixed below). To extend this set to a covering at depth $\boldsymbol{d}$, one only needs to replace step (4) by the following:

(4') Use an ILP software to minimize, with respect to binary variables $z_g, g \in G_0$ and $s^{(c)}, c \in S_k$, the function $\sum_{g\in G_0} w_g z_g$ subject to the constraints

$$z_g \geq z_g^{(0)}, \quad g \in G_k$$

$$\sum_{g\in G_k} z_g u_g^{(c)} \geq \boldsymbol{d}(1 - s^{(c)}), \quad c \in S_k$$

and

$$\sum_{c\in S_k} s^{(c)} \leq \alpha|S_k|.$$

5.2.1 Remarks

(a) Even though it was not designed with this intent, our marker set estimation algorithm can be interpreted as an asymmetric classification method to separate type k from all other types. In an approach akin to classical statistical testing, we determine a classifier with some free parameters (here, the gene set M) that we require to have a minimal sensitivity, $1 - \alpha$ (so that α is reminiscent of a type I error). The free parameters are then optimized in order to minimize an objective function that controls (among other things) the sensibility of the classifier.

(b) If we take $w_g = 1$ for all g, the previous algorithm only depends on training data associated with cell type k. The resulting algorithm provides a *minimal* gene set that covers the considered population, in the sense that all cells in that population (except a fraction α) has at least $\boldsymbol{d}$ active genes in the selected set. With a suitable definition of what is meant by being active, this covering algorithm was introduced in Ke et al. [38] and applied to the determination of important gene motifs in a population of tumor cell associated with a specific phenotype.

5.3 Over-representation Analysis

Over-representation analysis was performed using the Fisher's exact test using gene sets from the GO database. Over-represented gene sets were marked as significant by a FDR<0.1 cutoff using the Benjamini-Hochberg method. The background gene list was defined as all the genes present in the Telley dataset.

6 Acknowledgements

Data sharing and visualization via NeMO Analytics was supported by grants R24MH114815 and R01DC019370. The work of DG, LY, and LJ was partially supported by NIH National Cancer Institute Grant R01CA200859. The work of DG, LY, and AW was partially supported by NSF Award 2124230.

References

[1] Nemo analytics profile for cellcover, 2022. URL https://nemoanalytics.org/p?l=CellCover.

[2] Seth A Ament, Ricky S Adkins, Robert Carter, Elena Chrysostomou, Carlo Colantuoni, Jonathan Crabtree, Heather H Creasy, Kylee Degatano, Victor Felix, Peter Gandt, Gwenn A Garden, Michelle Giglio, Brian R Herb, Farzaneh Khajouei, Elizabeth Kiernan, Carrie McCracken, Kennedy McDaniel, Suvarna Nadendla, Lance Nickel, Dustin Olley, Joshua Orvis, Joseph P Receveur, Mike Schor, Shreyash Sonthalia, Timothy L Tickle, Jessica Way, Ronna Hertzano, Anup A Mahurkar, and Owen R White. The neuroscience multi-omic archive: a brain initiative resource for single-cell transcriptomic and epigenomic data from the mammalian brain. *Nucleic Acids Research*, 51(D1), 2023. URL https://doi.org/10.1093/nar/gkac962.

[3] Covering package online repository, 2022. URL https://github.com/lanlanji/CoveringPackage.

[4] Philipp Angerer, Lukas Simon, Sophie Tritschler, F Wolf, David Fischer, and Fabian Theis. Single cells make big data: New challenges and opportunities in transcriptomics. *Current Opinion in Systems Biology*, 4, 2017. doi: 10.1016/j.coisb.2017.07.004.

[5] Dominic Grün and Alexander van Oudenaarden. Design and Analysis of Single-Cell Sequencing Experiments. *Cell*, 163(4):799–810, nov 2015. ISSN 1097-4172 (Electronic). doi: 10.1016/j.cell.2015.10.039.

[6] Aviv Regev, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, Hans Clevers, Bart Deplancke, Ian Dunham, James Eberwine, Roland Eils, Wolfgang Enard, Andrew Farmer, Lars Fugger, Berthold Göttgens, Nir Hacohen, Muzlifah Haniffa, Martin Hemberg, Seung Kim, Paul Klenerman, Arnold Kriegstein, Ed Lein, Sten Linnarsson, Emma Lundberg, Joakim Lundeberg, Partha Majumder, John C Marioni, Miriam Merad, Musa Mhlanga, Martijn Nawijn, Mihai Netea, Garry Nolan, Dana Pe'er, Anthony Phillipakis, Chris P Ponting, Stephen Quake, Wolf Reik, Orit Rozenblatt-Rosen, Joshua Sanes, Rahul Satija, Ton N Schumacher, Alex Shalek, Ehud Shapiro, Padmanee Sharma, Jay W Shin, Oliver Stegle, Michael Stratton, Michael J T Stubbington, Fabian J Theis, Matthias Uhlen, Alexander van Oudenaarden, Allon Wagner, Fiona Watt, Jonathan Weissman, Barbara Wold, Ramnik Xavier, Nir Yosef, and Human Cell Atlas Meeting Participants. Science Forum: The Human Cell Atlas. *eLife*, 6: e27041, dec 2017. ISSN 2050-084X. doi: 10.7554/eLife.27041. URL https://doi.org/10.7554/eLife.27041.

[7] Valentine Svensson, Roser Vento-Tormo, and Sarah A Teichmann. Exponential scaling of single-cell RNA-seq in the past decade. *Nature Protocols*, 13(4):599–604, 2018. ISSN 1750-2799. doi: 10.1038/nprot.2017.149. URL `https://doi.org/10.1038/nprot.2017.149`.

[8] Yong Wang and Nicholas E Navin. Advances and applications of single-cell sequencing technologies. *Molecular cell*, 58(4):598–609, may 2015. ISSN 1097-4164. doi: 10.1016/j.molcel.2015.05.005. URL `https://pubmed.ncbi.nlm.nih.gov/26000845` `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4441954/`.

[9] Conor Delaney, Alexandra Schnell, Louis V Cammarata, Aaron Yao-Smith, Aviv Regev, Vijay K Kuchroo, and Meromit Singer. Combinatorial prediction of marker panels from single-cell transcriptomic data. *Molecular systems biology*, 15(10):e9005, 2019.

[10] Malte D Luecken and Fabian J Theis. Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular systems biology*, 15(6):e8746, 2019.

[11] Karthik Shekhar, Sylvain W Lapan, Irene E Whitney, Nicholas M Tran, Evan Z Macosko, Monika Kowalczyk, Xian Adiconis, Joshua Z Levin, James Nemesh, Melissa Goldman, et al. Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. *Cell*, 166(5):1308–1323, 2016.

[12] The Tabula Muris Consortium. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature*, 562(7727):367–372, 2018. ISSN 1476-4687. doi: 10.1038/s41586-018-0590-4. URL `https://doi.org/10.1038/s41586-018-0590-4`.

[13] Christopher Heje Grønbech, Maximillian Fornitz Vording, Pascal N Timshel, Casper Kaae Sønderby, Tune H Pers, and Ole Winther. scVAE: variational auto-encoders for single-cell gene expression data. *Bioinformatics*, 36(16):4415–4422, 2020. ISSN 1367-4803. doi: 10.1093/bioinformatics/btaa293. URL `https://doi.org/10.1093/bioinformatics/btaa293`.

[14] Rahul Satija, Jeffrey A Farrell, David Gennert, Alexander F Schier, and Aviv Regev. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, 33(5):495–502, 2015. ISSN 1546-1696. doi: 10.1038/nbt.3192. URL `https://doi.org/10.1038/nbt.3192`.

[15] Guo-Cheng Yuan, Long Cai, Michael Elowitz, Tariq Enver, Guoping Fan, Guoji Guo, Rafael Irizarry, Peter Kharchenko, Junhyong Kim, Stuart Orkin, John Quackenbush, Assieh Saadatpour, Timm Schroeder, Ramesh Shivdasani, and Itay Tirosh. Challenges and emerging directions in single-cell analysis. *Genome Biology*, 18(1):84, 2017. ISSN 1474-760X. doi: 10.1186/s13059-017-1218-y. URL `https://doi.org/10.1186/s13059-017-1218-y`.

[16] Etienne Becht, Leland McInnes, John Healy, Charles-Antoine Dutertre, Immanuel W H Kwok, Lai Guan Ng, Florent Ginhoux, and Evan W Newell. Dimensionality reduction for visualizing single-cell data using UMAP. *Nature biotechnology*, dec 2018. ISSN 1546-1696 (Electronic). doi: 10.1038/nbt.4314.

[17] Bianca Dumitrascu, Soledad Villar, Dustin G Mixon, and Barbara E Engelhardt. Optimal marker gene selection for cell type discrimination in single cell analyses. *Nature Communications*, 12(1):1186, 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-21453-4. URL `https://doi.org/10.1038/s41467-021-21453-4`.

[18] Tamim Abdelaal, Lieke Michielsen, Davy Cats, Dylan Hoogduin, Hailiang Mei, Marcel J T Reinders, and Ahmed Mahfouz. A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biology*, 20(1):194, 2019. ISSN 1474-760X. doi: 10.1186/s13059-019-1795-z. URL `https://doi.org/10.1186/s13059-019-1795-z`.

[19] Oliver Stegle, Sarah A Teichmann, and John C Marioni. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics*, 16(3):133–145, 2015. ISSN 1471-0064. doi: 10.1038/nrg3833. URL `https://doi.org/10.1038/nrg3833`.

[20] J B Angevine, Jr and R L Sidman. Autoradiographic study of cell migration during histogenesis of cerebral cortex in the mouse. *Nature*, 192(4804):766–768, November 1961.

[21] Madeline A Lancaster and Juergen A Knoblich. Spindle orientation in mammalian cerebral cortical development. *Curr. Opin. Neurobiol.*, 22(5):737–746, October 2012.

[22] Marion Betizeau, Veronique Cortay, Dorothée Patti, Sabina Pfister, Elodie Gautier, Angèle Bellemin-Ménard, Marielle Afanassieff, Cyril Huissoud, Rodney J Douglas, Henry Kennedy, and Colette Dehay. Precursor diversity and complexity of lineage relationships in the outer subventricular zone of the primate. *Neuron*, 80(2):442–457, October 2013.

[23] David V Hansen, Jan H Lui, Philip R L Parker, and Arnold R Kriegstein. Neurogenic radial glia in the outer subventricular zone of human neocortex. *Nature*, 464(7288):554–561, March 2010.

[24] Pasko Rakic, Albert E Ayoub, Joshua J Breunig, and Martin H Dominguez. Decision by division: making cortical maps. *Trends Neurosci.*, 32(5):291–301, May 2009.

[25] Ludovic Telley, Gulistan Agirman, Julien Prados, Nicole Amberg, Sabine Fièvre, Polina Oberst, Giorgia Bartolini, Ilaria Vitali, Christelle Cadilhac, Simon Hippenmeyer, et al. Temporal patterning of apical progenitors and their daughter neurons in the developing neocortex. *Science*, 364(6440), 2019.

[26] Gioele La Manno, Kimberly Siletti, Alessandro Furlan, Daniel Gyllborg, Elin Vinsland, Alejandro Mossi Albiach, Christoffer Mattsson Langseth, Irina Khven, Alex R Lederer, Lisa M Dratva, et al. Molecular architecture of the developing mouse brain. *Nature*, 596(7870):92–96, 2021.

[27] Trygve E Bakken, Jeremy A Miller, Song-Lin Ding, Susan M Sunkin, Kimberly A Smith, Lydia Ng, Aaron Szafer, Rachel A Dalley, Joshua J Royall, Tracy Lemon, et al. A comprehensive transcriptional map of primate brain development. *Nature*, 535(7612):367–375, 2016.

[28] Damon Polioudakis, Luis de la Torre-Ubieta, Justin Langerman, Andrew G Elkins, Xu Shi, Jason L Stein, Celine K Vuong, Susanne Nichterwitz, Melinda Gevorgian, Carli K Opland, et al. A single-cell transcriptomic atlas of human neocortical development during mid-gestation. *Neuron*, 103(5):785–801, 2019.

[29] Aparna Bhaduri, Madeline G Andrews, Walter Mancia Leon, Diane Jung, David Shin, Denise Allen, Dana Jung, Galina Schmunk, Maximilian Haeussler, Jahan Salma, et al. Cell stress in cortical organoids impairs molecular subtype specification. *Nature*, 578(7793):142–148, 2020.

[30] Andrew E Jaffe, Richard E Straub, Joo Heon Shin, Ran Tao, Yuan Gao, Leonardo Collado-Torres, Tony Kam-Thong, Hualin S Xi, Jie Quan, Qiang Chen, et al. Developmental and genetic regulation of the human cortex transcriptome illuminate schizophrenia pathogenesis. *Nature neuroscience*, 21(8):1117–1125, 2018.

[31] Isabella N Grabski and Rafael A Irizarry. A probabilistic gene expression barcode for annotation of cell types from single-cell RNA-seq data. *Biostatistics*, 23(4):1150–1164, October 2022.

[32] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939785. URL `http://doi.acm.org/10.1145/2939672.2939785`.

[33] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[34] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[35] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[36] Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck III, Shiwei Zheng, Andrew Butler, Maddie J. Lee, Aaron J. Wilk, Charlotte Darby, Michael Zagar, Paul Hoffman, Marlon Stoeckius, Efthymia Papalexi, Eleni P. Mimitou, Jaison Jain, Avi Srivastava, Tim Stuart, Lamar B. Fleming, Bertrand Yeung, Angela J. Rogers, Juliana M. McElrath, Catherine A. Blish, Raphael Gottardo, Peter Smibert, and Rahul Satija. Integrated analysis of multimodal single-cell data. *Cell*, 2021. doi: 10.1016/j.cell.2021.04.048. URL `https://doi.org/10.1016/j.cell.2021.04.048`.

[37] Ka Yee Yeung and Walter Ruzzo. Details of the Adjusted Rand index and Clustering algorithms Supplement to the paper "An empirical study on Principal Component Analysis for clustering gene expression data" (to appear in Bioinformatics). *Science*, 17, 2001.

[38] Qian Ke, Wikum Dinalankara, Laurent Younes, Donald Geman, and Luigi Marchionni. Efficient Representations of Tumor Diversity with Paired DNA-RNA Anomalies. *bioRxiv*, 2020. doi: 10.1101/2020.04.24.060129. URL https://www.biorxiv.org/content/early/2020/05/31/2020.04.24.060129.

[39] Alan Brown, Alexey Amunts, Xiao-Chen Bai, Yoichiro Sugimoto, Patricia C Edwards, Garib Murshudov, Sjors H W Scheres, and V Ramakrishnan. Structure of the large ribosomal subunit from human mitochondria. *Science*, 346(6210):718–722, November 2014.

[40] Huan Chen, Brian Caffo, Genevieve Stein-O'Brien, Jinrui Liu, Ben Langmead, Carlo Colantuoni, and Luo Xiao. Two-stage linked component analysis for joint decomposition of multiple biologically related data sets. *Biostatistics*, 23(4):1200–1217, October 2022.

[41] Crabp2 marking early npc state in the primate - a visualization using nemo, 2022. URL https://nemoanalytics.org/p?l=NeocortexEvoDevo&g=CRABP2.